UNIVERSITY OF KWAZULU-NATAL

Using educational data mining to predict sub-Saharan African science, technology, engineering, and mathematics students' academic performance: A systematic review

By

Langelihle Lucky Mhlongo

217071140

A dissertation submitted in fulfilment of the requirements for the degree of

Master of Commerce

School of Management, IT and Governance

College of Law and Management Studies

Supervisor: Prof Irene Govender

Co-supervisor: Dr. Rosemary Diane Quilling

2023

Declaration

I Langelihle Lucky Mhlongo ……declare that

(i) The research reported in this dissertation/thesis, except where otherwise indicated, is my original research.

(ii) This dissertation/thesis has not been submitted for any degree or examination at any other university.

(iii) This dissertation/thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

(iv) This dissertation/thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

a) their words have been re-written but the general information attributed to them has been referenced;
b) where their exact words have been used, their writing has been placed inside quotation marks, and referenced.

(v) Where I have reproduced a publication of which I am an author, co-author or editor, I have indicated in detail which part of the publication was actually written by myself alone and have fully referenced such publications.

(vi) This dissertation/thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the dissertation/thesis and in the References sections.

Signature:

Date: 10/12/2023

Dedication

I dedicate this thesis to my forever-loving, prayerful, and caring late mother for her love and support. You have invested a lot in this journey that I have embarked on. Unfortunately, you are not here to witness this day. Thank you for everything. I will always love you, and may you continue living in the bosom of the Lord.

Acknowledgements

I want to thank my supervisors, Professor Irene Govender and Dr. Rosemary Diane Quilling, for their assistance, motivation, advice, and support throughout this research. Their advice moulded the research idea into reality, and their support and empathy also contributed to the continuation and completion of this research. I also want to thank Ms. Deborah Cunynghame for her continuous assistance and help with the administrative tasks of my research. I also want to thank my family, friends, and colleagues who provided the much-needed moral and technical support. Finally, I want to thank the almighty God for giving me the idea, will, and a constant supply of strength to complete my research.

Abstract

The growing pervasiveness of technology in the modern world has expanded the critical roles of Science, Technology, Engineering, and Mathematics (STEM) fields to drive economic expansion through development and innovation, thus generating more jobs. This has caused an emergence of research in Educational Data Mining (EDM), which has enabled higher educational institutions (HEIs) to see areas that need improvement and to use available resources for early intervention efficiently. This research followed the systematic literature review (SLR) approach, aligning with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) framework. Focusing on a five-year (2017-2021) time frame of studies that concentrated on EDM's effectiveness on students' academic performance prediction in sub-Saharan African HEIs, 41 studies were obtained for further analysis. These studies were selected based on a set of criteria in the SLR process.

The findings of this SLR showed that interest in EDM has increased in recent years. Moreover, the findings revealed that the most frequently used factors to predict academic performance are categorised under five categories: students' previous and current class performances, demographics, socio-economic data, and e-Learning behaviours. According to the findings, Decision Trees, Nave Bayes, Artificial Neural Networks, Regression, Support Vector Machines, Random Forests, and K-Nearest Neighbours are the most frequently used EDM methods. Researchers were able to predict academic performance at three levels: course, year, and degree. While most research focused on degree and course level prediction, few studies focused on predicting academic performance at the year level.

Moreover, predictions conducted at the year level were less effective than the course and degree-level predictions. This suggests researchers may need to increase focus on year-level predictions, particularly on factors that increase or decrease prediction accuracy, to improve prediction accuracy at the year level. Additionally, this study developed the Higher Educational Data Mining (HEDM) conceptual framework to enhance longstanding efforts in the EDM field to advance knowledge production systematically and coherently for academic performance prediction.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Introduction

Preparing educated workers for careers in the Science, Technology, Engineering, and Mathematics (STEM) fields is crucial to every country's technological advancement, scientific innovation, competitiveness, and economic development (Bengesai & Pocock, 2021). Thus, higher educational institutions (HEIs) are one of the key parts of society, as HEIs are crucial to the development and growth of any country (Nwosu *et al.*, 2018). Every year HEIs produce large numbers of graduates and postgraduates. HEIs consider the students as their main assets. Hence, there is a growing need for student academic performance excellence in HEIs, which requires continual improvement. Student academic performance in HEIs is an essential benchmark to compare their quality (Adekitan & Salau, 2019). Benchmarking is essential to continual improvement. It includes investigating good practices that can be employed to improve students' academic performance. Even though HEIs employ best practices for teaching and learning, HEIs still face the issues of students dropping out, achieving low academic performance and being unemployed.

Lecturers, administrators, and other stakeholders should be concerned about students' academic performance (Bassi *et al.*, 2019). The concerns in South Africa focus on driving economic expansion by increasing productivity (Poh & Smythe, 2014). However, to achieve this economic expansion, there need to be enough qualified workers to meet the demands of every sector of the economy. Based on a report by Statistics South Africa, the unemployment rate increased by 0.2% from the fourth quarter of 2022 to 32.9% in the first quarter of 2023 (StatsSA, 2023). It is observed that the unemployment rate consists of many individuals that are either without matric, with matric, with some form of tertiary qualification, or a university dropout. These individuals shifted from being employed to not being economically active and to being unemployed throughout the two quarters, which is the cause for the increase of 0.2% in the unemployment rate (StatsSA, 2023). Dropouts negatively affect HEIs by reducing student enrolment and increasing the non-achievement of their objectives. Therefore, the focus should be on HEIs, and other stakeholders, as many students are failing to finish their degrees, contributing to the increased unemployment rate (Alban & Mauricio, 2019; Poh & Smythe, 2014).

Besides the unforeseen crisis of COVID-19, technologies such as Machine Learning (ML) and Artificial Intelligence (AI) have disrupted different job industries, making HEIs need to adapt and evolve to prepare students stepping into the fourth industrial revolution (4IR) era. Education increasingly relies on technology, and vast amounts of data about student activities are available in the academic environment. Records of student activities, ratings, and interactions with teachers and other students are now collected from learning management systems (LMS), including Blackboard and Moodle (Lee & McLoughlin, 2010; Raghavjee *et al.*, 2021). HEIs are starting to be more informed of the possibility of monitoring LMS data to improve teaching and learning quality (Miguéis *et al.*, 2018; Pelletier *et al.*, 2021). However, analysing students' academic performance is difficult because of the vast educational data that needs to be considered (Ofori *et al.*, 2020). Hence, practical tools such as data mining are needed to process student data.

The critical task of data mining (DM) uses different techniques to find and extract meaningful knowledge from data (Olaniyi *et al.*, 2017). When the data in a dataset originates from an educational context, the mining executed on such a dataset is labelled as educational data mining (EDM). Using statistical methods, EDM effectively reveals hidden patterns and meaningful information that may not be easily identifiable (Thakar, 2015). Educational data mining is done on a large scale as it enables HEIs to make data-driven decisions at all levels (Kovač & Oreški, 2018). Predicting students' academic performance is necessary as it can assist lecturers in identifying students who require educational intervention early (Hussain *et al.*, 2018).

Wanjau and Muketha (2018) state that predicting students' academic performance in HEIs is an intricate process of decision-making that relies on more than just exam scores. Literature suggests that students' academic performance, particularly in STEM, depends on a variety of factors, including family, socio-economic, personal, and other environmental factors (Girma, 2019; Wang, 2013; Wanjau & Muketha, 2018). Students in the STEM fields who cannot graduate on time negatively impact economic expansion due to a low throughput of entrepreneurs and skilled employees entering the economy in the 4IR era (Abe & Chikoko, 2020). EDM can assist HEIs in making informed decisions about their undergraduate students (Agarwal *et al.*, 2012). EDM can also help HEIs identify STEM students who will most likely not graduate on time. This study focused on EDM as it allows for the automated discovery of knowledge (Mgala, 2016). There was a lack of detailed reviews on EDM conducted in developing countries, and this research presented an opportunity to provide a scholarly review

in this domain. The study aimed to systematically review the effectiveness of EDM in predicting sub-Saharan African undergraduate STEM students' academic performance.

1.2 Problem Statement

In South Africa, public HEIs are colleges, comprehensive universities, and technology universities (Makombe & Lall, 2020). Private individuals or organisations own private HEIs. The state establishes and funds public HEIs using the Department of Higher Education and Training (DHET) (Makombe & Lall, 2020). These funds depend on the success rates of HEIs. Hence, HEIs must succeed financially to stay in business (Dumond & Johnson, 2013). The increasing pervasiveness of technology in modern society has expanded the critical roles of STEM fields to drive economic expansion and produce jobs through development and innovation (Aulck *et al.*, 2017). Additionally, the outbreak of COVID-19 has resulted in many HEIs shifting to an online mode of teaching and learning, which has brought about an increase in data collected and stored in different databases.

Sub-Saharan African HEIs, such as the University of Kwazulu-Natal (UKZN), have different systems that store student information. These systems include Student Information System (SIS) and Learning Management System (LMS). LMS, such as Moodle, is an online learning platform that dispenses knowledge simultaneously, as online behaviour is stored in each student's log (Hasan *et al.*, 2019). SIS stores related academic data, such as demographics and current or past academic performance (Raghavjee *et al.*, 2021). Student data held by these systems may help HEIs understand the dropout rates, uptake of STEM fields, and overall learning tendencies. However, HEIs may not be using this data meaningfully to assist in any decisions or policymaking for improving students' academic performance (Mashiloane, 2016). The problem of collecting and storing data is addressed in many cases, but not the problem of analysing the data for new knowledge generation (Hasan *et al.*, 2019). Hence, tools are needed to analyse the student data automatically to generate new knowledge.

EDM refers to applying DM methods to find valuable knowledge from data created in an educational context (Adekitan *et al.*, 2019; Dutt *et al.*, 2015). In contrast, Learning Analytics (LA) involves collecting, analysing, and reporting student data to improve learning and the contexts where learning takes place (Maphosa & Maphosa, 2020; Siemens & Baker, 2012). Information extracted from both EDM and LA brings new knowledge on trends that were unknown previously on students' academic performance and learning behaviours, quality and efficiency of education, and students' inclination to dropout (Adekitan & Noma-Osaghae,

2019; Kim *et al.*, 2018; Roy & Garg, 2017; Yang & Li, 2018). This brings the possibility to discover quality gaps, and improve strategies and policies (Adekitan & Salau, 2019; Osmanbegovic & Suljic, 2012). This study focused on EDM as it allows for the automated discovery of knowledge (Mgala, 2016).

Prediction in EDM typically involves creating a model, method, or algorithm that accepts specific variables and outputs expected results for the predicted variable (Nudelman *et al.*, 2018). Prediction in EDM generally has two meanings (Sokkhey *et al.*, 2020). First, prediction in EDM involves extracting key features such as age, prior experience, and matric scores that influence students' academic performance. Second, prediction in EDM involves predicting overall outputs such as academic scores, grades, dropout rates, and other performance measurements. In this study, the first meaning is used, as it is in line with the objective of the study.

Substantial work has been done using EDM methods, which fall into classification, clustering, and regression analysis (Shingari *et al.*, 2017). However, many untouched areas remain, and no unified approach is followed (Ugalde & Venkateswaran, 2018). There was a lack of detailed reviews on EDM conducted in developing regions, namely Sub-Saharan Africa. This research presented an opportunity to provide a scholarly study in this domain (Maphosa & Maphosa, 2020). This research aimed to provide a systematic literature review (SLR) on EDM's effectiveness in predicting sub-Saharan African undergraduate STEM student academic performance. This study explored factors such as student behaviour within LMS, academic results, and socio-economic and demographic factors. Additionally, this study examined various EDM methods, including Decision Trees, Support Vector Machines, k-Nearest Neighbours, Artificial Neural Networks, and Naive Bayes used to predict sub-Saharan African undergraduate STEM students' academic performance. This SLR aimed to create the prospect for researchers to use detailed reporting methodologies for further EDM research in the sub-Saharan African region and other developing countries, as it may reveal new knowledge for those contexts.

1.3 Research Questions

How effective is EDM in predicting sub-Saharan African undergraduate STEM students' academic performance?

- What key input factors are used in EDM to predict sub-Saharan African undergraduate STEM students' academic performance?
- What are the key student involvement factors used in EDM to predict sub-Saharan African undergraduate STEM students' academic performance?
- What are EDM's most frequently used tools and algorithms to predict sub-Saharan African undergraduate STEM students' academic performance?
- To what extent is EDM effective in predicting sub-Saharan African undergraduate STEM students' academic performance?

1.4 Research Objectives

To determine the effectiveness of EDM in predicting sub-Saharan African undergraduate STEM student**s'** academic performance.

- To identify the key input factors used in EDM to predict sub-Saharan African undergraduate STEM students' academic performance.
- To identify the key student involvement factors used in EDM to predict sub-Saharan African undergraduate STEM students' academic performance.
- To identify EDM's most frequently used tools and algorithms to predict sub-Saharan African undergraduate STEM students' academic performance.
- To ascertain the extent of EDM effectiveness in predicting sub-Saharan African undergraduate STEM students' academic performance.

1.5 Significance of Study

This research addressed the fundamentals of student academic performance and student success. The study was motivated by the need to gain a deeper understanding of EDM in predicting undergraduate students' academic performance taught in the disciplines of Science, Technology, Engineering, and Mathematics (STEM) or related disciplines such as statistics, as these students drive the economic expansion in the country. The research covered all students in STEM fields. Still, it focused more on Technology students in Information Systems (IS), Information Technology (IT), Software Engineering (SE), Computer engineering (CE), and

Computer Science (CS) courses (Uzoka *et al.*, 2013). Students in the STEM fields who cannot graduate on time negatively impact economic expansion due to a low throughput of entrepreneurs and skilled employees entering the economy in the 4IR era (Abe & Chikoko, 2020). Educational Data Mining (EDM) can assist HEIs in making informed decisions about their undergraduate students (Agarwal *et al.*, 2012). EDM can also help HEIs identify students who will likely not graduate on time.

There was a lack of detailed research on EDM conducted in developing countries, and this research presented an opportunity to provide a scholarly review in this domain. This research aimed to study the literature on EDM. Moreover, to determine the effectiveness of EDM in predicting students' academic performance by identifying the different factors such as student behaviour within LMS, academic results, and socio-economic and demographic factors. In addition, this study aimed to identify various EDM methods commonly used to predict sub-Saharan African undergraduate STEM students' academic performance. These EDM methods include Decision Trees, Support Vector Machines, k-Nearest Neighbours, Artificial Neural Networks, and Naive Bayes. This research provides a detailed report that uses the SLR process to review the EDM literature. This research encourages researchers to use detailed reporting methodologies for further EDM research in their contexts, as it may reveal new knowledge.

1.6 Contributions of this study

This research contributed to the EDM perspective and dimension of the overall research performed in sub-Saharan African HEIs. Most HEIs have similar problems of students being excluded, de-registering, failing, and dropping out. Adopting EDM may allow researchers in various HEIs to gain more knowledge about their students, including identifying the possibility of failures or dropouts (Jacob *et al.*, 2015). In this way, HEIs can assist students and improve their throughput levels.

1.7 Research Methodology

This study followed a systematic literature review (SLR) process. SLR is a commonly used method for literature review (Saa *et al.*, 2019). Using SLR protocols to guide researchers throughout the review process improves methodological transparency and allows future review replication (Mallett *et al.*, 2012). The most crucial part of SLR studies is identifying and locating all the literature related to the primary research questions of SLR studies (Paez, 2017).

Hence, a comprehensive search for literature through various databases is crucial for SLR studies as it helps prevent missing important literature.

Grey literature is evidence not published in academic and commercial distribution channels that may provide meaningful contributions to an SLR study (Paez, 2017). Grey literature includes research reports, theses and dissertations, policy statements, government reports, issue papers, bulletins and newsletters, and geophysical and geological surveys (Bellefontaine & Lee, 2014). Grey literature was included in this research due to its ability to record findings in emerging or niche research areas and findings that yield negative or null results (Adams *et al.*, 2016). The grey literature found in this research included conference proceedings and theses or dissertations. Eleven studies were theses or dissertations, and four studies were conference proceedings. This grey literature was part of the 41 selected studies included for further analysis that were published in the five years from 2017 to 2021. These studies focused on EDM's effectiveness in predicting STEM students' academic performance in the sub-Saharan African region.

1.8 Delimitations of this Study

This research had some delimitations. The study's scope was restricted to the sub-Saharan African region because there was insufficient EDM literature to conduct an adequate review in South Africa. Additionally, there was a lack of detailed reviews in the sub-Saharan African region; this provided an opportunity to provide a scholarly review in the EDM domain. Moreover, initial searches for studies were conducted using online databases, but this was not enough. Hence, other sources of evidence, by additional searching through various libraries, needed to be done using a revised search string. Additionally, only freely accessible studies were selected for further analysis. This may have resulted in some vital literature being excluded from further analysis because they were not freely accessible.

1.9 Structure of this study

This section covers how this dissertation was structured to achieve the objectives of this study. The following are the chapters covered in this dissertation:

Chapter One

Chapter One provides an introduction, defines the problem, identifies the research questions and objectives, and motivates the need for this study. This chapter summarises the critical points of this research.

Chapter Two

Chapter two presents the related literature on EDM. It provides an overview of existing literature on EDM from the sub-Saharan African perspective and identifies the literature gap. In addition, it also develops a conceptual framework using an improved development theory and the hypothesis formation step in knowledge discovery. It sets the tone for EDM's effectiveness in predicting students' academic performance and the crucial factors integral to the research topic.

Chapter Three

Chapter three is the Research Methodology which presents available research methods, the methods used by the researcher, and the reasoning behind the methodology choice. The PRISMA framework guides the research methodology of this study.

Chapter Four

Chapter four presents the results obtained from the systematic search for the studies related to EDM using the PRISMA framework. It provides the results of both the bibliometric analysis and the thematic analyses using Nvivo.1.7.1. The thematic analysis is broken down into two parts. This chapter covers the first part of the thematic analysis and focuses on academic performance analysis.

Chapter Five

Chapter five is a continuation of Chapter four, which covers the second part of the thematic analysis, and focuses on academic performance prediction. Moreover, it synthesises the two parts of the thematic analysis by addressing the challenges of academic performance analysis and prediction. Additionally, it explains the relationship between HEDM frameworks and thematic analysis and how this relationship allows researchers to interpret the results holistically and develop theoretical frameworks.

Chapter Six

Chapter six discusses the results obtained in Chapters four and five to answer this dissertation's research questions and objectives. Moreover, the discussion chapter comprehensively explores the multifaceted aspects of EDM, ranging from success prediction at the degree level, key input and student involvement factors, to the utilization of EDM tools and evaluation methods, shedding light on the intricacies and challenges of predicting sub-Saharan African undergraduate STEM students' academic performance.

Chapter Seven

Chapter seven provides a summary of all the work done in this research and highlights the key findings of this research in relation to answering the research questions. This also briefly mentions the limitations, threats to validity, future directions, and whether the research had data availability, conflict of interests, and funding.

1.10 Summary

This chapter outlined the background and importance of this research focusing on EDM's effectiveness in predicting students' academic performance. It introduced the research problem, questions, and objectives of this study. The research and analysis methods required for this research have been explored, and the delimitations. The next chapter provides an overview of existing evidence on EDM from the sub-Saharan African perspective to form the theoretical basis for this research.

Chapter 2: Literature review

2.1 Introduction

This chapter aims to synthesise existing literature related to this research, provide a deeper understanding of the topic, and place this research in the existing scholarship. Additionally, this chapter aims to inform and broaden the literature search conducted in phase one of this study which is discussed in the following chapter. The structure of this chapter is as follows: firstly, DM is defined and explored by covering the literature on the DM methods, techniques, tools, and application areas. Secondly, EDM is explored in detail to give the reader an in-depth understanding of the main concepts relevant to this research. Thirdly, EDM as a big data analytic tool in educational systems is explored by discussing the need for methods and tools to handle the enormous educational data. Fourthly, the justification for using EDM in this research is explored by providing academic defence for the choice of EDM over Learning Analytics.

Additionally, the focus of EDM in sub-Saharan Africa and STEM fields is explored, focusing on the factors that drive the need for EDM to be employed. The factors influencing student academic performance are also explored, focusing on the aspects relevant to the EDM community. Moreover, the effectiveness of EDM is examined, focusing on the relevant factors contributing to its efficacy. Furthermore, a conceptual framework is developed, explored, and justified to provide the reader with an in-depth understanding of the crucial constructs in this study. Finally, a conclusion is drawn to highlight how the literature has informed this study.

2.2 Data Mining

Over the last decade, finding meaningful trends from data has been labelled by various names, such as information discovery, pattern processing, information harvesting, data mining, and knowledge extraction (Chuddher, 2015). According to Adekitan and Salau (2019), management information systems, data analysts, and statistics communities commonly use DM. DM is defined as extracting meaningful hidden information from data through scientific methods and analysis to identify hidden patterns and trends inside that dataset (Adekitan & Salau, 2019). Hence, DM can also be defined as knowledge discovery.

The first knowledge discovery in databases (KDD) workshop in 1989 introduced the term knowledge discovery, emphasising that the discovery of knowledge is data-driven (Piatetsky-

Shapiro, 1991). KDD has been popularised over the past years in various fields, including AI and ML (Khedr & El Seddawy, 2015). The KDD process includes these main steps: selecting, cleaning, storing, data mining, and presenting (Sultan *et al.*, 2021). Firstly, various data sources are selected as the targeted data. Secondly, the data is cleaned to remove inconsistent and incomplete records to attain a dataset that is standardized. In addition, aggregate functions are used to transform and consolidate the dataset into appropriate forms for mining. Thirdly, the consolidated dataset is stored in a data warehouse. Fourthly, DM methods are applied to extract the data patterns. Lastly, the results are evaluated and interpreted to reveal the knowledge discovered. KDD is the entire process of meaningful discovery of knowledge from a dataset, and DM is a step in the KDD process (Alyahyan & Düştegör, 2020). The steps in the KDD process, as shown in Figure 2.2, are essential in ensuring that meaningful knowledge is derived from the data.

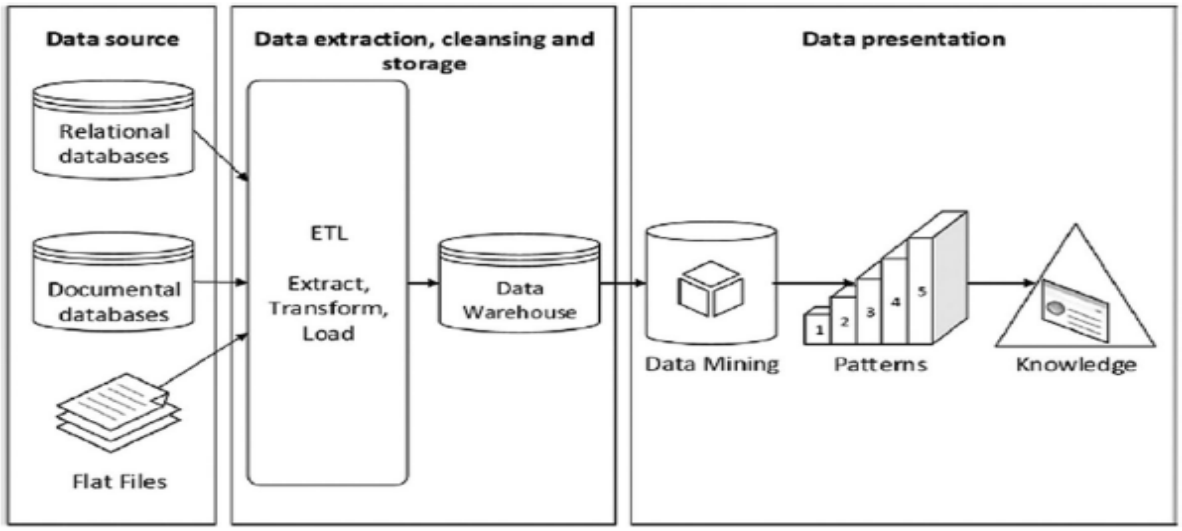

Figure 2.1: KDD process (Alyahyan & Düştegör, 2020, p. 2)

According to Adekitan and Salau (2019), discovering meaningful knowledge from data requires the application of DM models that associate computer science with statistics to mine new valuable knowledge from a jumble of meaningless data.

According to Pratheebha *et al.* (2021), there are five main DM methods: Classification, regression, association rule mining, clustering, outlier detection, and sequential patterns.

Classification

The Classification approach is the most frequently used DM method that employs a set of pre-defined classes that enable the development of models to classify large populations (bin Roslan & Chen, 2022). The classification models used in DM include Decision Trees, Bayesian Networks, Artificial Neural networks, Support vector machines, and Instance-based models. Classification models categorise dependent variables and assign new data to well-defined classes (Christy *et al.*, 2018). According to Naseem (2021), the primary purpose of the classification approach involves training datasets and developing accurate prediction models using the available features in a dataset. The prediction models' performance is measured by comparing expected and actual values in a test dataset. The predicted variables are usually binary or categorical. A prediction model is accepted only when it achieves the desired accuracy, which is then used to make new predictions.

Regression

The Regression approach is a statistical method mainly used for numeric predictions (Pratheebha *et al.*, 2021). Regression models the relationship between a dependent variable and one or more independent variables. In this method, the target values are already known. For example, Smita and Sharma (2014) used regression analysis to predict a child's behaviour with their family history. However, real-world target values are often tough to predict or unknown. For example, stock prices and sales volumes are challenging to predict due to their dependence on complex interactions with multiple predictor values. Hence, it may be necessary to use more sophisticated methods such as the Classification and Regression Trees (CART) model that can classify and predict both categorical and continuous variables.

Clustering

The Clustering approach consists of grouping abstract or physical objects into categories of similar things (Saraswat, 2016). According to Mirza *et al.* (2016), using the clustering approach can assist in determining the correlations and overall distribution patterns among dataset features. For example, groups of customers can be formed using their purchasing habits.

Association Rule Mining

Association rules are used for predicting a single or combination of variables from a dataset (Saraswat, 2016). The association rules can be derived from small and large datasets; hence, reasonable accuracy is critical to determining the preferred rules (Pratheebha *et al.*, 2021).

According to Hashima *et al.* (2018), association rule mining has two concepts defining its essential rules: confidence and support. The support concept determines the number of cases to make a correct prediction. In contrast, the confidence concept determines the number of cases predicted correctly.

Sequential patterns

A sequential pattern is a sequence of variables commonly occurring in a particular order. These items have the same transaction time value (Slimani & Lazzez, 2013). Sequential pattern mining involves examining relationships between sequential events to find sequential events that occur in a certain order (Saraswat, 2016). According to Wang *et al.* (2018), sequential pattern mining aims to identify the maximal sequences satisfying the pre-specified minimum support threshold. The maximal sequence is a critical pattern for identifying the sequential relationships between different item sets. Sequential pattern mining is widely used to analyse DNA sequences.

Outlier detection

According to Pratheebha *et al.* (2021), rare events are more intriguing than those occurring regularly in some applications. Various datasets used in DM may have data that does not conform to the general model (Mirza *et al.*, 2016). These data types are termed outliers. Analysing outlier data is called outlier or anomaly analysis. According to Modi and Oza (2016), outlier analysis is conducted in various datasets, such as graphical, numerical, and text. Identifying outliers can lead to discovering valuable and meaningful knowledge, such as detecting fraud and predicting abnormal values.

The DM methods discussed above fall into two types of purposes and goals: verification and discovery (Saraswat, 2016). Verification consists of the evaluation of a proposed hypothesis by external sources. It also includes traditional statistical techniques like the test of goodness of fit, hypothesis testing, and variance analysis. Discovery methods, on the other hand, are those that identify data patterns automatically. The discovery method branch has two categories: prediction and description (Khedr & El Seddawy, 2015). Prediction methods are oriented towards building behavioural models that can predict new and unseen sample values. In contrast, Description methods are oriented toward interpreting data, which involves understanding the relationships between data in the dataset. Description models can be evaluated along the dimensions of prediction accuracy, novelty, utility, and understandability.

Classification and regression are categorised under the prediction category, while clustering, outlier detection, sequential patterns, and association rule mining are classified under the description category. The predictive and descriptive DM goals or functions with their associated methods are shown in Figure 2.2.

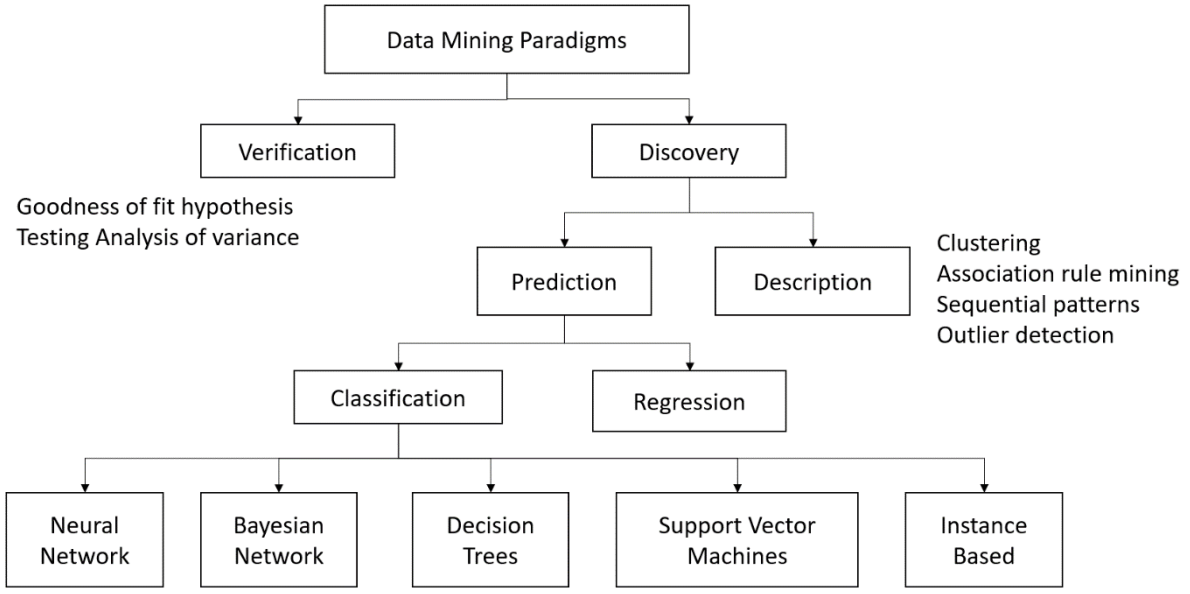

Figure 2.2: Data Mining Paradigms (Saraswat, 2016, p. 3)

According to Ashenafi (2017), various open-source and proprietary tools are available for developing and executing DM models, such as Rapid Miner, IBM SPSS Modeler, Oracle Data Mining, KNIME, Rattle, Python, Kaggle, WEKA, Teradata, and Orange. DM tools are mostly used for predicting patterns, behaviours, and trends, enabling organisations to make effective decisions driven by knowledge (Yadav & Pal, 2012). DM can be a process where the user guides the DM tool and decides the techniques to use. DM can also be an automated process where the user sets the DM tool to run automatically. DM can also be a combination of guided and automatic processes (Coronel & Morris, 2016). DM is a fast-growing area implemented in various sectors, such as customer behaviour, fraud detection, facility maintenance management, engineering, traffic management, genetics, and intrusion detection (Adekitan *et al.*, 2019). Not all sectors are on par regarding their growth, as DM has not been explored yet in some industries.

2.3 Educational Data Mining

In the last decade, interest in identifying and understanding the critical factors influencing students' academic performance prediction in HEIs has increased, particularly using EDM methods to predict these factors (Al Luhaybi, 2021). According to Alyahyan and Düştegör (2020), EDM is interdisciplinary and associated with education, statistics, and computer science, as shown in Figure 2.3. Additionally, EDM is associated with learning analytics, machine learning, and computer-based instruction.

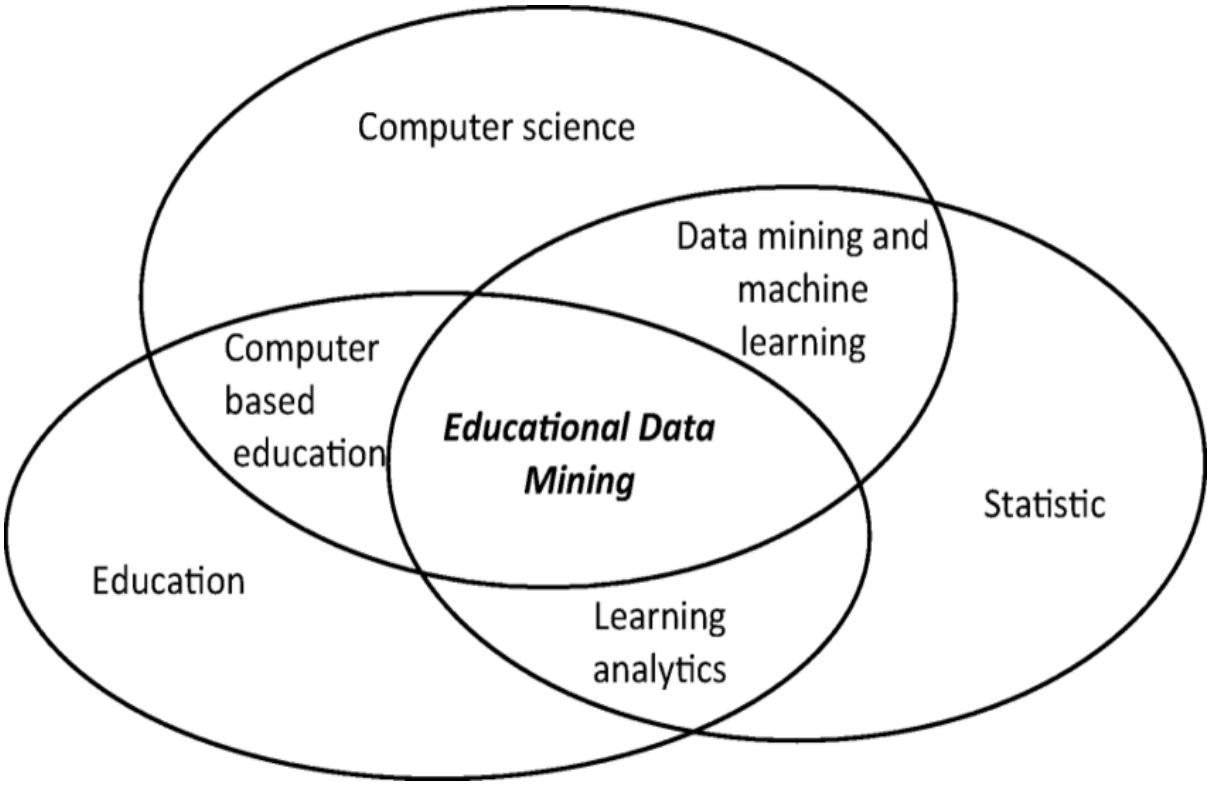

Figure 2.3: Educational data mining-related areas (Romero & Ventura, 2013, p. 17)

The EDM community website (www.educationaldatamining.org) defines EDM as a field associated with developing methods and algorithms for exploring enormous amounts of educational data to better understand students and the contexts where learning occurs (Saa *et al.*, 2019). Unlike DM, the EDM implementation process is iterative, which involves hypothesis formation, testing, and refinement, as presented in Figure 2.4 (Romero & Ventura, 2013).

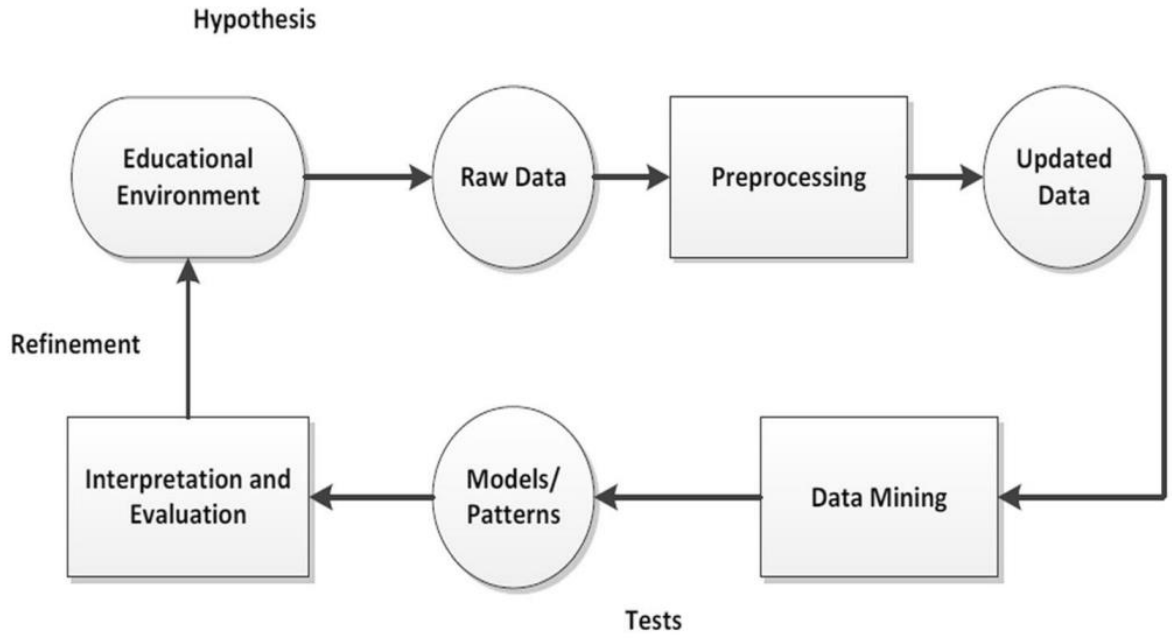

Figure 2.4: Knowledge discovery process in educational institutions (Romero & Ventura, 2013, p. 19)

Hypothesis formation provides a prerequisite hypothesis model based on educational theories, which defines EDM as part of the educational environment. According to Alyahyan and Düştegör (2020), EDM is not a disjointed dataset from the educational entity but a meaningful part of it. Refinement is the feedback that proposes educational decisions and improves or optimises the academic environment. The EDM iterative cycle is a robust model as it enables knowledge discovery and provides meaningful feedback for decision-making to improve the educational environment (Lei *et al.*, 2017).

Related literature on EDM ranges from predicting academic performance, grouping students according to behaviour, and associating the courses for which students enrol (Kumar & Chadha, 2012). The work in EDM has not been limited to only modelling the students' behaviour. EDM is also used to gather historical data in HEIs to find meaningful information patterns which generate new knowledge (Saa *et al.*, 2019). HEIs have also reflected on using historical data for predicting student academic performance. For example, Osmanbegovic and Suljic (2012) looked at monitoring student success and retention rates. In their work, Romero and Ventura (2020) describe how analysing massive amounts of data from educational contexts can assist HEIs in understanding student behaviour better. According to Mashiloane (2016), educators can use findings from EDM research to improve their course materials or programs. EDM is also used to identify patterns in data. EDM seeks to identify students more likely to

graduate or drop out (Yadav *et al.*, 2012). With all the knowledge EDM provides, HEIs can introduce innovative strategies that reduce the likelihood of students failing or dropping out. Table 2.1 provides a comprehensive overview of the areas explored in EDM literature. It uncovers the diverse factors that impact academic performance and contribute to decision-making processes, providing a comprehensive understanding of the educational landscape influenced by EDM.

Table 2.1: Overview of Key Areas in EDM Literature

Area	Overview	References
EDM in Educational Systems	The subsection delves into the application of EDM within educational systems, particularly focusing on online learning systems commonly known as Learning Management Systems (LMS). The discussion explores the EDM process within these systems, highlighting its significance in extracting meaningful insights from educational data.	(Adekitan & Salau, 2019; Chweya *et al.*, 2020; Coronel & Morris, 2016; Du *et al.*, 2020; Hasan *et al.*, 2020; Khedr & El Seddawy, 2015; Li & Xu, 2023; Liang *et al.*, 2016; Lottering *et al.*, 2020; Matsebula & Mnkandla, 2016; Meghji *et al.*, 2018, p. 2; Raghavjee *et al.*, 2021; Ray & Saeed, 2018; Romero & Ventura, 2013; Roy & Singh, 2017; Singh & Pal, 2020; Thakar, 2015; Vambe & Sibanda, 2016; Wanjau & Muketha, 2018; Zohair & Mahmoud, 2019)
Justification for the Focus on EDM over Learning Analytics	The subsection provides a rationale for prioritizing EDM over Learning Analytics (LA). It elucidates the reasons behind choosing EDM as the primary focus, offering insights into its advantages and distinct capabilities in the context of this study.	(bin Roslan & Chen, 2022; Mgala, 2016; Saa *et al.*, 2019; Siemens & Baker, 2012, pp. 252-253)
Focus on the Sub-Saharan African Context	The subsection highlights the unique characteristics and challenges of the sub-Saharan African educational context. It explores why this region serves as a specific focus within the study, shedding light on the factors that differentiate it from other educational landscapes.	(Bawack & Kamdjoug, 2020; Maphosa & Maphosa, 2020; Olaya *et al.*, 2020; Wanjau & Muketha, 2018)
Focus on Science, Technology, Engineering, and Mathematics	The subsection narrows the focus to STEM disciplines. It examines how EDM is applied and relevant within the domains of Science, Technology, Engineering, and Mathematics, providing insights into the specific dynamics of these fields.	(Abe & Chikoko, 2020; Aulck *et al.*, 2017; Bengesai & Pocock, 2021; bin Roslan & Chen, 2022; Darvas *et al.*, 2017; Khan & Ghosh, 2021; Naseem, 2021; Singh & Alhulail, 2022; Wanjau & Muketha, 2018)
Factors Examined in EDM Literature	The subsection reviews various factors that are subjects of examination within the EDM	(Al-Shabandar *et al.*, 2017; Alom & Courtney, 2018; Bokana & Tewari, 2014; Bussu *et al.*, 2019;

	literature. These factors include student behaviour within LMS, academic results, and socio-economic and demographic characteristics. This comprehensive analysis forms a foundational understanding of the variables considered in EDM research.	Cantabella *et al.*, 2019; Davidson, 2019; Hussain *et al.*, 2018; Iqbal *et al.*, 2017; Jacob *et al.*, 2015; Ndou *et al.*, 2020; Nguyen *et al.*, 2021; Patil *et al.*, 2017; Prabowo *et al.*, 2021; Saa *et al.*, 2019; Sorour *et al.*, 2014, p. 1; Taodzera *et al.*, 2017; Zacharis, 2015; Zollanvari *et al.*, 2017)
Effectiveness of EDM	The use of EDM in decision-making takes centre stage in this subsection. It discusses how, by leveraging qualitative and quantitative data, EDM effectively informs educational decisions. This effectiveness is explored through practical applications and case studies.	(Adekitan & Salau, 2019; Dutt *et al.*, 2017; Khedr & El Seddawy, 2015; Kim *et al.*, 2018; Lei *et al.*, 2017; Mgala, 2016; Romero & Ventura, 2020)
Effectiveness Metrics of EDM	The subsection shifts to examining the metrics used to gauge the effectiveness of EDM. Metrics such as academic performance, decision support, dropout and retention rates, and recommendations are dissected, offering a comprehensive understanding of how EDM's impact is quantified and measured.	(Adak *et al.*, 2016; Adejo & Connolly, 2017; Alban & Mauricio, 2019; Caruth, 2018; Chacon *et al.*, 2012; Chang *et al.*, 2016; Daramola *et al.*, 2014; Haverila *et al.*, 2020; Hegazi & Abugroon, 2016; Huebner, 2013; Iqbal *et al.*, 2017; Jembere *et al.*, 2017; Kabakchieva, 2012; Khan *et al.*, 2021; Makombe & Lall, 2020; Mayilvaganan & Kalpanadevi, 2014; Meghji *et al.*, 2018; Mengash, 2020, p. 55462; Mgala & Mbogho, 2015; Miguéis *et al.*, 2018; Nguyen *et al.*, 2021; Pratt *et al.*, 2019; Ramesh *et al.*, 2013, p. 38; Romero & Ventura, 2013; Sani *et al.*, 2020; Shahiri & Husain, 2015; Shuqfa & Harous, 2019; Vo & Nguyen, 2012)

2.3.1 EDM in Educational Systems

Technological advancements and the expanded use of online learning platforms reduce learning costs, as students and educators are not bound to a particular location (Hasan *et al.*, 2020). Online learning systems have various names, including Learning Management System (LMS), Learning Support System (LSS), Learning Platform (LP), Course Management System (CMS), Learning Content Management System (LCMS), and Managed Learning Environment (MLE) (Khedr & El Seddawy, 2015).

Online learning systems are generally termed LMS. Businesses, schools, and HEIs increasingly install LMS to complement traditional course programs (Romero & Ventura, 2013). The most generally used LMS open-source is Modular Object-Oriented Developmental Learning Environment (Moodle) (Li & Xu, 2023). It allows for creating powerful, engaging, flexible online experiences (Chweya *et al.*, 2020). LMS collects massive amounts of data useful for analysing student behaviour (Raghavjee *et al.*, 2021). Figure 2.5 presents different forms of educational data in LMS.

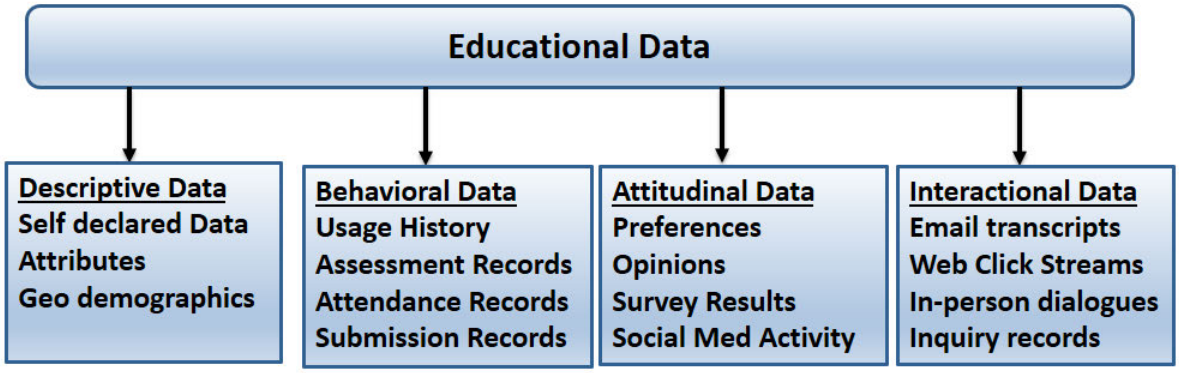

Figure 2.5: Different Forms of Educational Data in LMS (Meghji *et al.*, 2018, p. 2)

HEIs are generating large amounts of data from their LMS and applications. According to Wanjau and Muketha (2018), this data is not always processed and analysed to produce information and knowledge. Adekitan and Salau (2019) argue that one of the shortcomings of conventional data processing applications is that they cannot process data that is not expressed in quantitative terms. Data analytics refers to the collection of statistical, mathematical, and modelling methods used on data to extract knowledge (Coronel & Morris, 2016). Therefore, Big data analytics refers to the collection of methods and tools used to analyse large complex data sets that traditional applications cannot process from which actionable information may be obtained.

According to Ray and Saeed (2018), EDM can be used to analyse enormous educational datasets and can be conducted through two methods: EDM and LA. These methods enable the development of quantitative research to respond to the expanding need for evidence-based analysis regarding educational policies and practices (Adekitan & Salau, 2019). EDM allows HEIs to discover factors influencing students' academic performance and proactively address such factors, which will improve student academic performance, lower dropout rates, increase throughput, and improve institutional effectiveness (Du *et al.*, 2020).

EDM generally starts with collecting student data using learning management systems, log records servers, surveys, social media, and self-assessment modules (Romero & Ventura, 2013). Once data is collected, it is stored in a repository that can store huge volumes of data (Khedr & El Seddawy, 2015). The stored data needs to be cleansed and transformed before it can be loaded on EDM methods (Meghji *et al.*, 2018). These EDM methods include Classification, clustering, text mining, and association rule mining, as mentioned in section 2.1. EDM offers an excellent opportunity to retrieve hidden meaningful knowledge from the student data, as presented by the EDM workflow in Figure 2.6.

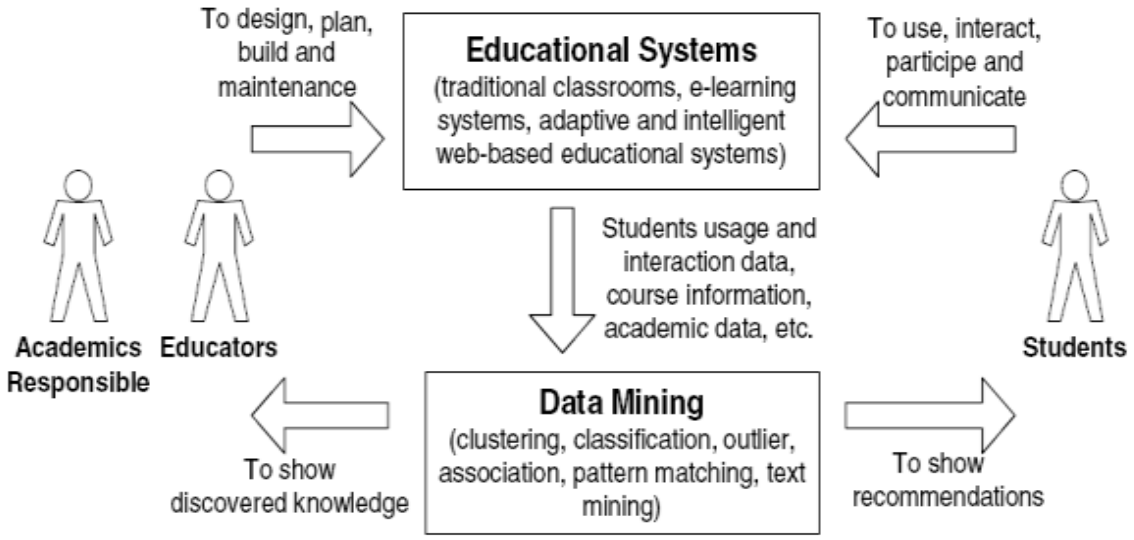

Figure 2.6: EDM Workflow (Khedr & El Seddawy, 2015, p. 25)

EDM can assist HEIs in understanding the relationship between students and educators. It can predict the potential for negative student behaviours (Adekitan & Salau, 2019) and the potential for dropout and students' academic performance (Vambe & Sibanda, 2016). The effectiveness and quality of decision-making can be significantly improved using EDM. Likewise, essential feedback from students can be evaluated using EDM to discover areas of need and lapses to improve the educational environment (Hasan *et al.*, 2020).

The most frequently used EDM approach is classification, which is based on a set of pre-defined classes that enable the development of models that can classify large data records (Singh & Pal, 2020). Classification classifies data based on class labels and training sets to predict categorical values (Matsebula & Mnkandla, 2016). Unusual behaviours may occur when generating data from EDM, which creates anomalous data. The detection of anomalies or deviations is called Outlier analysis. Outlier analysis discovers abnormal patterns in data or deviations regarding the expected behaviour (Liang *et al.*, 2016).

Clustering is frequently used for machine learning, image analysis, bioinformatics, and pattern recognition (Thakar, 2015). In clustering, objects from the data are assigned into clusters of similar things (Zohair & Mahmoud, 2019). Association analysis involves discovering patterns, causal structures, correlations, or associations in enormous datasets (Roy & Singh, 2017). According to Lottering *et al.* (2020), Classification is the most suitable method for classifying students' academic performance into pre-defined classes. In contrast, association and clustering are not suitable in such cases.

2.3.2 Justification for the Focus on EDM over Learning Analytics

Two fields that focus on analysing educational data have been identified to understand students and the environment where their learning occurs. EDM and LA are these fields. They are defined as cited in the literature:

According to Siemens and Baker (2012, p. 252), "EDM involves the development of methods that explore enormous educational datasets to understand students better and the contexts where learning occurs." Meanwhile, "LA involves collecting, measuring, analysing, and reporting student data and the context that learning occurs to understand and improve the educational environment" (Siemens & Baker, 2012, pp. 252-253).

The two fields focus on similar issues: understanding students better and the contexts in which their learning occurs. However, this study chose EDM based on similar motivations as Siemens and Baker (2012) provided. Firstly, EDM involves the use of advanced statistical and machine-learning techniques to discover hidden patterns and relationships in educational data. It goes beyond basic descriptive analytics and identifies intricate trends and correlations within large datasets. Meanwhile, LA typically involves more basic, descriptive analytics focusing on summarizing and visualizing data to provide insights into learning behaviours and outcomes. When the goal is to uncover hidden patterns and gain a deeper understanding of complex relationships within educational data, EDM's advanced analytical techniques make it a more suitable choice. EDM allows for sophisticated modelling, clustering, and prediction, enabling HEIs to extract valuable insights that may not be apparent through simpler analytical methods (Mgala, 2016).

Secondly, one of the strengths of EDM is its ability to build predictive models. These models can predict students' future academic performance, identify at-risk students, and recommend personalised interventions based on historical data patterns (bin Roslan & Chen, 2022). While LA can provide valuable insights into past and current learning behaviours, its focus on

descriptive analytics limits its capacity for predicting future trends or student outcomes. Hence, when aiming to proactively address challenges and enhance educational outcomes by predicting future trends, EDM's predictive modelling capabilities offer a distinct advantage. This is particularly crucial for HEIs seeking to implement early intervention strategies and personalized learning approaches.

Thirdly, EDM often involves analyzing granular data at the individual student level (Saa *et al.*, 2019). This allows for a more detailed examination of learning behaviours, preferences, and challenges. Meanwhile, LA may aggregate data at a higher level, providing insights into overall class performance or trends but potentially overlooking individualized learning needs. When the objective is to tailor educational experiences to individual students and address their specific needs, EDM's capability to analyze granular data on an individual level is essential. This level of granularity enables the identification of personalized learning paths and targeted interventions (Mgala, 2016).

In summary, while LA offers valuable insights through descriptive analytics, EDM provides a more powerful and robust approach by incorporating advanced statistical and machine-learning techniques. When the goal is to uncover complex patterns, predict future outcomes, and enable personalized learning experiences, EDM is a more justified choice.

2.3.3 Focus on the sub-Saharan African Context

The focus on the sub-Saharan African context in this section aims to highlight the need to address the specific challenges STEM students face in the region. Sub-Saharan Africa comprises diverse countries with varying socio-economic conditions, cultural practices, and educational systems (Olaya *et al.*, 2020). Within this complexity, STEM education faces unique challenges, including insufficient infrastructure, limited access to quality education, and socio-economic factors that affect educational opportunities. These challenges create a distinct educational environment that demands targeted research and interventions.

Understanding the impact of EDM within the sub-Saharan African context is crucial for several reasons. Firstly, the socio-economic factors and cultural differences prevalent in the region may influence how students engage with educational materials and respond to interventions (Wanjau & Muketha, 2018). Therefore, tailoring EDM strategies to these specific circumstances is essential for their effectiveness. Secondly, the potential insights derived from EDM can inform targeted interventions that address the specific needs of STEM students in Sub-Saharan Africa, ultimately contributing to improved graduation rates and academic

outcomes (Maphosa & Maphosa, 2020). By focusing on this context, the research seeks to provide valuable insights that can be applied to enhance educational outcomes in similar global contexts.

Despite the evident importance of focusing on sub-Saharan Africa, there is a noticeable scarcity of comprehensive research in this specific context, particularly concerning EDM in STEM education (Bawack & Kamdjoug, 2020). This scarcity underscores the urgency and significance of the current research endeavour. To bridge this gap, a systematic literature review on EDM's effectiveness in predicting sub-Saharan African undergraduate STEM student academic performance was conducted. A systematic literature review aligns seamlessly with the outlined challenges, the importance of the sub-Saharan African context, and the scarcity of existing research. It enables a comprehensive assessment of challenges, identifying patterns and trends across diverse studies (Maphosa & Maphosa, 2020). Importantly, a systematic literature review facilitates tailoring EDM strategies to the sub-Saharan African context by synthesizing evidence on how socio-economic and cultural factors influence student engagement. Furthermore, it informs targeted interventions by providing insights into effective strategies in similar contexts, potentially contributing to improved graduation rates and academic outcomes.

In summary, the focus on the sub-Saharan African context in EDM research is justified by the region's unique challenges, the relevance of understanding EDM's impact, and the scarcity of existing research. The systematic literature review on EDM's effectiveness in predicting academic performance among sub-Saharan African undergraduate STEM students plays a pivotal role in addressing this gap, offering a consolidated understanding of the challenges, tailoring interventions to the specific context, and contributing valuable insights to the broader field of EDM and global education strategies.

2.3.4 Focus on Science, Technology, Engineering, and Mathematics

Enrolment in sub-Saharan African HEIs, particularly in STEM disciplines, has grown faster than on any other continent over the past decade; however, fewer than two out of ten students graduate (Darvas *et al.*, 2017). Moreover, Bengesai and Pocock (2021) found that out of 1370 registered STEM students surveyed, 50% were enrolled in STEM programs, namely Engineering. Around 10% of students had completed their registered engineering program, 40% had persisted, and 50% had dropped out. Wanjau and Muketha (2018) argued that regardless of the interest in STEM fields, enrollment and throughput rates in mathematics and

science are low compared to non-STEM areas. The enrolment of students in science courses is comparatively low globally, and this has caused a shortage of skills in the STEM industry (Aulck *et al.*, 2017).

According to Naseem (2021), only one out of the three sub-Saharan African students enrol in STEM courses, while the remaining students opt for other non-STEM programs. Previous literature states that high student dropouts are caused by factors such as prior educational knowledge, financial background, and demographics (Singh & Alhulail, 2022). sub-Saharan African students' challenges influence their decision not to choose STEM courses (Wanjau & Muketha, 2018). These challenges are categorised into school, home, and community. Within the school environment, where there may be limited resources and a shortage of qualified teachers, EDM can play a crucial role in identifying patterns and trends in students' academic performance, helping to tailor educational strategies to address specific weaknesses and challenges (Abe & Chikoko, 2020). For instance, EDM methods can be employed to analyze academic performance data and identify areas where additional resources or targeted interventions are needed. Moreover, EDM can assist in understanding the impact of negative perceptions surrounding the difficulty of STEM subjects, offering insights into the effectiveness of different teaching methods and interventions in changing these perceptions.

Within the home environment, where parental influence and gender stereotypes play a role, EDM can contribute by providing data-driven insights into the factors influencing students' educational decisions (Khan & Ghosh, 2021). Analyzing data on course selection patterns and performance based on gender and parental involvement can help identify areas where targeted interventions, such as awareness campaigns or mentorship programs, are needed to challenge stereotypes and encourage diverse participation in STEM fields. Additionally, community factors, including the absence of visible role models and limited access to STEM-related activities, can be addressed through EDM-driven initiatives (bin Roslan & Chen, 2022). By analyzing data on community engagement and the impact of outreach programs, EDM can guide the development of effective strategies to promote STEM education, creating an environment that supports and encourages students to pursue STEM courses.

In the context of this systematic literature review on EDM's effectiveness in predicting academic performance among sub-Saharan African undergraduate STEM students, EDM emerges as a diagnostic tool and a proactive solution. By synthesizing existing evidence on the effectiveness of EDM interventions within the unique challenges of Sub-Saharan Africa, the

review provides valuable insights into the potential of data-driven strategies to predict and enhance STEM students' academic performance. The systematic review acts as a foundational resource, informing future EDM initiatives and contributing to a comprehensive understanding of how EDM can be optimized to address the specific needs of STEM students in Sub-Saharan Africa.

2.3.5 Factors Examined in EDM Literature

Educational Data Mining (EDM) provides higher educational institutions (HEIs) with an opportunity to discover unique relationships and patterns about students that could help study, predict, and improve students' academic performance (Jacob *et al.*, 2015). The most common variables or factors considered in EDM research include student behaviour within LMS, academic results, and socio-economic and demographic characteristics (Cantabella *et al.*, 2019).

Student Behaviour Within Learning Management Systems (LMS)

The standard at present in HEIs consists of including an LMS as an essential methodological tool that generates data that aids in developing prediction models to discover valuable knowledge (Cantabella *et al.*, 2019). Many studies have been conducted using EDM with LMS data. Al-Shabandar *et al.* (2017) found that student academic success strongly correlated with their click-stream LMS behaviours and demographics. Moreover, Zacharis (2015) found that the frequency of solving quizzes, viewing files, and reading and posting messages were significant predictors of student success.

In addition, Sorour *et al.* (2014, p. 1) found that the students' LMS comments after lessons predicted student success with an accuracy of 82.6%. Furthermore, Hussain *et al.* (2018) analysed students' learning behaviour using student demographic data, LMS activity data, and end-of-course assessment survey data. The study found that students' LMS behaviour correlates significantly with academic success. Therefore, using different forms of data to analyse student behaviour is more meaningful and helps to identify new relationships (Cantabella *et al.*, 2019). These studies have proven that using EDM with data from LMS and surveys benefits HEIs.

Academic Results

Using the academic results obtained by students in the past to predict future outcomes and discover patterns or relationships in big sets of institutional data is perhaps the most expected variable or factor used in EDM. Many would assume that students who previously achieved low results are more likely to attain low marks in the future, but this may not always be true (Nguyen *et al.*, 2021). EDM provides the opportunity to simultaneously consider past academic results and other factors or variables that may have influenced the results – resulting in more accurate predictions (Saa *et al.*, 2019). For instance, HEIs can use EDM to predict students' academic performance, allowing HEIs to identify students with poor grades. Thus, HEIs can help them achieve better academic performance through various intervention programs, improving overall students' academic performance (Zollanvari *et al.*, 2017).

According to Prabowo *et al.* (2021), the American Grade Point Average (GPA) is frequently used to measure students' academic performance. Students' GPA is predicted using time series or tabular data. In most literature, tabular data is the EDM model input (Davidson, 2019). Tabular data includes student background data, including educational programs registered for, year of enrollment, admission type, achievement test scores, and academic performance in high school. Meanwhile, the time-series data refers to student academic grades, including students' GPAs, which were reported biannually throughout their active study period. For example, Patil *et al.* (2017) used students' previous semester performance to predict their Cumulative Grade Point Average (CGPA). Furthermore, their findings suggested that combining historical GPA and tabular data was useful (Iqbal *et al.*, 2017).

Demographic and Socio-economic Factors

Demographic and socio-economic factors are among the most common factors in the literature surrounding EDM. Many EDM studies have considered factors such as ethnicity, gender, and age. Alom and Courtney (2018) studied the influence of gender on Australian students' success and found that gender plays a critical role in success rates, specifically in certain states. Taodzera *et al.* (2017) found that school province and ethnicity played an important role in predicting student success in engineering at a South African university.

Literature in EDM has also focused on socio-economic factors. According to Ndou *et al.* (2020), socio-economic factors include parents' education, occupation, and family income but are not limited to these factors. Additionally, Bokana and Tewari (2014) also found socio-economic and other factors, such as institutional environment, intellectual leadership, learning

environment, and psychological attitudes, to be significant predictors of students' success. In another study by Bussu *et al.* (2019), family characteristics and financial support were reported as important factors explaining why students drop out of HEIs. Socio-economic factors can improve a model's accuracy or efficiency. However, these factors do not provide a straightforward measure of the student's potential, making predicting students' academic performance a multifaceted and more complex process (Ndou *et al.*, 2020). Hence researchers need to be precise in selecting these factors as they affect the accuracy or efficiency of their models.

2.3.6 Effectiveness of EDM

HEIs have massive amounts of educational data that can be collected for decision-making, including academic scores, record data on students' activities, course syllabi, and administrative documents (Khedr & El Seddawy, 2015). This educational data makes decision-making data-driven (Mgala, 2016). The educational data may be helpful for HEIs or policymakers in responding to the demands of students and teachers to improve teaching and learning (Adekitan & Salau, 2019). Students may use self-evaluation data to govern their study habits (Kim *et al.*, 2018). Meanwhile, teachers may use assessment data to improve their lecturing styles and intervention strategies (Dutt *et al.*, 2017).

Educational data provides no interpretation. Hence, it cannot be used directly for decision-making (Mgala, 2016). Usually, the analysis and interpretation of educational data are conducted through various methods such as EDM or LA. Depending on the decision-making required, these methods may need to be adapted (Lei *et al.*, 2017). Hence, using EDM improves the decisions made by HEIs as it provides new knowledge that is easily understandable for improving decision-making. From the EDM research perspective, there are two general approaches: qualitative and quantitative. The EDM process consists of the quantitative approach that shows mining outcomes used for qualitative analysis and interpretation (Romero & Ventura, 2020). The quantitative approach consists of the mined outcomes from the educational data in a numeric format, which may not be helpful or easily understood by HEIs stakeholders. The qualitative approach consists of the interpreted quantitative data in a meaningful and easily understood format for HEI stakeholders. The qualitative analysis enriches the quantitative access of EDM.

2.3.7 Effectiveness metrics of EDM

Various private and government funds depend on the success rates of HEIs (Makombe & Lall, 2020). Hence, HEIs are adopting common practices such as EDM to improve students' success rates and increase institutional effectiveness to attain those funds and stay in business. Several factors contribute to measuring EDM's effectiveness in HEIs, such as academic performance, decision support, dropout and retention rates, and recommendations.

Academic Performance

Various definitions have been given to academic performance in the literature. Academic performance is linked with the achievement and graduation of students. Academic performance measures students' ability to graduate from current or future courses (Meghji *et al.*, 2018). This often relates to whether the student is high-performing or struggling (Kabakchieva, 2012). High-performing students generally perform well in their previous and current courses. On the contrary, struggling students usually perform poorly their entire academic career.

Moreover, students' grades also are of interest to course coordinators. Previous literature has used EDM with students' previous academic data to provide recommendations in the subsequent phases of their studies. Mayilvaganan and Kalpanadevi (2014) assessed various classification methods, including Naïve Bayes, C4.5 Decision Trees, and k-Nearest Neighbours for students' academic performance prediction. They discovered that the most efficient method was C4.5 Decision Trees. Shahiri and Husain (2015) reviewed EDM methods, including Support Vector Machines, Decision Trees, Naïve Bayes, k-Nearest Neighbours, and Artificial Neural Networks for predicting students' academic performance. Their findings revealed that CGPA is a commonly used factor, with the Decision Trees model being the most effective.

In contrast, Miguéis *et al.* (2018) showed that the Random Forests were more efficient in predicting students' academic performance with first-year performances. Artificial Neural Networks are also effective in predicting students' academic performance in some studies. Mengash (2020, p. 55462) predicted performance using a dataset that included secondary school GPA and admission scores. The findings revealed that Artificial Neural Networks were the most effective solution, accurately predicting 79% of students' academic performance. Furthermore, Ramesh *et al.* (2013, p. 38) used EDM to predict various features influencing students' academic performance, such as personality, psychological, and socio-economic factors. Their findings showed an accuracy of 72.38% using multilayer perceptron models. Applying various factors such as psychological, personality, and socio-economic factors can

effectively predict academic performance, allowing HEIs to support students and improve teaching and learning.

EDM for Decision-making Support

EDM includes analysing educational processes such as admissions, course selections, and alumni relations. Moreover, applying specific EDM methods, including classification, association rule mining, multivariate statistics, and web mining, is crucial for predicting and forecasting students' academic performance and institutional improvement needs (Chacon *et al.*, 2012). EDM methods can model students' individual differences and provide measures to respond to them, improving students' academic performance (Hegazi & Abugroon, 2016). Literature notes that a solid data warehousing strategy is crucial for EDM's success. Huebner (2013) states that having readily available information for decision-makers is crucial in HEIs. HEIs need a reliable data warehousing strategy for EDM to be successful or effective (Romero & Ventura, 2013). Data warehouse strategies are created to increase competitiveness and improve reporting quality to external stakeholders such as parents, community leaders, and legislators (Hegazi & Abugroon, 2016).

Vo and Nguyen (2012) proposed an educational decision-support system that is knowledge-driven, which was found to be useful for educators, students, and HEIs. In addition, Mgala and Mbogho (2015) state that decision support that is knowledge-driven is useful for educational stakeholders in making effective decisions to improve students' academic performance. Hence, the decision-making process is more meaningful for educational stakeholders when relevant data or knowledge is available.

Dropout and Retention Rates

One of the critical challenges in HEIs is student retention. Retention rates are an important benchmarking tool in HEIs (Caruth, 2018). According to Haverila *et al.* (2020), HEIs strive to increase or maintain the number of students enrolled by improving the quality of programs offered to enhance students' academic performance and prevent them from dropping out of their studies. Dropouts negatively affect HEIs by reducing student retention rates and hindering institutional effectiveness (Alban & Mauricio, 2019).

Many empirical studies investigate the determining factors leading students to drop out. The literature identifies financial concerns due to low or costly educational funding as a significant factor that leads to low retention rates in HEIs (Pratt *et al.*, 2019). Furthermore, a key indicator

of HEIs failure to complete their mission is low retention rates (Shuqfa & Harous, 2019). Hence, the failure of HEIs to prepare young adults for various professions leads to massive economic development losses as fewer professionals are produced from HEIs (Sani *et al.*, 2020).

A student with slow progress is identified as an at-risk student (Shuqfa & Harous, 2019). According to Khan *et al.* (2021), predicting students' likelihood of dropping out brings many advantages to HEIs and students. Additionally, forecasting students' academic performance successfully assists HEIs in identifying the characteristics of outstanding students. These features, including students' attendance and engagement levels, can be set as the main principles for every student enrolling in HEIs to identify at-risk students early and provide strategic interventions for improving students' academic performance. The goal is to improve performance, increase retention, reduce attrition, and increase graduation rates (Adejo & Connolly, 2017).

Recommendations

For students to obtain desirable outcomes, it is essential that they choose suitable courses (Nguyen *et al.*, 2021). Various factors, such as interest, demands, and competence, influence students' decisions on course enrollment. Recommendations on the appropriate programs through the assessed student grades assists them in making the best choices in their studies. Hence, this has resulted in researchers focusing on developing EDM recommendation systems to assist students in choosing suitable courses. Chang *et al.* (2016) collected student data through surveys and successfully predicted students' academic performance using the collaborative filtering method to recommend suitable programs.

Jembere *et al.* (2017) discovered that the Matrix Factorization method effectively predicts students' academic performance and recommends relevant courses. Iqbal *et al.* (2017) found that combining matrix factorisation and collaborative filtering methods, such as Restricted Boltzmann Machines provided better recommendations. Adak *et al.* (2016) predicted students enrolled in compulsory programs to recommend relevant programs in the following semester using EDM methods such as Fuzzy logic and Decision Trees. Daramola *et al.* (2014) implemented and tested EDM on a recommendation system that suggests appropriate programs based on students' previous academic performances. They found that 93% of the results were suitable for automated recommendations. Therefore, the quality of the suggestions or feedback contributes to the effectiveness of EDM.

2.4 Developing a Conceptual Framework

This study explores various factors influencing EDM's effectiveness in predicting students' academic performance in HEIs, including student behaviour within Learning Management Systems (LMS), academic results, and socio-economic and demographic factors. In addition, it explored EDM methods influencing EDM's effectiveness in predicting students' academic performance, including Decision Trees, Artificial Neural Networks, Regression, k-Nearest Neighbour, Support Vector Machines, and Naïve Bayes. This was done to discover EDM's effectiveness in predicting sub-Saharan African undergraduate STEM students' academic performance and the existing frameworks applicable to such a topic. This section explores the various theoretical or conceptual frameworks developed and used in the EDM domain, such as the knowledge discovery in a database process model, the cross-industry standard process for data mining, and the hybrid data mining model. This study also covers frameworks for student development in higher educational environments, such as the Input-Environment-Outcomes model, the student involvement model, and the student development model. These frameworks are covered in this study because they attempt to explain the interaction between the students and the educational environment and how this interaction influences students' academic performance. This interaction also influences the prediction of students' academic performance.

Knowledge discovery in databases (KDD) process model

The focus of KDD is the knowledge discovery process that includes accessing and storing data, scaling DM models to massive datasets to predict effectively, and interpreting and visualising data. According to Alyahyan and Düştegör (2020), KDD finds understandable patterns that can be construed as meaningful knowledge. It also focuses on DM models' robustness and scaling properties for large datasets. Figure 2.7 shows the steps of the KDD process.

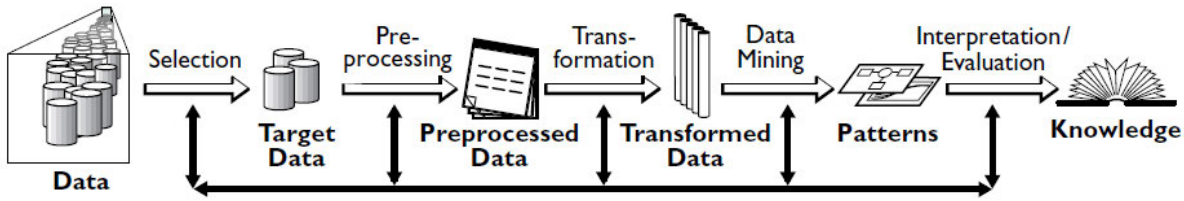

Figure 2.7: KDD Process model (Fayyad *et al.*, 1996, p. 29)

The KDD process is organised into six phases, as outlined in the work of Fayyad *et al.* (1996): selecting a target dataset, preprocessing, transforming data, mining, interpreting, and discovering knowledge.

The cross-industry standard process for data mining (CRISP-DM) model

In EDM research, the CRISP-DM model is a commonly used framework (Pressman, 2015). It provides clear guidance about the different phases of DM and how they should be conducted (Schröer *et al.*, 2021). Figure 2.8 shows the stages of CRISP-DM as outlined in the work of Chapman *et al.* (1999). The phases are business understanding, data understanding, data preparation, modeling, evaluation, and deployment. These phases are a blueprint to follow when planning and conducting DM projects.

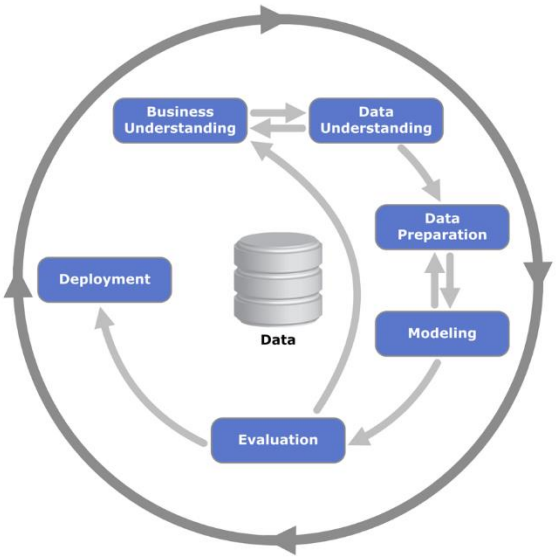

Figure 2.8: Cross-industry standard process for data mining (CRISP-DM) model (Yaacob *et al.*, 2020, p. 4)

Hybrid Data Mining process model

The academic models, such as KDD, and industrial models, such as CRISP-DM, developed by EDM researchers have created new hybrid models (Girma, 2019). Cios and Kurgan (2005) developed the hybrid DM process model based on the CRISP-DM and KDD models. Figure 2.9 presents the hybrid process model's six steps. The phases include understanding the problem domain, understanding the data, preparing the data, mining the data, evaluating knowledge discovered, and using the found information.

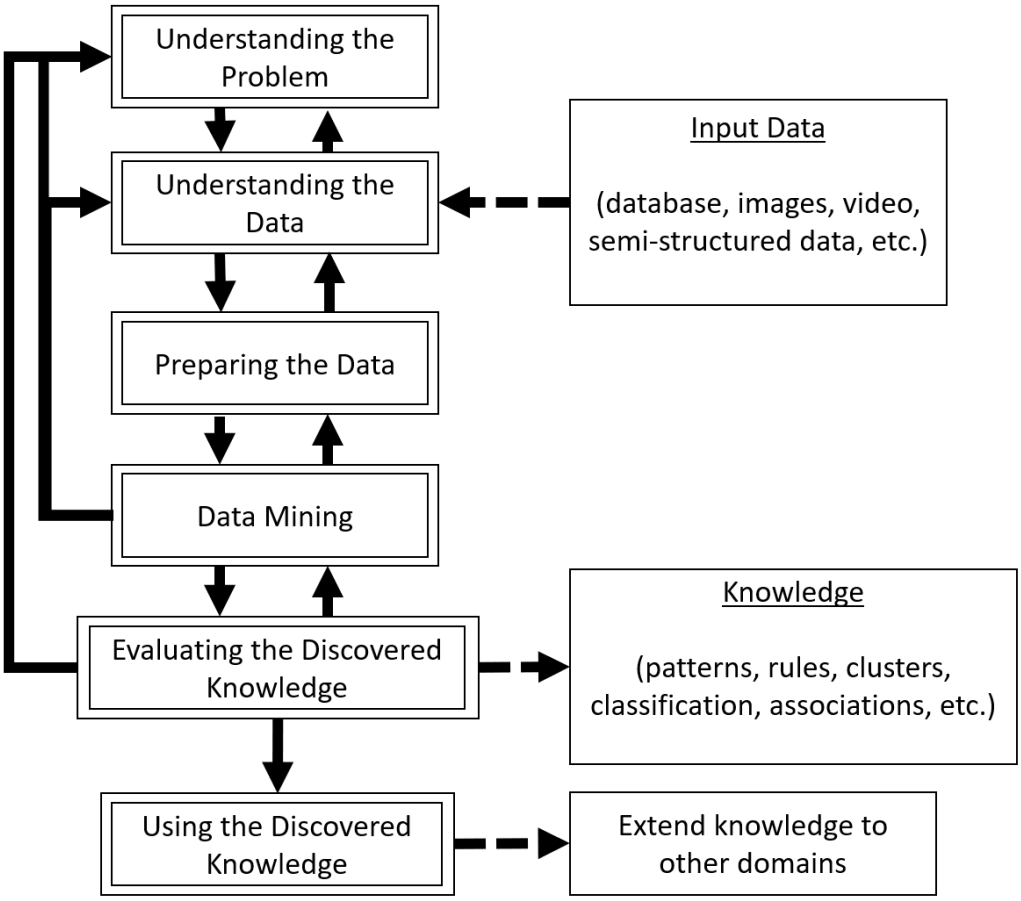

Figure 2.9: Hybrid DM process model (Cios & Kurgan, 2005, p. 6)

In contrast to Data Mining (DM) frameworks such as the KDD process model, CRISP-DM model, and the Hybrid DM model, this study extends its exploration to include two established models that specifically describe students' development in higher educational environments. The two established models include the Student Involvement Theory (Astin, 1984) and the Inputs-Environment-Outcomes model (Astin, 2012).

DM frameworks excel in providing a structured and systematic approach to data analysis, facilitating tasks such as data preprocessing, feature selection, and model evaluation. These DM frameworks are particularly robust in handling large datasets and automating the discovery of patterns and trends. In contrast, while not focused on data-driven processes, the Student Involvement Theory and the Inputs-Environment-Outcomes model contribute valuable qualitative and quantitative insights into students' interactions with their educational environment.

The Student Involvement Theory, proposed by Astin (1984), emphasises the importance of engagement and participation in educational activities, providing a qualitative lens to understand students' experiences. The Inputs-Environment-Outcomes model, proposed by Astin (2012), contributes to a comprehensive understanding of the factors influencing student outcomes, offering quantitative rigour to studying student development. Integrating these established developmental models into the study's framework aims to create a more comprehensive approach, leveraging the strengths of DM frameworks in systematic data analysis while benefitting from the insights provided by well-established theories of student development in higher education (Lei et al., 2017).

Student Involvement Theory

Most conventional academic theories classify students as a black box, meaning that the focus is on the inputs and outputs but not the internal mechanics or details of how these inputs are transformed into outputs (McCollum, 2018). The basis of the student involvement theory is that students must not be considered a type of Black Box. The student involvement theory exposes some details or internal mechanics within the supposed Black Box. Student development is intertwined with student involvement, resources, and individual abilities, which the student involvement theory links from the students' perspective (Lei *et al.*, 2017). Figure 2.10 represents the five postulates in the student involvement theory proposed by Astin (1984).

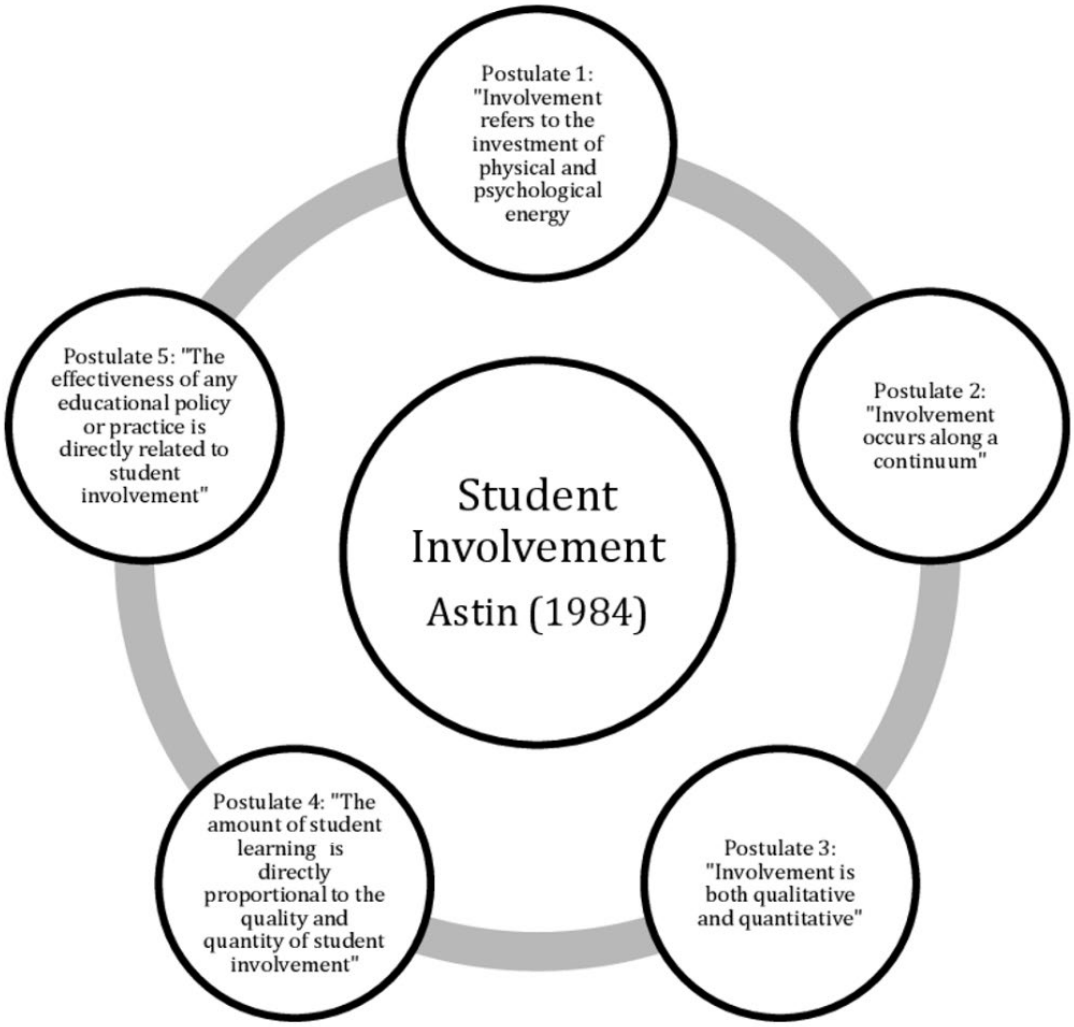

Figure 2.10: Theory of Involvement postulates

Student involvement is the quantity and quality of physical and psychological energy students invest in educational experiences. The energy students invest varies from student to student; hence, student involvement is a continuous variable. Involvement has qualitative and quantitative features. The skills or knowledge that students gain from their involvement is strongly correlated with the extent of their participation. Various types of involvement influence students' development, including place of residence, student governance, athletic involvement, academic involvement, and student-faculty interaction (Jaafar *et al.*, 2012). Some evidence shows that the effectiveness of educational courses or programs for students encourages student involvement and improves future development and student success (Lei *et al.*, 2017).

Leadership and policymakers in HEIs can govern these involvement factors to improve retention and the various outcomes (Kim, 2017). While this theory emphasizes the significance of student involvement, it progressively transforms student development from Black Box into

White Box (Lei *et al.*, 2017). According to McCollum (2018), this theory still needs more evidence and application to reveal its truths. It also needs to focus more on the variables' inputs, outcomes, and interactions with the educational environment.

Inputs-Environment-Outcomes Model

Astin (2012) states that from the HEIs perspective, a course or program is considered a model that includes three attributes: student inputs, educational environment, and outcomes, as shown in Figure 2.11. The first variable, inputs, refers to the personal characteristics that a student initially brings to a program or course, such as race, gender, level of competency at the time of entry, educational experiences, family background, and personal expectations of future courses (Bard, 2016).

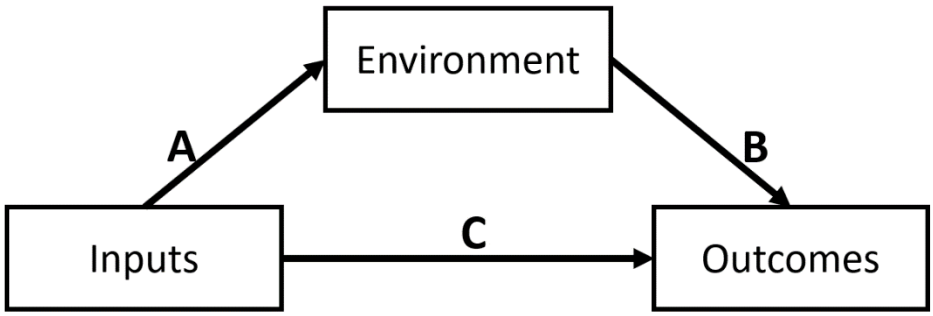

Figure 2.11: Inputs-Environment-Outcomes Model

The second variable is the environment, which consists of students' actual experiences during a program (Lei *et al.*, 2017). It focuses on recording students' development activities in an educational environment. The third variable is outcomes, which include school retention, career orientation, incremental knowledge, and graduation throughput. Outcomes refer to the abilities that HEIs are attempting to develop throughout the current course or program (Chigbu & Nekhwevha, 2021). Outcomes can be defined by time outcome and outcome type. Outcomes variables are influenced by both the educational environment and the inputs. Inputs inform students' attitudes and choices in their educational environment and significantly shape their academic outcomes (Patton *et al.*, 2019).

Improved Student Development Model

The improved student development model combines the student involvement theory and the Inputs-Environment-Outcomes model, as shown in Figure 2.12. The enhanced student development model is strengthened by some advantages of both models, which appeal to this study for two main reasons, as outlined in Lei *et al.* (2017).

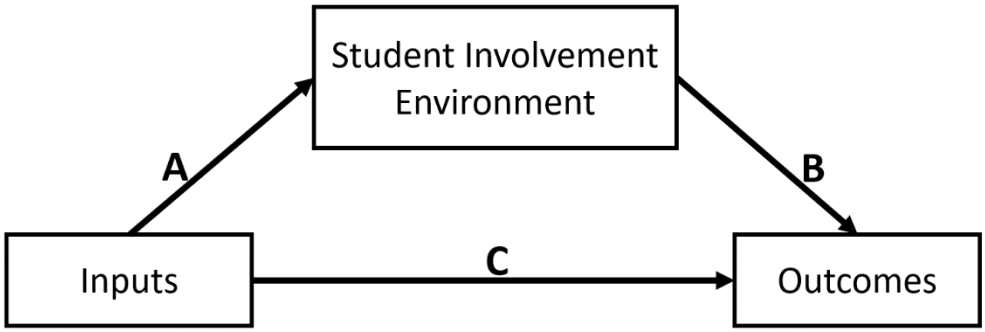

Figure 2.12: Improved Student Development Model

Firstly, the subjectivity of students in the educational involvement environment is respected better by the improved model. Secondly, the enhanced model focuses on static document data in HEIs, such as exam papers and course content, and dynamic data, such as student interactions with peers and faculty. Furthermore, the enhanced student development model is inclined to a quantitative study from the academic research perspective.

Although various frameworks have been developed in the DM and academic domains, no existing theoretical framework can be directly applied to achieve this study's objectives. Therefore, a conceptual framework was created for this study by determining relevant variables from the literature and incorporating them into this study's conceptual framework.

2.5 Conceptual Framework

The student development theory and the hypothesis formation phase in the knowledge-discovery model underpin the conceptual framework of this research. Hence, this research developed a conceptual framework to understand EDM's effectiveness in forecasting students' academic performance. The Higher Educational Data Mining (HEDM) framework includes three major parts: inputs, student development environment, and outcomes, as shown in Figure 2.13.

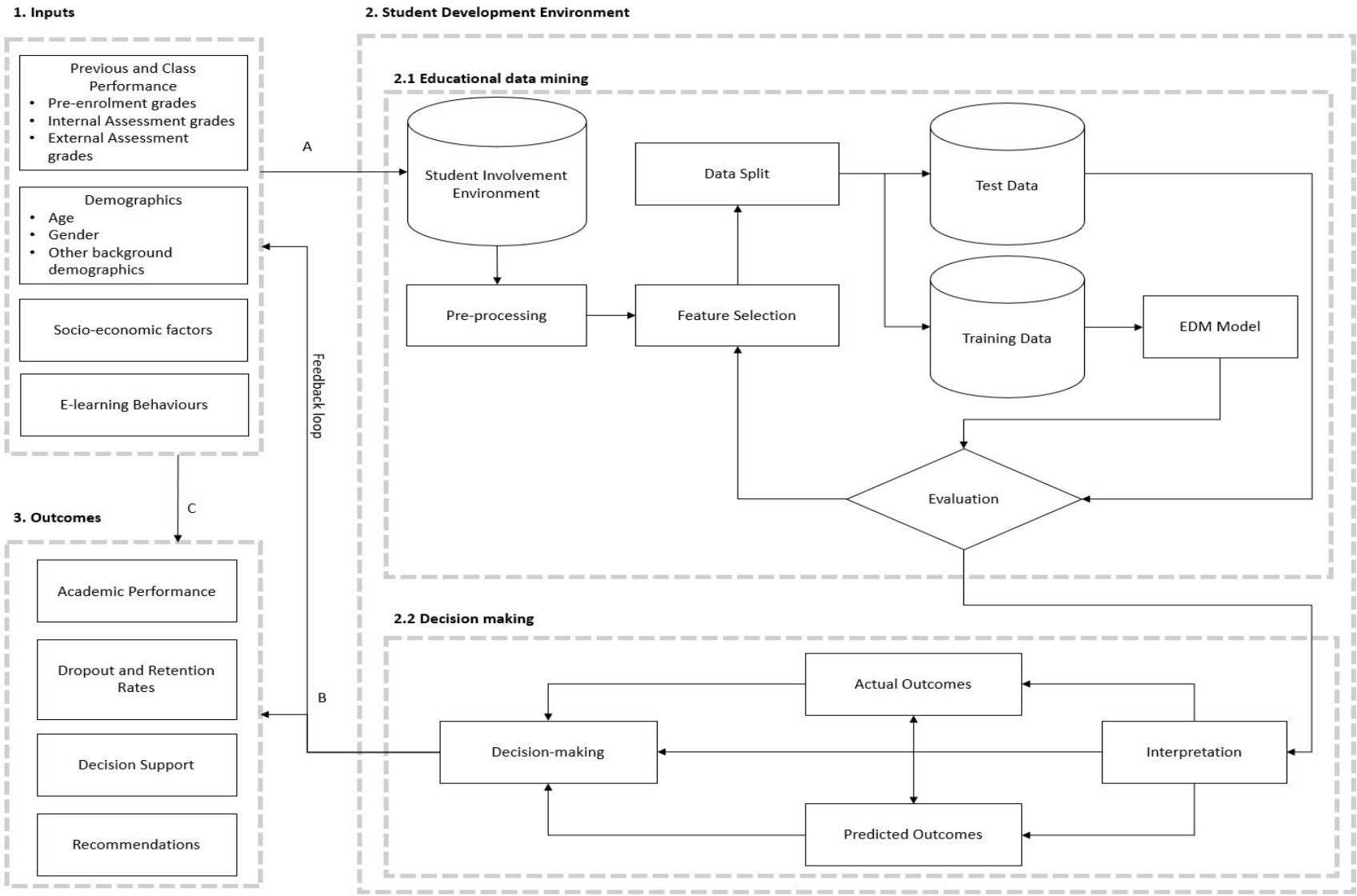

Figure 2.13: HEDM framework for decision making

Inputs

In the HEDM model, the improved student development theory is the basis for higher educational hypothesis formation. Students' inputs (relationship A) influence the students' involvement environment (relationship B), and relationships A and B both impact students' outcomes (relationship C) directly. In other words, input and involvement data are independent variables, and are indirectly influenced by only educational decisions (Astin, 2012). Outcomes are a dependent variable influenced by input and involvement data. Outcomes data is used to determine the impact of decision-making and to evaluate the quality of the student development environment (Lei *et al.*, 2017). Therefore, all three variables have effects on each other.

Inputs include students' demographics, background, and any previous experiences. The environment consists of students' experiences and involvement during an educational program. The student development environment includes LMS and SIS. Lastly, outcomes are comprised of students' attitudes, beliefs, values, and knowledge after completing a program.

Student Development Environment

The student development environment is where the student interacts with teachers. It contains two meaningful parts, which are EDM and decision-making.

Educational Data Mining: According to the HEDM model, EDM is not disjointed from an educational entity but is a meaningful part of it. Multiple and complex sources exist in academic environments (Khedr & El Seddawy, 2015). In many cases, the data obtained from educational settings can be structured and unstructured data formats that are complicated to process. Thus, data from the student involvement environment must be pre-processed or cleaned before it is used for EDM. Converting data to a suitable form for solving particular student development problems is vital (Lei *et al.*, 2017). The quality of student data directly affects accuracy in predictive models. There are various methods to clean data, but they are not a single approach. Still, cleaning methods must be selected to suit the data's actual problem domain to help manage abnormal, duplicate, and missing values (Alyahyan & Düştegör, 2020).

Once data is cleaned, feature selection is applied to educational data. Feature selection consists of choosing the best available attributes based on certain evaluation criteria (Xiao *et al.*, 2021). Redundant and irrelevant attributes that have a low impact on outputs can be eliminated to increase the performance of EDM models. In EDM, datasets are commonly separated into two segments: training and testing. The training part is used for model fitting. At the same time, the

testing dataset evaluates the performance of the constructed model from the trained data. In most cases, data is split into 70% for training EDM models, while 30% is used for model testing, where the accuracy of models is calculated (Alwarthan *et al.*, 2022).

Next, predictive or descriptive algorithms are selected and used in building accurate models that transform input data into anticipated outcomes (Khedr & El Seddawy, 2015). Predictive algorithms are supervised DM methods that yield inferences from a student dataset for future predictions. In contrast, descriptive algorithms are unsupervised DM methods that use data in a dataset to find unknown knowledge and present it to relevant stakeholders. If the evaluation metrics are unmet, the evaluation stage produces feedback for stakeholders to adjust the selected features to improve model performance. The accuracy, recall, precision, ROC area, and F-score are some of the evaluation metrics used in EDM (Alwarthan *et al.*, 2022). If the evaluation criteria are met, a model is built and used to predict the outcomes from student data. It is crucial to interpret the data mining results and present educational suggestions. From the EDM perspective, the interpretation step transforms from quantitative results to qualitative suggestions (Lei *et al.*, 2017).

Decision Making: Decision-making involves the interpretation step, a transformation process from quantitative data to qualitative recommendations. According to Lei *et al.* (2017), stakeholders may use the transformed data to achieve various objectives depending on the demand levels influencing the decision-making. Lecturers may use the performance data to reflect on their intervention strategies and teaching practices (Romero & Ventura, 2013). In contrast, students may use the transformed data to boost academic performance and manage study habits (Khedr & El Seddawy, 2015). Administrators or policymakers may also use the transformed data to respond to the demands of students and teachers, provide timely interventions, and improve education quality (Alwarthan *et al.*, 2022).

The accuracy of the predicted outcome data towards the actual outcome data is compared and shows EDM's effectiveness in predicting students' academic performance and decision-making. The final decisions and outcomes may be used for improving the student development system; this is the feedback loop of the HEDM model. The outcome data are stored in the student development environment, which is critical in the following decision-making series. The student development environment, including EDM, becomes more robust and intelligent as this conceptual framework is an iterative process.

Outcomes

Outcomes reflect the quality or effectiveness of decision-making, as shown by relationship B. Moreover, as relationship C demonstrates, they are also directly influenced by input data and the student involvement environment. The outcomes from decision-making consist of academic performance, decision support, dropout and retention rates, and recommendations provided. Academic performance measures the ability of a student to graduate from current or future courses (Meghji *et al.*, 2018). Thus, EDM uses students' historical and current academic performance data to provide recommendations for the following phases of their courses.

Recommendations on appropriate programs through the assessed student grades will assist them in choosing more suitable courses throughout their program (Nguyen *et al.*, 2021). Therefore, decision support driven by knowledge is useful for educational stakeholders in making more reasonable and appropriate decisions (Mgala & Mbogho, 2015). Additionally, successfully forecasting students' academic performance can assist educators in identifying the aspects of outstanding students (Khan & Ghosh, 2021). These attributes can be the main principles assigned to students enrolling in HEIs. The aim is to improve performance, increase retention, reduce attrition, and increase the throughput rates (Adejo & Connolly, 2017).

2.6 Summary

This chapter gave the reader an understanding of the use of DM, which was then defined and explored, including explaining its application areas. The relevant topics associated with EDM, such as EDM as a Big data analytic tool in educational systems and the focal area in this research, were explored. Additionally, this literature review analysed EDM and Learning Analytics (LA), covering differences and similarities between the two research areas to justify the choice of EDM for this research. Additionally, the focus of EDM in sub-Saharan Africa and STEM fields was explored. Moreover, this literature review also identified the factors that influence students' academic performance and are relevant to the EDM community. Furthermore, the effectiveness of EDM was also explored, focusing on the relevant factors that contribute to its efficiency.

Additionally, a conceptual framework was developed for this study due to the lack of an existing and relevant applicable framework. The HEDM conceptual framework was a combination of the improved student development theory and hypothesis formation in DM. This literature review noted a lack of detailed reviews in the field of EDM in sub-Saharan

African HEIs, with only a few published studies on using and implementing EDM in sub-Saharan Africa and developing countries. This suggested a gap in the literature that needed further exploring to establish the potential benefits for HEIs and other stakeholders.

This literature review explored factors such as student behaviour within Learning Management Systems (LMS), academic results, and socio-economic and demographic factors. Additionally, the literature review explored various EDM methods, including Decision Trees, Support Vector Machines, Artificial Neural Networks, k-Nearest Neighbours, and Naive Bayes used to predict sub-Saharan African undergraduate STEM students' academic performance. This study aimed to provide a systematic literature review on EDM's effectiveness in predicting sub-Saharan African undergraduate STEM student academic performance. To create the prospect for researchers to use detailed reporting methodologies for further EDM research in sub-Saharan Africa and other developing countries, as it may reveal new knowledge for those contexts

Chapter 3: Research Methodology

3.1 Introduction

This chapter describes the chosen methodology for this study. The research methodology selected reflects how the objectives of this research are achieved. This research was planned and designed to answer this study's research questions by collecting viable research data. The research design informed the data collection method used in this study.

3.2 Research Design

Research design is a strategy guiding a study (Sekaran & Bougie, 2016). The research design is integral to a dissertation or thesis. It ensures consistency between chosen methods, techniques, and underlying philosophy.

3.2.1 Research Philosophy

A research philosophy consists of assumptions or beliefs about knowledge development (Saunders *et al.*, 2015). These beliefs or assumptions include but are not limited to axiological, epistemological, and ontological assumptions. Ontological assumptions refer to the nature of the reality being investigated in one's research (Alharahsheh & Pius, 2020). Epistemological assumptions refer to how knowledge of this reality is acquired and communicated to other human beings (Alharahsheh & Pius, 2020; Mauthner, 2020). Axiological assumptions refer to how one's values influence the research process (Saunders *et al.*, 2015). According to Ryan (2018), coherent beliefs or assumptions lead to a plausible research philosophy underpinning the chosen methodology and data analysis process. Hence, a cohesive set of beliefs or assumptions is essential, leading to plausible research questions and the chosen methodology.

According to Saunders *et al.* (2015), there are five significant research philosophies: critical realism, postmodernism, pragmatism, interpretivism, and positivism. Positivism takes on a natural scientist standpoint, which involves working with the observable reality within society and producing generalizations (Al Riyami, 2015). Positivism focuses on the facts without the influence of human bias when interpreting results (Alharahsheh & Pius, 2020; Pham, 2018). Interpretivism differs from positivism as it focuses on experienced events and interpreting those events rather than providing generalizations about them (Alharahsheh & Pius, 2020). Interpretivism is a subjectivist philosophy that views people as detached from physical events (Saunders *et al.*, 2015).

Critical realism is a philosophical branch that differentiates between the observable and natural worlds (Pawson, 2013; Singh, 2019). The natural world exists independently from human perceptions, constructions, and theories and cannot be observed (Gray, 2016). On the other hand, the observable world comprises experiences and perspectives from everything observable. Critical realism states that unobservable compositions cause visible experiences; hence, if researchers examine these compositions generating visual experiences, it will lead to them understanding the social world (Gray, 2016; Singh, 2019). Therefore, the researcher's task is to develop a reliable or accurate domain-specific account of the patterns from observed experiences.

Postmodernism rejects the possibility of one having objective knowledge (Kroeze, 2012). This philosophy focuses on communities' subjective values rather than a single researcher's authoritative voice (Farhan, 2019). It focuses on questioning what has been accepted to reveal hidden explanations, contradictions, inconsistencies, oppositions, and dominations (Creswell & Poth, 2016). Pragmatism is based on the notion that philosophical debates on the nature of reality can be explained through research to understand the challenges of the real world (Kelly & Cordeiro, 2020). A research onion was developed by Saunders *et al.* (2015) focused on the research philosophies mentioned previously. The research onion helps organize the research and design by following each step of its layers. The research layers resemble an onion and peel away to culminate at the centre, as shown in Figure 3.1.

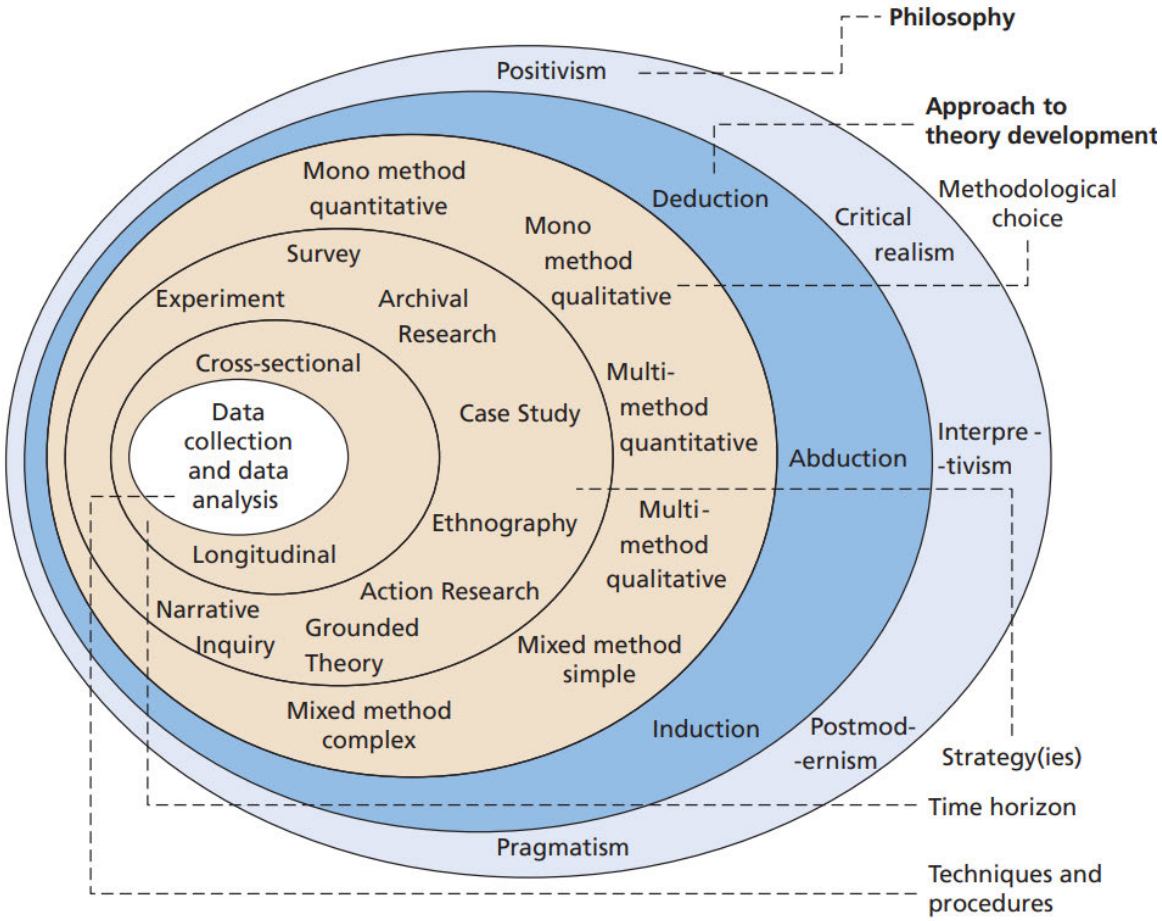

Figure 3.1: Research Onion (Saunders *et al.*, 2015, p. 6)

This study used the critical realism philosophy from the first layer. Critical realism enables the assessment of the domain where the research is conducted and the involved researchers' voices (Gray, 2016). The researcher was interested in teasing out the crucial opportunities and challenges for further analysis. According to Durning and Carline (2015), an inductive study consists of shifting from specific data regarding certain phenomena to an abstract or general conceptualization. The researcher discussed EDM and then developed the HEDM conceptual framework in section 2.5. Hence, this research followed an induction approach from the second layer.

A monomethod study utilizes only qualitative or quantitative methods (Molina-Azorin, 2012). The qualitative method was chosen from the third layer as an inductive approach underpinned this research. The researcher collected secondary data for this research mainly in textual forms and analysed thematically using qualitative data analysis techniques and tools, namely, NVivo. According to Khan (2014), the qualitative research approach involves conceptual thinking and theory building as it explores the possible antecedents that little is known about. The choice

from the second and third layers led to the grounded theory aspect being selected from the fourth layer. According to Charmaz (2014), grounded theory involves collecting data and developing analytic categories or codes from data, not from logically deduced hypotheses. This theory makes comparisons in each stage of the analysis and advances theory development (Creswell & Poth, 2016). Thus, researchers develop a grounded theory inductively from data.

Cross-sectional studies are a research design where a researcher simultaneously collects data from many individuals (Saunders *et al.*, 2015). In contrast, in longitudinal studies, a researcher repeatedly collects data from the same subjects over time, mainly concentrating on a subset of individuals with similar characteristics (Setia, 2016). These approaches help answer different kinds of research questions. However, cross-sectional studies are more straightforward when initially collecting data and identifying patterns that can be examined further in longitudinal studies (Page *et al.*, 2016). Therefore, this study chose the cross-sectional approach from the fifth layer. The researcher planned to select studies published in the last five years that relate to the effectiveness of EDM in predicting undergraduate sub-Saharan African STEM students' academic performance. These five layers discussed above end with the centre being the data collection aspect and data analysis techniques which stem from selecting the research method that is discussed next.

3.2.2 Research Method

Empirical studies are frequently undertaken to explore various phenomena in EDM (Imtiaz *et al.*, 2013). However, as every investigation is limited in scope, researchers must rigorously collect, analyse and interpret the results from all related empirical studies regarding the domain of interest to present a meaningful general overview of the existing literature (Niazi, 2015). The evidence-based paradigm has been applied in several disciplines, including clinical medicine, information systems, education, and social policy (Brereton *et al.*, 2007). This paradigm encourages researchers to use a systematic literature review approach to objectively evaluate and synthesize empirical results about the research questions and integrate the results into practice (Verner *et al.*, 2014).

A critique mostly posed to EDM researchers relates to the little or no usage of the available experiences and methods from other disciplines (Niazi, 2015). Therefore, this study implements a systematic literature review (SLR) process in the EDM domain to examine the EDM's effectiveness in predicting sub-Saharan African STEM students' academic performance. In addition, this SLR seeks to identify areas of improvement of current EDM

infrastructure and practices. SLR is a commonly used method for literature review (Saa *et al.*, 2019). Researchers use SLR as a guide throughout the review process, which expands the comprehensibility of the methodology and allows future repetition of an SLR study (Albreiki *et al.*, 2021).

SLR was appropriate for this study since there were numerous EDM-related primary studies. Quantifying available research results across articles was necessary to generate the maximum possible evidence for the research questions in section 1.3. SLR comprises three main phases that are planning, conducting, and reporting. To ensure clarity and quality of the methodology. Each step of SLR has sub-steps that must be performed when conducting this type of study, as shown in Figure 3.2.

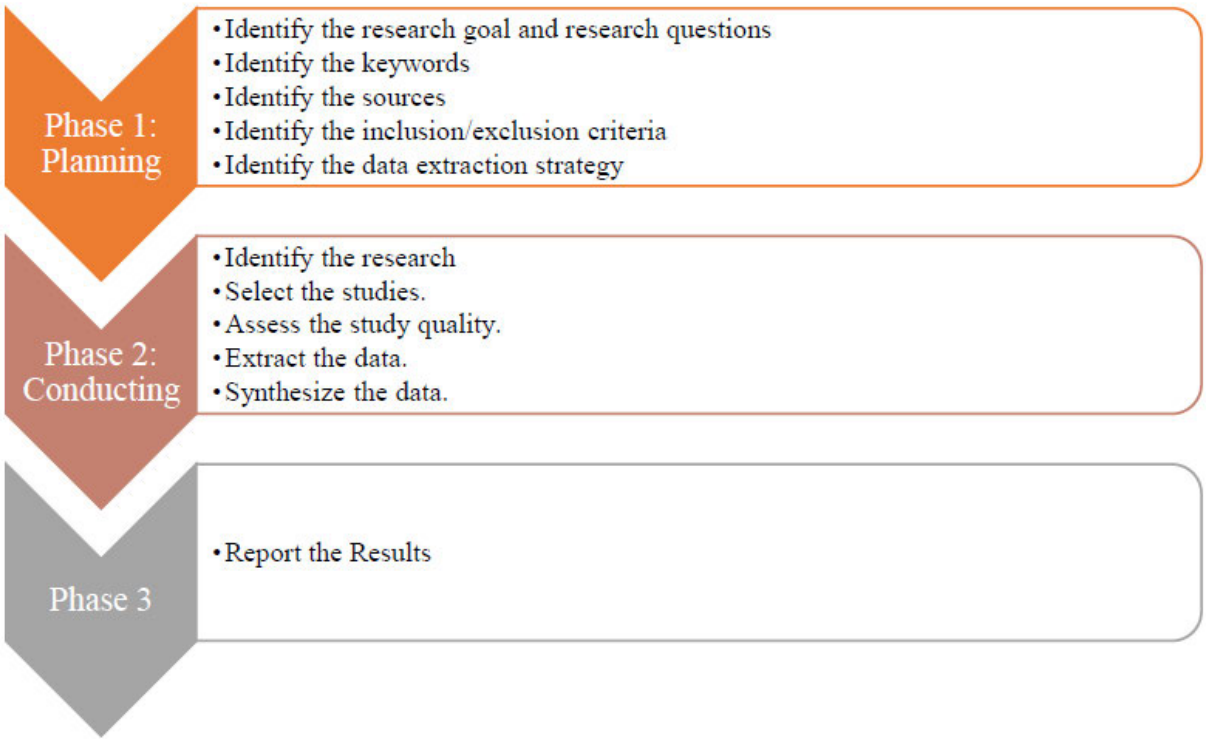

Figure 3.2: SLR process (Saa *et al.*, 2019, p. 8)

Identifying and locating all the literature related to the main research question is a vital part and challenge in SLR studies (Paez, 2017). Hence, a transparent search for literature through various databases is crucial for SLR studies not to miss important literature. The information retrieval or literature search process not only informs the SLR findings but also defines the available data for the synthesis (Rethlefsen *et al.*, 2021). This ensures that the SLR is both robust and reproducible and minimizes bias (Moher, 2018). According to Moher (2018),

current reporting methods help researchers in achieving a rich study. Various guidelines exist for conducting literature searches and reporting findings in SLR (Lefebvre *et al.*, 2019).

The most commonly used reporting guideline for SLRs is the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA), which covers SLRs' literature search and reporting components (Rethlefsen *et al.*, 2021). Moher et al. (2009) state that the PRISMA framework addresses poor SLR reporting. It consists of twenty-seven recommended items for SLR reporting. In addition, Moher *et al.* (2009) provide further reporting advice on all the items and examples of reporting in their work. Therefore, the PRISMA framework outlined in the work of Moher *et al.* (2009) guides the reporting of this SLR. The researcher followed the recommended checklist of 27 items for reporting, as shown in Table A.1 in Appendix B, to ensure a clear and transparent research methodology.

3.3 Phase 1: Planning

This section explores the steps taken in conducting this research, including identifying the keywords, databases, eligibility criteria, and data extraction strategy.

3.3.1 Registration and Protocol

Registering a detailed research plan before data collection is crucial to achieving a transparent SLR. The primary or secondary objectives and methods are established in advance, and readers are informed about changes to the process (Stewart *et al.*, 2015). Since this study followed a secondary data collection approach, the analysis was document-based. The published articles were freely available from various online databases, such as Google Scholar, Scopus, ACM, etc. The researcher did not have to request access or pay for some published full-text articles. Moreover, there were no respondents, but the researcher was still required to acquire ethical clearance for the study before collecting data. This was done through the UKZN Research Office. The researcher requested ethical clearance from the Humanities and Social Sciences Research Ethics Committee (HSSREC) as this was a non-medical study, as shown in Appendix A.

3.3.2 Search Strategy

The search strategy describes all the information sources included in the literature search. Peer-reviewed articles are found in databases for research, such as ACM, IEEE Xplore, Taylor & Francis Online, Google Scholar, Springer Link, ProQuest, Scopus, EBSCOhost, Emerald Insight, Web of Science, and ScienceDirect. This study selected three commonly used

databases: Scopus (https://www-scopus-com.ukzn.idm.oclc.org/), Google Scholar (https://scholar.google.co.za/), and ProQuest (https://www.proquest.com/). These databases are widely used in EDM reviews and involve studies investigating the predictive modelling of STEM student outcomes (Namoun & Alshanqiti, 2021). In addition, the free accessibility to the databases influenced the choice of the selected databases. Moreover, these databases provided access to grey literature as well (Magini *et al.*, 2022, p. 1).

Grey literature is evidence that has not been commercially published in academic distribution and publishing channels, which may provide meaningful contributions to an SLR study (Paez, 2017). According to Bellefontaine and Lee (2014), grey literature includes research reports, theses and dissertations, policy statements, government reports, issue papers, bulletins and newsletters, and geophysical and geological surveys. Grey literature may reduce publication bias, increase the timeliness and comprehensiveness of reviews, and provide a clear view of available studies (Mahood *et al.*, 2014). The diverse audiences and formats of grey literature can offer several challenges when searching for literature for SLRs (Bellefontaine & Lee, 2014). However, the benefits of using grey literature are more significant than the time, cost, and resources required to search for grey literature (Paez, 2017).

Grey literature is crucial and must be included in an SLR study. A grey literature search strategy that is carefully thought out may be an invaluable part of an SLR study (Paez, 2017). Furthermore, the comprehensiveness of an SLRs literature search is increased by including grey literature in the search strategy (Adams *et al.*, 2017). Additionally, it may enhance the SLR's results and lower the possible bias of publications. In the EDM domain, no minimum number of studies is required to conduct a practical SLR study (Okoli & Schabram, 2010). Hence, this study aimed to consider at least 40 studies for selection as a general number of selected studies in SLR studies in the EDM domain (Rodríguez-Triana *et al.*, 2017). The keywords used to identify literature are:

"educational data mining," "machine learning," "learning analytics," "academic analytics," "effectiveness of educational data mining," "predicting students' performance," "factors affecting students' performance" "higher educational institution," "university," "tertiary," "STEM courses," "classification," "clustering," "sub-Saharan Africa," "developing countries," and "Africa." Connectors of keywords searched, such as AND, OR, and truncation (*), are utilized to broaden or reduce the results. The following search string should form an initial

search string that is used to search for the literature, as shown in Table 3.1. It was limited from 2017 to 2021 and sorted by relevance for all databases searched.

Effectiveness AND (“data mining” OR “educational data mining” OR “learning analytics”) AND (“classification” OR “clustering”) AND (“predicting student academic performance”) AND (“science” OR “technology” OR “engineering” OR “math*”) AND (“course*” OR “program*”) AND (“undergraduate” OR “university”) AND (“Africa*” OR “Sub-Saharan Africa*” OR “developing countr*”)

Table 3.1: Search terms for SLR

Connector	Research string
	Effectiveness
AND	(“data mining” OR “educational data mining” OR “learning analytics”)
AND	(“classification” OR “clustering”)
AND	(“predicting student academic performance”)
AND	(“science” OR “technology” OR “engineering” OR “math*”)
AND	(“science” OR “technology” OR “engineering” OR “math*”)
AND	(“Africa*” OR “Sub-Saharan Africa*” OR “developing countr*”)

3.3.3 Eligibility Criteria

Tricco *et al.* (2018) state that eligibility criteria are the features used to assess the suitability of a study in the selection process. Table 3.2 shows this study’s eligibility criteria involving inclusion and exclusion conditions. This research considered primary research studies for inclusion published in the five years, from 2017 to 2021, as any study before 2017 might have outdated information due to rapid changes in EDM (Chanthiran *et al.*, 2022). Primary research studies collect data directly from the source rather than depending on data collected from previously done research (Driscoll, 2011). This study included literature on EDM’s effectiveness in predicting undergraduate STEM students’ academic performance in sub-Saharan Africa.

Table 3.2: Eligibility Criteria

Inclusion Criteria	Exclusion Criteria
Should satisfy the conditions of the identified keywords	Not a Primary research study
Must be classified as STEM or related discipline EDM research	Not in higher education or tertiary level
Must be published between 2017 and 2021	Not conducted in sub-Saharan Africa
Must be conducted in higher educational institutions in Sub-Saharan Africa	Not educational data mining research
It should be written in English	Not available in full text

3.3.4 Data Extraction Strategy

Tricco *et al.* (2018) state that the data collection strategy consists of data extraction from articles. This study used manual data collection through a data extraction table consisting of the following data items: ID, Author, Country, Source, Journal, Objective, Factor Category, Findings, Data Mining Approach, Data Mining Algorithm, Data Collection Method, and Dataset Size. The data items and their descriptions are adapted from the work of Saa *et al.* (2019), as shown in Table 3.3.

Table 3.3: Data Items with their descriptions (Saa *et al.*, 2019, p. 13)

Item	Item Description
ID	Each ID number identifies each study for easy referencing during the review.
Author	The researcher and year of publication of the study.
Country	The country that the study was done in.
Source	The database source of the study.
Journal	The journal that published the study.
Objective	The primary goal of the study.
Factor Category	The categories of the factors include student demographics, online learning activities, student social information, etc.
Findings	The list of significant factors found in the study.
Data Mining Approach	The DM approaches that the study uses, such as clustering, classification, and regression.
Data Mining Algorithm	The DM methods used in the study include k-Nearest Neighbour, Support Vector Machines, and Decision Trees.
Data Collection Method	Methods used to gather data in the study, such as e-Learning systems, surveys, and student information systems.
Dataset Size	The number of participants in the dataset used in the study.

3.3.5 Assessment of Quality

According to Keele (2007), assessing the quality of selected studies includes: providing more detailed inclusion and exclusion conditions; investigating whether variances in the quality of studies explain variances in the findings; methods of measuring the significance of each study when findings are synthesized; guiding the interpretation of results and determining the importance of deductions; and providing suggestions for further research. According to Keele (2007), a single method to determine study quality is not agreed upon. However, the Cochrane Reviewers Handbook and CRD Guidelines suggest that quality is the extent to which a study reduces bias and increases external and internal validity (Booth *et al.*, 2010; Higgins *et al.*, 2019).

Bias refers to producing results that systematically deviate from accurate ones. Internally valid results are Unbiased. Internal validity refers to how a study's design will likely prevent bias. Internal validity is a precondition for external validity. External validity refers to how the study's observed effects apply outside the research. Most quality checklists include questions to assess how much articles have addressed bias and validity (Muthukrishnan *et al.*, 2017). Tricco *et al.* (2018) state that quality assessment measures the quality of articles against a recognized measurement scale. This study used a quality assessment checklist of nine conditions to evaluate the studies' quality further. The checklist was adapted from Kitchenham *et al.* (2009), shown in Table 3.4. Every question is scored using a three-point scale, "yes" equals 1 point, "partially" equals 0.5 points, and "no" equals 0 points. Studies that scored 4.5 points and above passed the assessment and were used in further synthesis.

Table 3.4: Quality Assessment

	Question
Q1.	Does the study introduce a model or framework of predictive modelling?
Q2.	Does the study explain why it chooses specific modelling or analytic techniques?
Q3.	Does the study predict students' academic performance relating to the framework or model introduced?
Q4.	Is the study focusing on sub-Saharan African undergraduate students in STEM fields?
Q5.	Is the data collection adequately explained in detail?
Q6.	Is the reliability of the study explained?
Q7.	Does the study interpret the results of the predictive analysis?
Q8.	Are the conclusions adding to the literature?
Q9.	Is the study expanding your understanding of EDM?

3.4 Phase 2: Conducting

In this section, the steps of conducting the SLR are described in detail, including identifying the research, selecting the studies, assessing study quality, extracting data, and synthesizing data.

3.4.1 Selecting the studies

In this research, the studies were identified and selected in line with the PRISMA framework mentioned in section 3.2.2. PRISMA shows the information flow in the different phases of the SLR and indicates the number of studies identified, included, excluded, and the reasons for the steps taken (Moher *et al.*, 2009). Figure 3.3 shows the PRISMA process where the researcher selected the studies that met the eligibility conditions mentioned in section 3.3.3 and examined their contents to verify their selection eligibility.

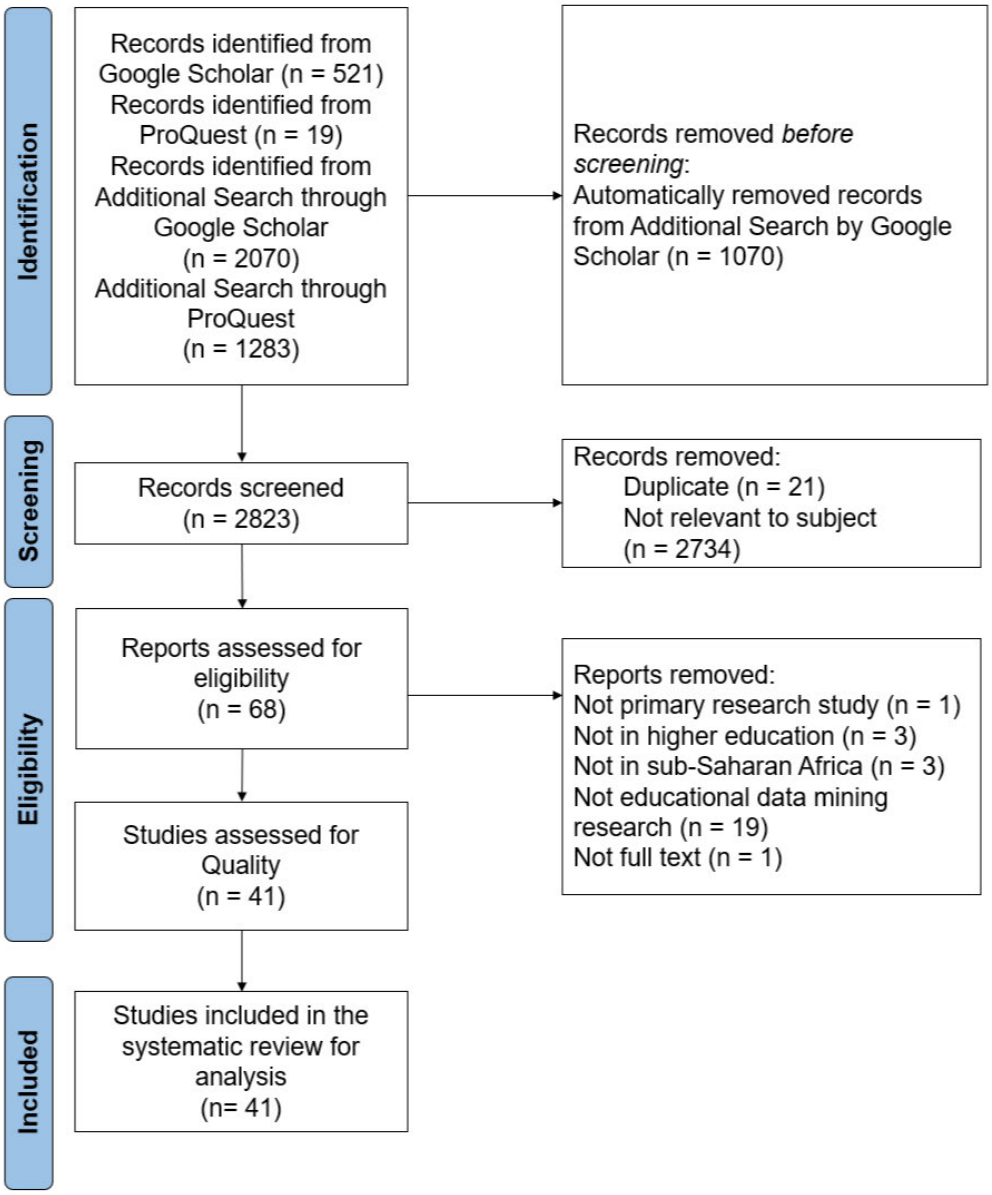

Figure 3.3: PRISMA flowchart

The researcher worked independently and built a quality assessment process into each stage to ensure transparency in implementing the SLR methodology.

Identification

On the 1st of May 2022, the researcher began the search with the search string in section 3.3.2. The researcher decided to add limiters to the online databases, so the search considered studies in the five years, from 2017 to 2021, according to the inclusion criteria specified in section 3.3.3. The results were sorted by their relevance. This would ensure that the most relevant studies were selected, with the remainder flagged as less appropriate. Google Scholar initially returned 521 results. ProQuest was then searched and produced 19 results. The researcher also

searched Scopus, which returned zero number of studies. The researcher discovered that the search string specified in section 3.3.2 was too broad and was not producing a useful number of results across the selected databases. The researcher thus had to revise the search string so that it could return valid results across the selected databases. The revised search terms are outlined in Table 3.5. The researcher added an asterisk to the search terms to maximize the possible results based on the alternatives caused by suffixes. The asterisk denoted that the word was pruned from its root form.

("impact" OR "effective*") AND ("education* data mining" OR "predictive modelling") AND ("academic" AND ("performance" OR "success")) AND ("Africa*" OR "developing countries")

Table 3.5: Revised Search terms for SLR

Connector	Research string
	("impact") OR ("effective*")
AND	("academic") AND ("performance" OR "success")
AND	("Africa*") OR ("developing countries")

The researcher conducted an additional search using the revised search string on 10 May 2022. The results were limited from 2017 to 2021 and were sorted by relevance. The researcher searched Google Scholar, which returned 2070 results. Upon manually capturing the studies from Google Scholar, the researcher discovered that Google Scholar is restricted to returning only the first 1000 relevant results of a search string by default. Google Scholar automatically excluded the rest of the 1070 results as it considered them irrelevant to the search string. The researcher then proceeded to ProQuest and exported the results to an Excel file. ProQuest returned 1283 results. Scopus returned zero results also with this revised search string. Scopus was then excluded from the list of selected databases to be searched, as it had zero studies produced for both search strings. Google Scholar and ProQuest produced a total of 2823 results minus the 1070 results excluded by Google Scholar, as shown in Table 3.6.

Table 3.6: Search results limited from 2017 to 2021

Databases	Results
Initial Search String	**540**
Google Scholar	521
ProQuest	19
Additional Search using Revised Search String	**2283**
Google Scholar	2070
Irrelevant results excluded by Google Scholar	- 1070
ProQuest	1283
Total	**2823**

Screening

The screening process was conducted in three stages due to a considerable number of initial results. Stage one consisted of the automatic selection of studies, which involves examining studies through their keywords, abstracts, and titles. The researcher removed 21 duplicate studies. Additionally, the researcher removed 2734 studies as they were unrelated to this SLR study's subject and did not meet the research keywords. Stage two consisted of reviewing studies based on the content and categorizing them under one of the following categories:

- excluded – not primary research.
- excluded – not higher education.
- excluded – research criteria.
- progressed to screening stage three.

The "excluded – not primary research" category was only used when the studies did not collect their data directly from the source but used secondary data. The "excluded – research criteria" category was only used when an abstract did not meet any of the clauses of the inclusion conditions specified in section 3.3.3. In stage three, the quality assurance consisted of reviewing all articles again, ensuring the researcher did consistent analysis. A study was progressed for full-text review when the researcher deemed the abstract inconclusive.

Eligibility

After screening the articles, 68 studies were reviewed for further assessment using the eligibility criteria. The researcher then manually reviewed the full-text articles. One was a review or SLR study, three were not higher education studies, three were not in sub-Saharan Africa, nineteen were not EDM research, and one was not available in full text. These studies were also excluded. The final set of studies selected for review or analysis comprised 41

studies. These studies are indicated and referred to as [S1], [S2], etc. The complete set of studies is contained in Table A.2 in Appendix B.

The researcher also assessed the quality of the selected studies by answering the questions in Table 3.4 of section 3.3.5 for each selected study. The researcher carried out the analysis of the studies collected by analyzing each study manually. Each study's quality assessment questions scores were totalled out of 9. Table A.3 in Appendix B presents the sum of the quality assessment scores for all the selected studies and their percentage results. Studies with a score of less than 4.5/9 (50%) were excluded from selection due to poor content and low quality. However, all 41 studies attained a score equal to or above 4.5 (50%). The top-scoring studies are [S7], [S19], and [S23], with scores equal to 9 (100%).

The researcher imported Table A.3 in Appendix B into SPSS version 27 for reliability analysis to ensure a reliable quality assessment of the nine questions. This study used Cronbach's alpha to measure the reliability of the quality assessment. According to Kanetaki *et al.* (2022), Cronbach's alpha seeks to discover the degree of connectedness of the results derived from the synthesis. The researcher did not use the Fleiss Multi-rater Kappa statistic tool to assess the overall agreements in researchers' ratings since the researcher worked independently in evaluating the quality of the selected studies. The attributes used in calculating Cronbach's alpha coefficient were the scale attributes that referred to the nine quality assessment questions mentioned in Table 3.4 of section 3.3.5. According to Fan and Thompson (2001), if the number of items measured is more than ten and Cronbach's coefficient is higher than .70 in the reliability analysis, then researchers can conclude that the research instrument is reliable. Furthermore, if items are less than ten, then Cronbach's alpha should be higher than .50 for the research instrument to be reliable (Pallant, 2020). In this study, nine items were in the research instrument. Thus, Cronbach's alpha coefficient is benchmarked at >.50. The Cronbach's alpha coefficient provided reliable results based on the above criteria, exhibiting internally reliable attributes amounting to 64.4%, as represented in Table 3.8.

Table 3.7: Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.644	.627	9

Included

The researcher extracted the 41 selected studies' findings according to the data items in Table 3.3 of section 3.3.4. Moreover, the researcher focused on determining the factors affecting the students' performance. More importantly, the significant factors and EDM techniques used in EDM research for predicting student academic performance so HEIs can see areas that need improvement and efficiently use available resources for early intervention. Table A.6 in Appendix B provides a summary of the data extracted from each study used in further analysis.

3.4.2 Synthesis

This research synthesized the findings from the selected studies using two types of analyses: bibliometric and thematic. Bibliometric analysis can be a critical component in determining the impact of research on the scientific community and society. It can be used for both structural (Quantitative) and conceptual (Qualitative) purposes (Bornmann, 2014). According to Guzmán-Valenzuela *et al.* (2021), bibliometric analyses provide descriptive statistics that reveal a particular topic's most crucial publication trends.

The researcher conducted a bibliometric analysis to identify the distribution of the selected studies by publication year, factor category, data collection technique, and type of EDM approach. Additionally, the researcher used Microsoft Excel Version 2304 Build 16.0.16327.20324 to generate tables and graphs to demonstrate the descriptive statistics for the bibliometric analysis. Given (2008) states that thematic analysis is a qualitative method that seeks to find current and vital themes. A thematic analysis follows a deductive-inductive process (Guzmán-Valenzuela *et al.*, 2021). The deductive categories include the themes of ethical issues, stakeholders, data analysis, and EDM methods. Moreover, inductive categories include structural factors and data governance. The thematic analysis addresses EDM's critical challenges in sub-Saharan African HEIs in a thematic format.

The researcher used NVivo release 1.7.1 software to import the selected studies from Endnote X9 and store them as cases. The researcher then analyzed each selected study case and coded them to their relevant nodes. In NVivo, the nodes are containers for people, places, themes, or other areas of interest. Many researchers have used NVivo for content analysis and view it as a leading qualitative tool (Abdous *et al.*, 2012). Therefore, it was appropriate to be used in the thematic analysis of this research.

3.5 Summary

This chapter showed the significance of research philosophy and design regarding this research. It helped the researcher to discover essential steps in data collection and analysis for a transparent investigation. This chapter started with a discussion of the research methodology used in the study, then proceeded to the analysis and sampling performed in this SLR, also the data collection strategy and quality assessment. This chapter created the foundation for the results of this research.

Chapter 4: Analysis Part 1

4.1 Introduction

This chapter is part of Phase 3 of the SLR process mentioned in the previous chapter, as shown in Figure 3.2. The researcher analysed 41 selected studies published in the five years, from 2017 to 2021, as discussed in section 3.3.3. This chapter seeks to report the results from the analysis to address the research questions and elaborate on the findings from the extracted data of this SLR research study. The analysis is organised into two sections. The first section, covered in this chapter (four), consists of the bibliometric analysis and the first part of the thematic analysis. The second section of the analysis consists of the remaining portion of the thematic analysis covered in Chapter five.

4.2 Bibliometric Analysis

A bibliometric analysis provides descriptive statistics relating to the common trends in a particular research phenomenon (Bornmann & Mutz, 2015). For this study, bibliometric analysis was performed on the 41 selected studies based on year of study, country of study, and publication type, as shown in Table 3.7 in section 3.4.1.

4.2.1 Distribution of Research by Year

This SLR analysed the 41 selected studies from 2017 to 2021, as shown in Figure 4.1. The researcher did this to observe the EDM publication evolution and new findings in the current research area.

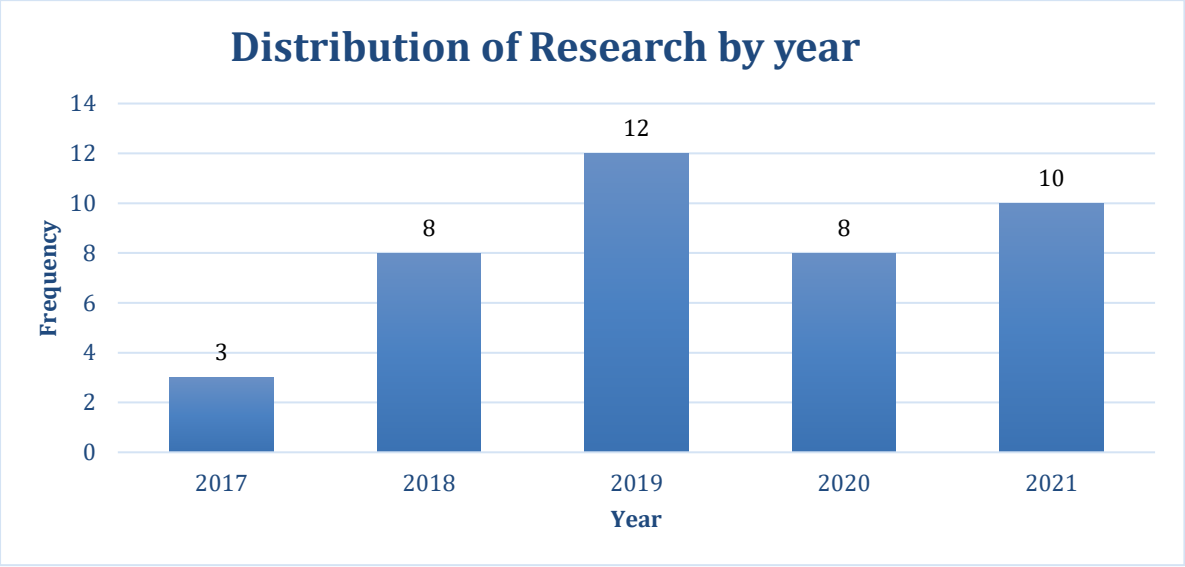

Figure 4.1: Distribution by year

In 2017 there were three studies (7%); in 2018, there was a rise to eight studies (20%); in 2019, there were 12 studies (29%) which was the highest; in 2020, there was a decrease to eight studies (20%) which may have been due to COVID-19; and in 2021, there was a rise again to 10 studies (24%). As seen, the rate of published studies on this topic has rapidly increased over the past five years, but it is still in its infancy (Khan & Ghosh, 2021).

4.2.2 Distribution of Research by Country

This study focuses on EDM's effectiveness in predicting students' academic performance in Sub-Saharan Africa. Generally, sub-Saharan African countries refer to all African countries, excluding Libya, Tunisia, Morocco, Algeria, Djibouti, Egypt, Sudan, and Somalia (Muthukrishnan *et al.*, 2017). Figure 4.2 shows the distribution of studies by country, South Africa had 23 studies (56%), and Nigeria had 11 studies (27%). These two counties dominate the reviewed studies on EDM in HEIs in the sub-Saharan African context. This finding aligns with Zawacki-Richter *et al.* (2019), which is unsurprising as South Africa and Nigeria are the two largest economies in sub-Saharan Africa (Maphosa & Maphosa, 2020).

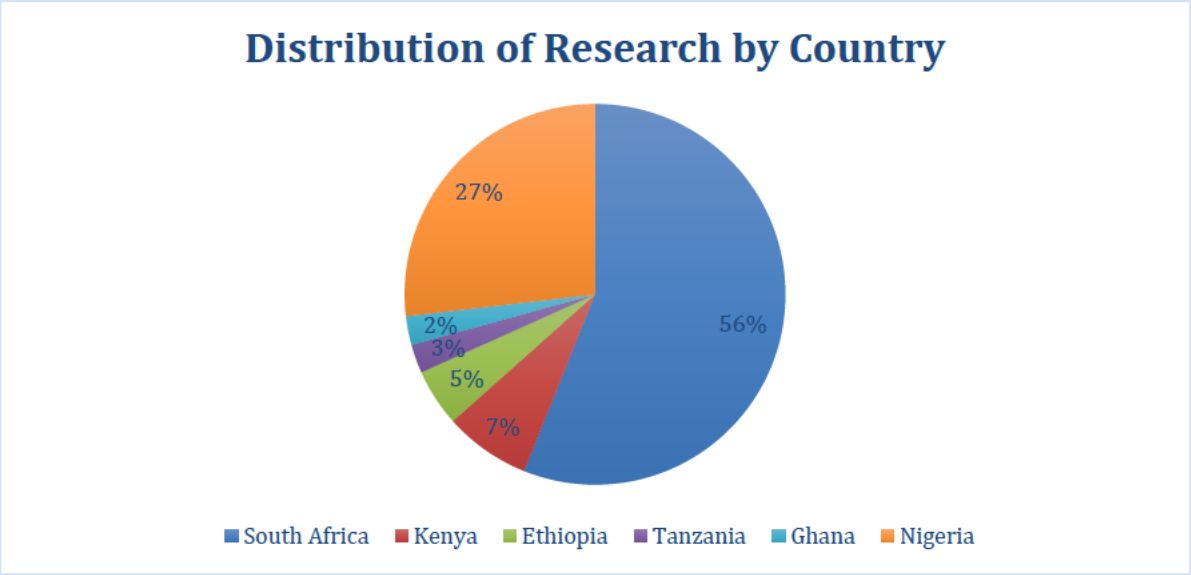

Figure 4.2: Distribution by country

There were three studies (7%) from Kenya, there were two studies (5%) from Ethiopia, there was one study (3%) from Tanzania, and there was one study (2%) from Ghana. The fact that the reviewed studies were conducted in six sub-Saharan African countries suggests that this field has not been embraced by most countries in this region, probably because of the state of economies in these countries (Maphosa & Maphosa, 2020).

4.2.3 Distribution of Research by Publication Type

This SLR focused on journal articles (23 studies), case studies or thesis papers (13 studies), conference papers (four studies), and a book section (one study), as shown in Figure 4.3. In contrast to current SLR studies focusing primarily on conference and journal articles (Shafiq et al., 2022), this SLR included thesis and case study articles as it assists in looking at the student's academic performance prediction problem from different views. The researcher included studies that presented a plausible methodology and comprehensive results from the prediction model experiments.

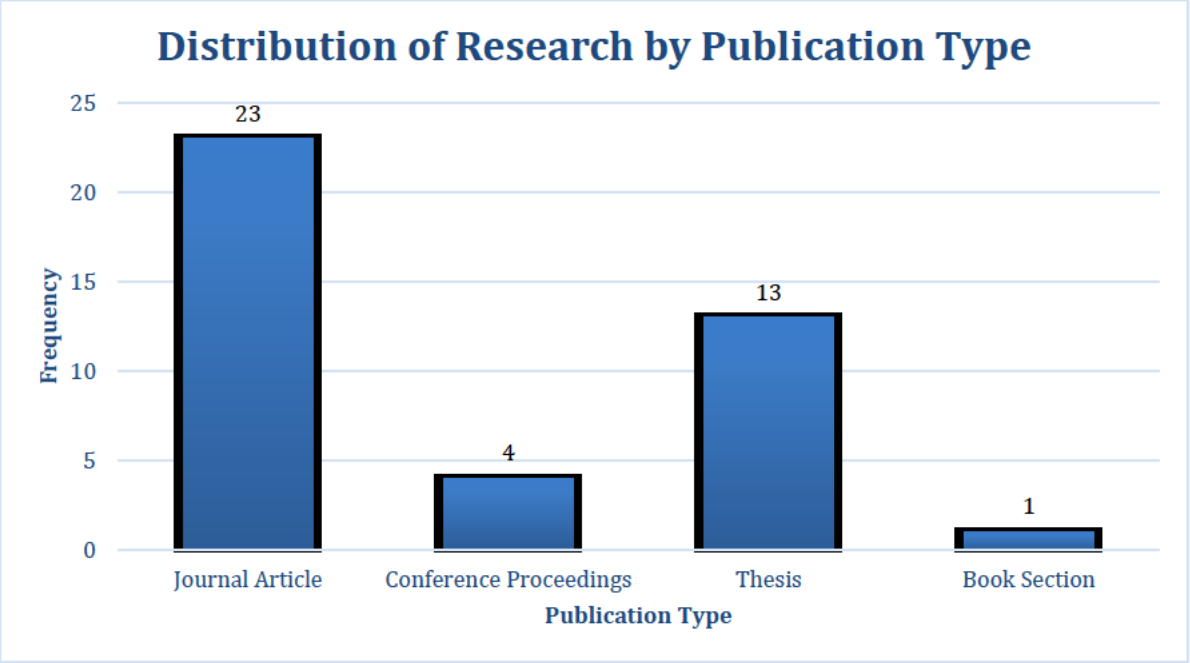

Figure 4.3: Distribution by publication type

4.3 Thematic Analysis

A thematic analysis of existing literature is presented in this section, which attempts to identify significant factors influencing students' academic performance prediction and the effectiveness of EDM. Figure 4.4 shows the organisational structure of this thematic analysis extracted from NVivo. Student academic performance analysis and prediction are EDM's two commonly investigated research areas. Even though their goals vary, the effects of performance analysis considerably influence performance prediction. These two aspects and the challenges they bring will be explained later in the analysis to bring them into theoretical alignment, thus creating a holistic conclusion.

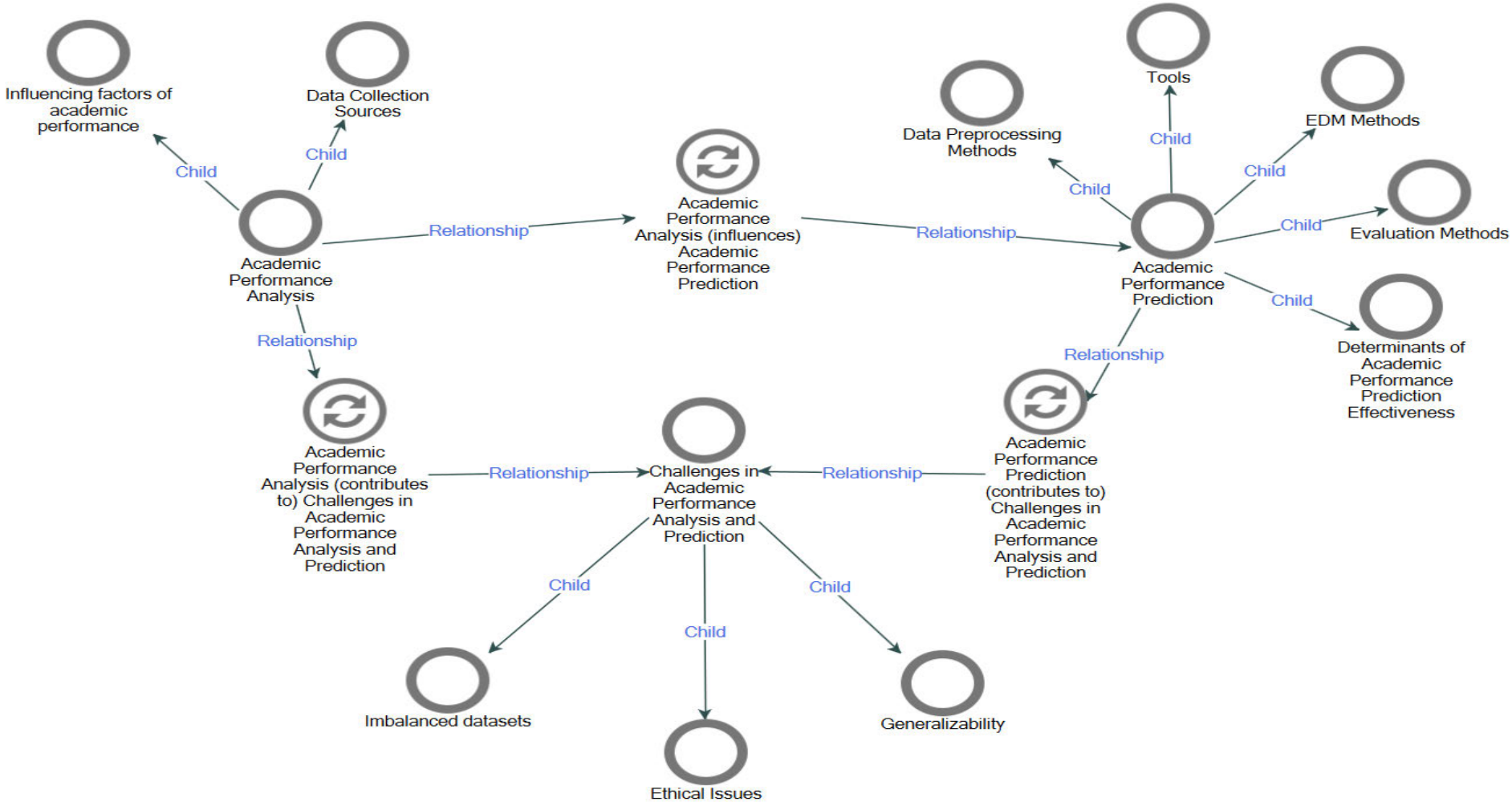

Figure 4.4: Thematic analysis map

This thematic analysis presents studies highlighting the importance of particular factors in the predictions. These studies are listed in bullet points arranged in ascending order based on each study's ID. The findings of this section are summarised in Appendix B Table A.4 – A.7. Additionally, the tables in Appendix B provide information on the list of factors used, methods used, and the key factors influencing students' academic performance prediction of each study. Moreover, the tables in the thematic analysis include information on each study's ID, predictors used, data collection technique, and dataset size. The predictors used are shown as (+) if significant or (-) if not significant. The thematic analysis map (figure 4.4) is represented in a structured format indicating the main themes, sub-themes and items in Table 4.1 for ease of reading.

Table 4.1: Detailed thematic analysis structure

Main Themes	Items	Sub-Items
Academic Performance Analysis	▪ Previous and current class performance ▪ Demographics ▪ Socio-economic factors ▪ E-Learning Behaviours	▪ Pre-enrolment Grades ▪ Internal Assessment Grades ▪ External Assessment Grades ▪ Age ▪ Gender ▪ Other Demographic Features
Academic Performance Prediction	▪ Data cleaning ▪ Data Transformation ▪ Feature selection ▪ Data Segmentation ▪ Classification ▪ Regression ▪ Clustering ▪ Aim of prediction ▪ Level of prediction	▪ Information Gain ▪ Gain Ratio ▪ Relief-F ▪ Success prediction ▪ Final Grade prediction ▪ Degree Level ▪ Course Level ▪ Year Level
Challenges in Academic Performance Analysis and Prediction	▪ Issue of Imbalanced Datasets ▪ Ethical issues ▪ Generalizability of data	

4.3.1 Academic Performance Analysis

Choosing the correct predictors is crucial in achieving an efficient prediction task. Therefore, factors influencing the process of knowledge discovery must be identified before the prediction of students' academic performance (Helal *et al.*, 2018). Moreover, the increase in student dropouts and the decrease in their success are the main concerns of various HEIs. A preliminary analysis of student performance can assist HEIs in this case (Khan & Ghosh, 2021). This

section presents existing literature identifying and measuring influencing factors of students' academic performance analysis.

4.3.1.1 Factors influencing academic performance

In exploring the determinants of academic performance, a crucial facet involves examining the distinctive input factors employed within EDM to predict the academic performance of sub-Saharan African undergraduate STEM students. This investigation directly addresses sub-research question one: *What key input factors are used in EDM to predict sub-Saharan African undergraduate STEM students' academic performance?*

According to Zollanvari *et al.* (2017), identifying input factors as predictors for subsequent use in predicting students' academic performance is difficult, as several factors influence academic performance in non-linear ways. These factors dictate the data that must be collected and analysed. The most common factors researchers explore include previous and current class performance, demographics, socio-economic factors, and e-learning behaviours. The findings of this section are summarized in Appendix B Table A.4.

Previous and current class performance

The previous and current class performance theme is the most frequently occurring theme in this analysis, as it was cited across all 41 selected studies. The previous and current class performance factors include pre-enrolment grades and internal and external assessments.

Pre-enrolment Grades: These factors relate to students' scores or grades before enrolling, including grades earned in prerequisite courses, matric scores, and admission point scores (APS). Regarding using previous grades for students' academic performance prediction, the commonly used features are matric and APS scores ([S1], [S15], [S21], [S32], [S36], [S40]). Although matric results have been essential in predicting students' academic performance, some studies do not regard it as necessary ([S25], [S27]). The following are studies that reported pre-enrolment grades as being an effective predictor of students' academic performance, as presented in Table 4.2:

Table 4.2: Studies focusing on pre-enrolment factors

ID	Predictors used
[S1]	(+) Background data and (-) pre-university data of students enrolled in a science degree.
[S15]	(+) Average matriculation results, (+) type of high school the student attended, (+) first semester first-year university results, (+) attendance, (+) self-study hours, and (+) lecturers' competency.
[S21]	(+) Lecturer (likes, name, number), (+) subject (likes, name, code), (+) grade (average).
[S25]	(-) National Benchmark Test (NBT), (+) School Quintile, (+) Number of years in degree, (+) Gender, (+) Age at first year, (+) From Rural / Urban, (+) Home Country, (+) Home Province, (+) Year Started, (+) Race, (+) Mathematics Major, (+) Plan Description, (+) Home Language, (-) Matric Maths grades.
[S27]	(+) High school GPA, (+) placement test scores, (+) course GPA, (+) year of admission, (+) year of study, and (+) majors.
[S32]	(+) Gender, Race, (+) Residence, (+) Home language, (+) Second language, (+) APS score, (+) Mathematics score.
[S36]	(+) High school results, (+) first-year results, (+) demographics, and (+) socio-economic factors.
[S40]	(+) Career Flexibility, (+) High School Final Grade, (+) Math Grade, (+) Pre-University awareness, (+) Teacher Inspiration, (+) Financial Aid, (+) Extracurricular, (+) Parent Career, (+) Societal Expectation, (+) Career Earning, (+) Self-Efficacy, (+) Gender, (+) Age, and (+) Family Income.

- [S1] used two datasets to develop a recommendation system to guide students on their academic path, one containing the undergraduate science course registration data and another containing the corresponding matric results. The data also includes an indication of the final grade of the student and the number of years taken to obtain that outcome. The results show that EDM classification models were highly accurate in predicting whether a new student was likely to qualify in 3 years, more than three years, or fail to meet the requirements using their matric results and enrolment data.
- In this study, [S15] developed a predictive model using attributes including first-semester university results, average matriculation, lecturer competency, self-study hours, and attendance. The results of this study show that these predictors used in the EDM model can predict a first-year university student's academic performance before their final exams.
- [S21] used students' APS scores and lecturer data for students' academic performance prediction. The findings suggested the possibility of predicting students who will succeed in an engineering course they choose to enrol in.
- In contrast, [S25] predicted undergraduate students' likelihood to drop out using the student's schooling, biographical, and individual characteristics. These factors were

found to be significant. However, they rejected using only pre-university factors such as the National Benchmark Test (NBT) and matric results as a primary schooling system measure for predicting students' academic performance. They suggested that these pre-university factors should be used with the university performances, especially the first-year level performances.

- [S27] conducted prediction modelling to enhance students' academic performance in computer courses. The findings showed that high school grades and previous computer knowledge were not significant in predicting students' academic performance.
- According to [S32], using discrete-time models also enabled testing the influence of Mathematics and APS scores on students' dropout risk. It was assumed that the dropout risk was more prominent in the first year of study and lessened over time. The findings revealed that the effects were not substantial and remained constant over time.
- [S36] aimed to discover factors influencing first-year engineering students' academic performance. The significant factors were ethnicity, school province, age, and mathematics and physics grades.
- [S40] developed a general EDM framework using ensemble methods for predicting STEM students' academic performance. Five significant factors significantly influenced the choice of students enrolling in STEM programs. These factors included the high school final exam score, beliefs in competency for succeeding in a STEM-related career, inspiration from the high school teachers, and expected career flexibility.

Internal Assessment Grades: Internal assessments in the literature are related to the internally evaluated grades in class tests, quizzes, assignments, or other related elements. Most literature focusing on this aspect positively influenced final performance ([S2], [S13], [S14], [S18], [S23]). The results also indicated a significant influence of class attendance ([S13]) and group assignment scores ([S23]) on students' academic performance prediction. The following are studies that reported internal assessment grades as being an effective factor in predicting students' academic performance, as shown in Table 4.3:

Table 4.3: Studies focusing on internal assessments

ID	Predictors used
[S2]	(+) Class grade, (+) summative assessment, (+) formative assessment, and (+) socio-demographics.
[S13]	(+) Health, (+) Background knowledge, (+) Student Attendance, (+) Lecture time, (+) Family stress, (+) Parent education, (+) Family income, (+) Student Attitude, (+) Lecturer Attitude, (+) Student Fear and Perception, (+) Student Study Habit, (+) Tutorials, (+) Extra Classes, (+) Lecturer Dedication, (+) Lecturer Availability, (+) Teaching aids, (+) Teaching Style, (+) Communication Skills, (+) Class population, (+) Electricity, and (+) Facilities.
[S18]	(+) GPA of students throughout six semesters.
[S23]	(+) Study hours per week, (+) bursary, (+) class attendance, (+) full-time or part-time classes, (+) number of modules registered, (+) language, (+) test marks, (+) group assignment marks, and (+) number of employed parents or guardians.
[S38]	(+) Two formal semester test marks, (+) online cumulative average quiz mark, (+) final period mark, (+) grade obtained in a prerequisite module, (+) repeating the module, (+) gender, and (+) high school quintile ranking.

- [S2] used predictors such as individual grades, year of completion, year of admission, grades achieved from the courses enrolled, and the degree class to predict the final grades of Communication and Information Technology students. This study found that timely completion of the first two courses leads to a high grade in computer architecture programs. The discrete mathematics and programming course grades significantly influence students' academic performance prediction.
- [S13] explored the factors that influence undergraduate programming students' academic performance and developed EDM methods for students' academic performance prediction. The findings indicated that the most influential factors included student health, university facilities, erratic power supply, student attendance, fearful perception of programming courses, and the attitude of students and lecturers.
- [S18] predicted students' academic performance using only the raw scores or results of students. The results revealed that the students' raw GPA scores over six semesters and previous semester results significantly predicted students' academic performance.
- [S23] used EDM to predict students' success. The results showed that group assignments positively correlated with the pass rate. Hence, students must be encouraged to participate more actively in group assignments.
- [S38] implemented five EDM models to identify at-risk Business Statistics students early to provide strategic interventions to improve performance. The significant factors

were marks achieved in a prerequisite course, semester test one marks, and average quiz marks.

External Assessment Grades: EDM researchers frequently explore Cumulative Grade Point Average (CGPA) as a form of external assessment for predicting student performance. Notably, CGPA focuses on all previous performances of students in their program or degree and combines them into a single value. Hence, the majority of literature establishes CGPA as an important factor in students' academic performance prediction ([S3], [S11], [S17], [S37], [S39], [S41]). However, some studies have also argued that CGPA is ineffective when used alone in students' academic performance prediction ([S7], [S32]). The following are studies that reported external assessment grades as being an effective predictor of students' academic performance, as shown in Table 4.4:

Table 4.4: Studies focusing on external assessments

ID	Predictors used
[S3]	(-) Year of entry, (+) course of study, (+) first three years GPA, and (+) Final CGPA.
[S7]	(+) High school average, (-) end-of-semester examination marks, and (-) CGPA.
[S11]	(+) Age, (+) Sex, (+) Marital status, (+) Secondary school area, (+) Secondary school type, (+) Attended primary school, (+) Years before admission, (+) Sports activeness, (+) Weekly study time, (+) Sponsor type, (+) Sponsor income, (+) Sponsor support, (+) Sponsor qualification, (+) University accommodation, (+) Work and Study, (+) Family size, (+) Course interest, (+) post-UTME score, (+) JAMB score, (+) Average SSCE score, (+) CGPA, and (+) Postgraduate degree.
[S17]	(+) CGPA, (+) course GPA, and (+) raw matric grades.
[S32]	(+) Gender, (+) Race, (+) Residence, (+) Home language, (+) Second language, (+) APS score, (+) Mathematics score, and (+) CGPA.
[S37]	(+) Field Choice Interest, (+) Ethiopian University Entrance Examination result, (+) High school GPA, (+) First-Semester First-Year grades, and (+) CGPA.
[S39]	(+) Gender, (+) admission grades, (+) first-year GPA, (+) students' registration numbers, (+) department, and (+) academic standing.
[S41]	(+) Gender, (+) marital status, (+) country of origin, (+) date of birth, (+) level of study, (+) telephone number, (+) home address, (+) email address, (+) name of previous school, (+) admission date, (+) JAMB score, and (+) CGPA.

- A study by [S3] used the first three years' GPA of students to predict the final year CGPA. This study found a significant influence of first-year GPA on engineering students' academic performance.
- In contrast, [S7] found a weak correlation between students' best six GPAs and CGPA.

- [S11] predicted students' academic performance only for students with a CGPA of less than 3.0 and categorised them into high-risk and low-risk. The findings showed that average Senior Secondary Certificate of Education (SSCE) score, post-Unified Tertiary Matriculation Examination (UTME) score, Joint Admissions and Matriculation Board (JAMB) score, secondary school type, weekly study time, sponsor type, sponsor support, sponsor income, sponsor qualification, secondary school area, years before admission, work and study, university accommodation, postgraduate degree, course interest, and sports activeness were significant factors for predicting students' CGPA.
- [S17] used EDM to analyse students' performance, including their CGPA and class of grade. The results revealed that out of the 21 first-year courses offered, only seven courses, such as mathematics, physics, and chemistry, commonly occur together as failed courses, and they significantly influence students' academic performance prediction.
- Another study by [S32] found that CGPA predicts a reduced graduation potential than what occurs when modelled alone.
- [S37] used predictors such as preparatory school GPA, entrance examination results, and first-semester first-year academic performance to predict the CGPA achieved by students. The results showed that these factors significantly influenced students' academic performance prediction.
- [S39] predicted students' academic performance using a dataset with attributes such as level of matric achieved, gender, program, first semester GPA, second semester GPA, CGPA, and current standing. These factors proved significant in predicting students' academic performance.
- [S41] predicted the students' academic performance using factors such as high school exam scores, region, gender, age, and CGPA to enable efficient and timely implementation interventions. This study found that students with high JAMB grades are more likely to attain a high CGPA. Additionally, this study found that students in their third and fourth years were more likely to attain high CGPA.

The researcher noted that academic performance factors such as pre-enrolment, internal assessment, and external assessment grades are essential factors that should be considered in any EDM project, as shown by the evidence above. The evidence presented also suggested that future work should focus on students' academic performance prediction using pre-enrolment

and enrolment data using EDM early, such as the first year of a course, to assist HEIs in making timely and meaningful decisions to support low-performing students.

Demographics

Exploring students' academic and demographic factors are common in students' academic performance analysis. The most frequent demographic characteristics in the analysis were age and gender. Other demographic factors, such as ethnicity and race, were used as predictors and were found to be significant when paired with other demographic features.

Age: The commonly used demographic factor to predict academic performance is age. Many researchers have found a strong correlation between age and performance ([S1], [S5], [S16], [S25], [S29], [S33], [S36]). These studies suggested that the strong correlation was caused by older students being more experienced, highly motivated, and possessing effective studying practices. One study found that age does not significantly influence students' academic performance prediction ([S41]). There were 24 studies identified that used age for predicting students' academic performance. However, only eight studies reported its significance on students' academic performance prediction. The following are studies that reported age as being an effective predictor of students' academic performance, as shown in Table 4.5:

Table 4.5: Studies focusing on the Age factor

ID	Predictors used
[S1]	(+) School quintile, (+) Age at first year, (-) high school grades, and (+) province.
[S5]	(+) UTME score, (+) age, (+) SSCE score, (+) 100 level grade.
[S16]	(+) Age, (+) Sex, (+) Employment status, (+) Type of school attended, (+) Preparatory completion year, (+) Preparatory attended region, (+) Field of study, (+) Financial sources, (+) Admission scores, (+) Year of admission, and (+) students' status.
[S25]	(-) NBT, (+) School Quintile, (+) Number of years in degree, (+) Gender, (+) Age at first year, (+) From Rural / Urban, (+) Home Country, (+) Home Province, (+) Year Started, (+) Race, (+) Mathematics Major, (+) Plan Description, (+) Home Language, (-) Matric Maths grades.
[S29]	(+) Admission point score, (+) home province, (+) home language, (+) home country, (+) stress and time pressure, (+) age at first year, (+) dwelling value, (+) academic self-efficacy, (+) work status, (+) financial support, (+) parents' education, (+) family income, (+) parents' occupation, (+) college activity participation, (+) sense of loneliness, (+) class communication, (+) organization and attention to study, (+) interest in sports, and (+) statistics major.
[S33]	(+) Attendance of the student in a lecture room, (+) end semester exam mark, (+) friends and family support, (+) economic status, (+) living location, and (+) parents' qualification.
[S36]	(+) Matric results, (+) first-year results, (+) demographic information, (+) age at first year, and (+) socio-economic information.
[S41]	(+) Gender, (+) marital status, (+) country of origin, (+) date of birth, (+) level of study, (+) telephone number, (+) home address, (+) email address, (+) name of previous school, (+) admission date, (+) JAMB score, and (+) CGPA.

- [S1] built a recommendation system to advise students on their future academic paths based on the selected courses. The study found that the significant factors were students' age at first year, school quintile, and province.
- [S5] developed an EDM classification model for students' academic performance prediction. The findings indicated that age was essential in predicting students' academic performance.
- [S36] implemented EDM methods to predict first year students' academic performance. The results revealed that age, school province, ethnicity, and physics and mathematics grades were significant factors.
- [S16] used an EDM predictive model to predict undergraduate students' attrition or retention in HEIs. Based on the results, the age of students was identified as a significant contributing factor behind student attrition and retention.
- [S25] found the students' age in their first year is among the eight most significant demographical attributes in students' academic performance prediction.
- [S29] sought to identify at-risk students early using prediction tasks. The findings revealed that the influential features of students' academic performance included the age at first year, matric results, school quintile, the year started, plan description, plan code, and majors.
- [S33] predict students' academic performance using attributes such as course enrolled in, satisfaction level, Age, and parental factors. The results showed that these factors significantly influenced students' academic performance prediction.
- In contrast, [S41] used various predictors for students' academic performance prediction. Their findings showed that age was not significant in predicting students' academic performance.

Gender: The second most used demographic factor in students' academic performance prediction is gender. This is unsurprising as the relationship between students' gender and academic performance has been broadly debated in EDM literature (Alturki *et al.*, 2022). Nineteen studies were identified that used gender to predict students' academic performance. However, only six studies reported its influence on the prediction ([S5], [S25], [S26], [S32], [S34], [S41]). Some researchers found students' gender as not significant ([S26], [S32]). However, in other studies, it was significant, with either the males or the females performing

better in predictions ([S5], [S25], [S34], [S41]). The following are studies that reported gender as being an effective predictor of students' academic performance, as shown in Table 4.6:

Table 4.6: Studies focusing on the Gender factor

ID	Predictors used
[S5]	(+) Gender, (+) Unified Tertiary Matriculation Examination (UTME) score, (+) age, (+) Senior Secondary Certificate of Education (SSCE) score, and (+) 100 level grade.
[S25]	(-) NBT, (+) School Quintile, (+) Number of years in degree, (+) Gender, (+) Age at first year, (+) From Rural / Urban, (+) Home Country, (+) Home Province, (+) Year Started, (+) Race, (+) Mathematics Major, (+) Plan Description, (+) Home Language, (-) Matric Maths grades.
[S26]	Students' (+) loan allocation data, (+) living locations, (+) coursework, (+) remarks, (+) final examination results, (+) mathematics grades, (+) gender, (+) age, and (+) placement scores.
[S32]	(+) Gender, (+) Race, (+) Residence, (+) Home language, (+) Second language, (+) APS score, (+) Mathematics score, and (+) CGPA.
[S34]	(+) Quintile, (+) province, (+) gender, (+) financial assistance, (+) township school, (+) grade, (+) outcome, (+) discussion, (+) message, and (+) time.
[S41]	(+) Gender, (+) marital status, (+) country of origin, (+) date of birth, (+) level of study, (+) telephone number, (+) home address, (+) email address, (+) name of previous school, (+) admission date, (+) JAMB score, and (+) CGPA.

- [S5] utilized EDM models to deduce the rules to classify students' academic performance. The study found that students' gender was a crucial factor in students' academic performance prediction.
- In a study by [S25], the gender of students was found to be among the eight most important demographical attributes that play an essential role in classifying students' academic performance.
- In contrast, [S26] used EDM methods to address the issue of low-performing students in a management program. Their results revealed that some predictors, such as gender, living location, and sponsorship, did not affect students' academic performance.
- [S32] used EDM methods to predict the timing and occurrence of student dropout in undergraduate engineering courses. The study found that the influence of mathematics scores, APS scores, and gender was not significant in the prediction.
- [S34] used EDM to predict students' likelihood of dropping out. The study found that economic and cultural capital combined with the gender factor provided better performances in students' academic performance prediction.

- [S41] used EDM for predicting students' academic performance, enabling efficient and timely interventions. Their results suggested that female students are more likely to gain higher marks than males.

Other Demographic Features: The other demographics available in the literature include school quintile, province of the student, country, home language, deferment rate, employment status, financial sources, race, and ethnicity, which have often been used as predictors. These factors have been widely explored in literature and found to influence students' academic performance prediction significantly ([S1], [S8], [S16], [S25], [S32]). In one study, ethnicity is not significant in predicting students' academic performance ([S4]). In another study, marital status was not significant in students' academic performance prediction ([S11]). It was also found that urban students performed better than rural students ([S41]). The following are studies that reported other demographics as being an effective predictor of students' academic performance, as shown in Table 4.7:

Table 4.7: Studies focusing on other demographic factors

ID	Predictors used
[S1]	(+) Background data and (-) pre-university data of students enrolled in a science degree.
[S4]	(+) Preadmission scores, (+) college, (-) geopolitical zone, (+) CGPA, and (+) year of graduation.
[S8]	(+) Rural/Urban School, (+) School Quintile, (+) Home language, (+) Class communication, (+) Home Country, (+) APS, (+) Age, (+) year of study, (+) Time management, (+) Home Province, (+) English, (+) Plan Description, (+) Student Absent Days, (+) Financial support, (+) Accounting, (+) Science, (+) Interest in sports, (+) Visited Resources, (+) International, (+) Mathematics, (+) Cognitive Difficulties.
[S11]	(+) Age, (+) Sex, (+) Marital status, (+) Secondary school area, (+) Secondary school type, (+) Attended primary school, (+) Years before admission, (+) Sports activeness, (+) Weekly study time, (+) Sponsor type, (+) Sponsor income, (+) Sponsor support, (+) Sponsor qualification, (+) University accommodation, (+) Work and Study, (+) Family size, (+) Course interest, (+) post-UTME score, (+) JAMB score, (+) Average SSCE score, (+) CGPA, and (+) Postgraduate degree.
[S16]	(+) Age, (+) Sex, (+) Employment status, (+) Type of school attended, (+) Preparatory completion year, (+) Preparatory attended region, (+) Field of study, (+) Financial sources, (+) Admission scores, (+) Year of admission, and (+) students' status.
[S20]	(+) Enrolment, (+) Drop out, (+) deferment, (+) progression rates.
[S23]	(+) Study hours per week, (+) bursary, (+) class attendance, (+) full-time or part-time classes, (+) number of modules registered, (+) language, (+) test marks, (+) group assignment marks, and (+) number of employed parents or guardians.
[S25]	(-) NBT, (+) School Quintile, (+) Number of years in degree, (+) Gender, (+) Age at first year, (+) From Rural / Urban, (+) Home Country, (+) Home Province, (+) Year Started, (+) Race, (+) Mathematics Major, (+) Plan Description, (+) Home Language, (-) Matric Maths grades.

[S32]	(+) Gender, (+) Race, (+) Residence, (+) Home language, (+) Second language, (+) APS score, (+) Mathematics score, and (+) CGPA.

- [S1] used EDM models to predict whether a new student will likely qualify in three years, more than three years, or drop out. The results revealed that background factors, including students' age at first year, school quintile, and province, had a noticeable influence on academic performance prediction.
- [S4] examined the influence of the geopolitical zone of origin on students' academic performance prediction. The results showed that the ethnicity of students was not statistically significant in predicting their academic performance. Hence, these findings suggested that using ethnicity in EDM predictions was inappropriate.
- [S8] aimed to predict the students at risk of not completing their year of study. The features used for this characteristic are commonly home province, home country, whether the student is from rural or urban, race, etc. The home province feature was among the most contributing factors to students' academic performance.
- [S11] found that the attributes shared by two or more feature selection algorithms, such as marital status, family size, gender, and attended high school, may contribute least to students' classified group.
- [S16] developed an EDM predictive model to predict undergraduate students' retention and attrition in HEIs. Their results identified student demographic factors such as types of preparatory attended school, division, department, preparatory completion year, and financial sources as the most significant student retention and attrition features.
- [S20] predicted students' progression rate in a program. The factors used in the study were the deferment rate, the dropout rate, and the enrolment rate. The results showed that the number of deferments of students highly determined the progression rate of students. More students were deferring than those dropping out of their courses altogether.
- [S23] used EDM to predict students' success. The results showed that bursary and group assignments positively correlated with the pass rate.
- [S25] found that the most contributing factors were biographical and individual characteristics such as race, home language, home country, and home province which played an essential role in predicting students' academic performance.

- [S32] used EDM methods to predict the timing and occurrence of student dropout in undergraduate engineering courses. The study found that the dropout risk for students of other racial backgrounds is considerably more than that of White students. The findings also showed that English first language students were more at-risk of dropping out than those who used it as a second language. This is a very interesting result, as language is often considered a limiting factor in student success, so this is the opposite of what is commonly assumed in the literature.
- Another study by [S41] found that students from urban areas were more likely to attain a high CGPA.

Socio-economic factors

Various EDM researchers have explored socio-economic attributes ([S29]). Moreover, many researchers have shown a positive relationship between socio-economic status and academic success ([S11], [S16], [S32]). The socio-economic attributes examined in the studies include family characteristics such as family income, parent or guardian occupation, and financial support. These socio-economic factors affect students' academic performance, specifically students' dropout behaviours ([S10], [S16], [S26], [S29]). In some studies, place of residence was shown to be highly influential and valuable in influencing student dropout behaviour ([S32]). The following are studies that reported socio-economic factors as being an effective predictor of students' academic performance, as shown in Table 4.8:

Table 4.8: Studies focusing on socio-economic factors

ID	Predictors used
[S10]	(+) Motivation, (+) Family support structure, (+) specialization, and (+) employment status.
[S11]	(+) Age, (+) Sex, (+) Marital status, (+) Secondary school area, (+) Secondary school type, (+) Attended primary school, (+) Years before admission, (+) Sports activeness, (+) Weekly study time, (+) Sponsor type, (+) Sponsor income, (+) Sponsor support, (+) Sponsor qualification, (+) University accommodation, (+) Work and Study, (+) Family size, (+) Course interest, (+) post-UTME score, (+) JAMB score, (+) Average SSCE score, (+) CGPA, and (+) Postgraduate degree.
[S16]	(+) Age, (+) Sex, (+) Employment status, (+) Type of school attended, (+) Preparatory completion year, (+) Preparatory attended region, (+) Field of study, (+) Financial sources, (+) Admission scores, (+) Year of admission, and (+) students' status.
[S26]	Students' (+) loan allocation data, (+) living locations, (+) coursework, (+) remarks, (+) final examination results, (+) mathematics grades, (+) gender, (+) age, and (+) placement scores.
[S32]	(+) Gender, (+) Race, (+) Residence, (+) Home language, (+) Second language, (+) APS score, (+) Mathematics score, and (+) CGPA.

- [S10] found that family support structure, change in program offering, and employment status were the most influential factors in students' academic performance prediction.
- [S11] examined the factors leading to students' poor performance using EDM in a Nigerian university. The identified factors included limited family educational background, poor study plans, lack of employment for graduates, limited support from guardians and parents, financial income, and a lack of interest in the course.
- [S16] found that poor social integration, the field of placement, mismatched areas of interest, poor commitment, and poor preparation are significant factors leading to students dropping out.
- [S26] used the predictor variables such as living locations, annual family income, family sizes, family status, parents' occupations, gender, enrolled programs, admission types, qualifications, medium of teaching, and grades in high school to classify if students will fail or pass. These factors were found to be crucial in predicting students' academic performance.
- [S32] used EDM methods to predict the timing and occurrence of student dropout in undergraduate engineering courses. The study found that the outcomes of the type of residence differ with time. For example, first-year students in private accommodations were more likely to drop out than on-campus students. In contrast, third-year students living in private-based accommodations were less likely to drop out than on-campus students.

E-Learning Behaviours

Various HEIs use learning management systems to enhance knowledge distribution (Khan & Ghosh, 2021). The commonly extracted data from LMS include e-learning behaviours such as solving quizzes, viewing files, content submissions, and reading and posting messages, which are significant predictors of students' academic performance ([S34]). Other predictors include opening resources and responding to course evaluation surveys ([S9], [S12]). Engagement data obtained from LMS is one of the primary metrics identified for measuring students' academic performance ([S35]). This analysis found a strong relationship between e-learning behaviours and students' academic performance ([S30]). The following are studies that reported E-learning

behaviours as being an effective predictor of students' academic performance, as shown in Table 4.9:

Table 4.9: Studies focusing on e-learning behaviours

ID	Predictors used
[S9]	(+) Discussion, (+) Gender, (+) Nationality, (+) Relation, (+) Semester, (+) Topic, (+) Section ID, (+) Grade ID, (+) raise hands, (+) Visual resources, (+) Announcement views, (+) Parent School satisfaction, (+) Parent Answering Survey, (+) Student Absent Days.
[S12]	(+) Viewing Announcements, (+) Visited Resources, (+) Topic, (+) section ID, (+) raised hand in Class, (+) Discussion groups, (+) Place of Birth, (+) Gender, (+) Nationality, (+) Parent School Satisfaction, (+) Parent responsible, (+) Student Absence Days, (+) Educational Stage, (+) Parent Participation, (+) Semester, (+) Parent Answering Survey, and (+) class grade.
[S14]	(+) Final Mark, (+) Assignment Mark, (+) Tutorial Mark, (+) Test Mark, (+) Bonus Mark, (+) NBT Math, (+) Math Prerequisites, and (+) Math High School.
[S30]	Course, staff, and student data in the format: (+) username, (+) time, (+) description, (+) IP address, (+) origin, (+) event name, (+) event context, (+) affected user, and (+) components.
[S34]	(+) Quintile, (+) province, (+) gender, (+) financial assistance, (+) township school, (+) grade, (+) outcome, (+) discussion, (+) message, and (+) time.
[S35]	(+) Student Info, (+) courses, (+) student Registration, (+) assessments, (+) student Assessments, and (+) student VLE.

- [S9] used EDM models with student log data collected from an LMS for student academic performance prediction. The findings showed that student absence days, visited resources, announcements view, raised hands, parents answering a survey, discussion, and parent-school satisfaction were significant in students' academic performance predictions.
- [S12] used a clustering technique with data from an LMS to group the low-performing students into two groups for knowledge discovery. Their results revealed that low-level absentee students with parents who do not participate enough in their studies are more likely to perform poorly than those with parents who participate.
- A study by [S14] used EDM to identify academic factors, such as engagement variables and course assessments, at specific intervals in a semester and provide helpful recommendations at those intervals. Moreover, the study used students' prior academic information, such as high school grades, for student academic performance prediction. The findings revealed that the final mark, assignment mark, tutorial mark, test mark, bonus mark, National Benchmark Test (NBT) math mark, math prerequisites, and math high school were significant in predicting students' academic performance.
- A study by [S30] applied EDM on student logs obtained from Moodle LMS to identify the influence of LMS usage patterns in improving students' academic performance.

This study's findings indicated a significant relationship between using LMS resources and students' academic performance at a 0.01 significance level.

- [S34] examined features from an LMS system, such as personality, behaviour, and background, to predict students' academic performance using EDM methods. These e-learning behavioural factors include uploaded assignments, time spent working on courses, and participation in online group discussions. The findings showed that students' e-learning behaviours and personalities significantly influenced their academic performance more than their background.
- The study by [S35] developed ensemble EDM models for early and accurate prediction of at-risk students' performance using students' demographic and weekly Virtual Learning Environment (VLE) data. The results showed that demographic and weekly VLE data significantly influenced students' academic performance.

4.3.1.2 Data Collection Sources in Academic Performance Analysis

The researcher analysed the 41 selected articles regarding their data collection sources and dataset sizes in this section, as shown in Table A.5 in Appendix B. Additionally, the researcher identified three data collection techniques in the literature: Moodle learning management system (LMS) logs, students' information system (SIS), and survey. As shown in Figure 4.5, most researchers have used data records and sources available through SIS, the most frequently used data collection technique cited in 35 studies (85%). The SIS included student information such as enrolment, background, and admission data. Moodle LMS logs were the least used data collection method, cited in six studies (15%). Researchers obtained Moodle LMS logs containing student information, including engagement and interaction data. Finally, the survey method was the second most used data collection method, cited in four studies (10%). The survey was a questionnaire handed directly to students to fill in. Some researchers used a mixed mode of data collection where they collected data from SIS and survey ([S6]) and SIS and LMS logs ([S9]) to further increase accuracy, even though these data sources are easy to access and provide reasonable accuracy in a prediction model. Since some studies used more than one data collection source, the total percentage of the data collection technique is not equal to 100%.

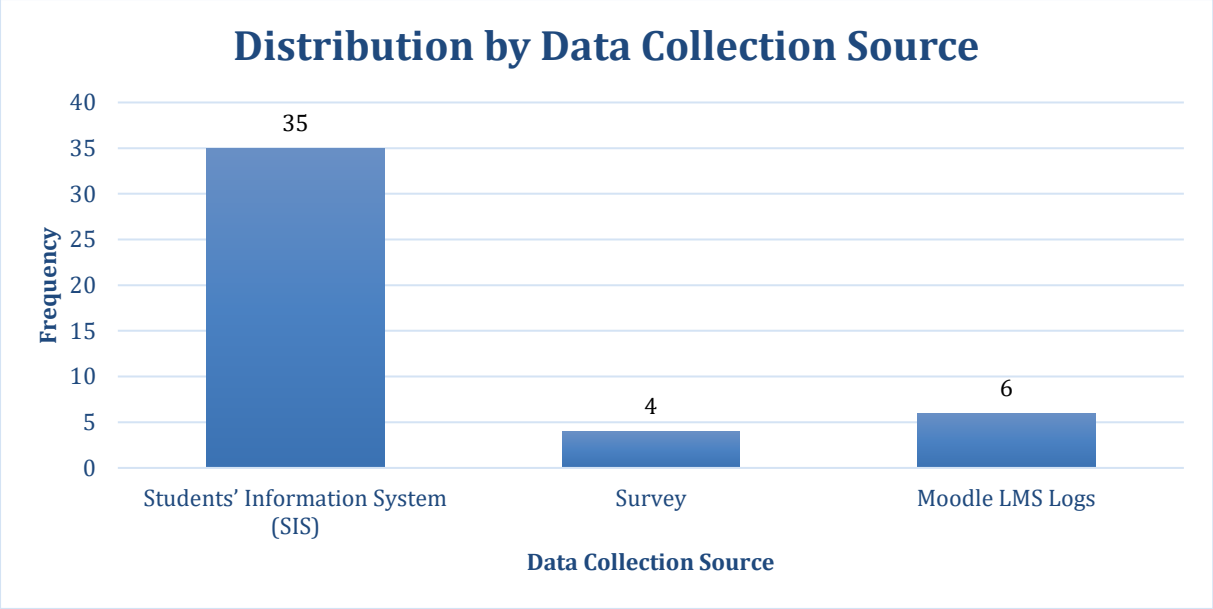

Figure 4.5: Distribution by data collection source

EDM researchers widely use Students Information Systems (SIS) in retrieving previous and current class performance and student demographics such as ethnicity, gender, and age. However, socio-economic factors may not be readily available. In such cases, researchers can acquire these factors from students through surveys. EDM researchers can also extract the instructor attributes and prior knowledge data from the survey method. Finally, students' e-learning behaviour can be gathered from Moodle LMS systems, as shown in Table 4.10.

Table 4.10: EDM Data Collection Sources

Data	Source
Previous and current class performance	SIS
Student Demographics	SIS, Survey
Socio-economic factors	SIS, Survey
E-Learning Behaviours	Moodle LMS Logs

Figure 4.6 shows a bar graph of the frequently used studies focused on their dataset size. Many of the studies (32 studies) used a dataset with less than 5000 records, and only six studies used datasets with greater than or equal to 5000 records for training predictive models ([S16], [S24], [S27], [S35], [S37]). However, most EDM algorithms in the literature need enormous datasets to perform well and provide accurate predictions (Alturki *et al.*, 2022). These results have introduced a potential research gap. Future studies can focus on identifying the best low-cost data collection methods that do not require too much effort to obtain the necessary data. Additionally, future studies need to focus on identifying the best methods of increasing dataset

sizes in data collection so that the EDM models can achieve greater accuracy and results can then be more generalizable.

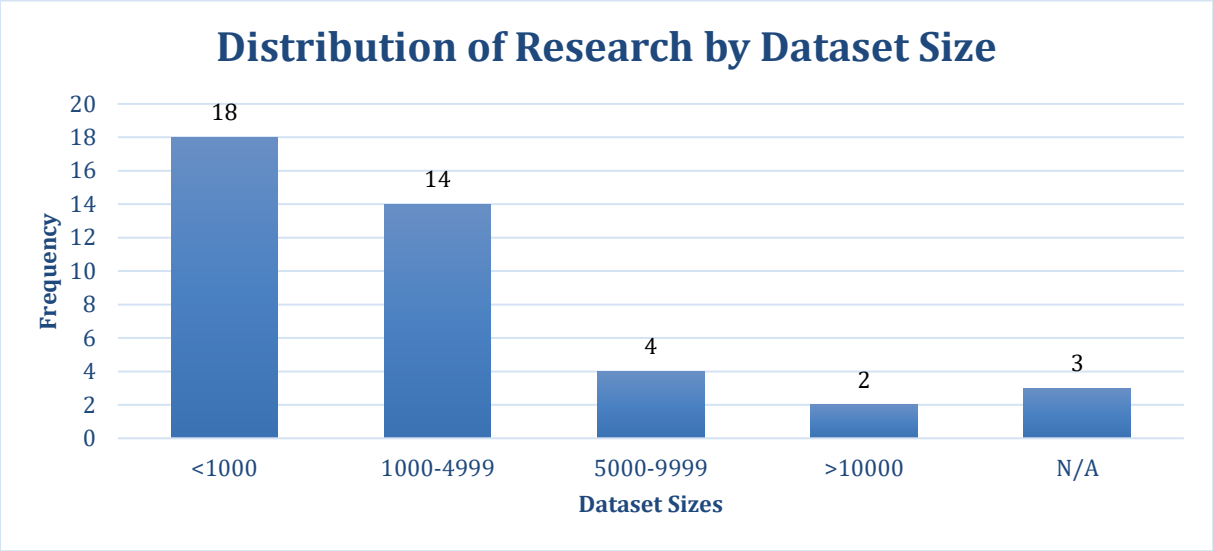

Figure 4.6: Distribution of research by Dataset size

4.4 Summary

This chapter presented the bibliometric analysis and the first part of the thematic analysis. This study identified the frequently studied factors influencing students' academic performance prediction and identified the primary data collection sources used by researchers to extract the influencing factors in the academic performance analysis theme.

Chapter 5: Analysis Part 2

5.1 Introduction

This chapter is a continuation of the thematic analysis from the previous chapter (four), which seeks to report the results of the analysis to address the research questions and elaborate on the findings from the data extracted from this SLR research study. Moreover, the previous chapter focused on answering Sub-Research Question One. As the analysis enters its focal point, it is necessary to reiterate the core research questions guiding this study:

Main-Research Question: How effective is EDM in predicting sub-Saharan African undergraduate STEM students' academic performance?

Sub-Research Question One: What key input factors are used in EDM to predict sub-Saharan African undergraduate STEM students' academic performance?

Sub-Research Question Two: What are the key student involvement factors used in EDM to predict sub-Saharan African undergraduate STEM students' academic performance?

Sub-Research Question Three: What are EDM's most frequently used tools and algorithms to predict sub-Saharan African undergraduate STEM students' academic performance?

Sub-Research Question Four: To what extent is EDM effective in predicting sub-Saharan African undergraduate STEM students' academic performance?

This chapter focuses on academic performance prediction. Moreover, the chapter consolidates academic performance analysis and prediction by discussing the challenges arising from these aspects. Finally, the chapter explores the relationship between the HEDM framework and the thematic analysis towards theory-building and enabling the discovery of new knowledge. The chapter is designed to extend beyond the immediate confines of the study, contributing to the broader understanding of how EDM can influence academic performance in the sub-Saharan African STEM educational landscape. The subsequent sections unfold systematically, each dedicated to a specific research question, offering in-depth analyses that aim not only to answer the posed questions but to deepen the understanding of the underlying phenomena.

5.2 Academic Performance Prediction

This section establishes guidelines in the literature on the various steps to take while using EDM for students' academic performance prediction, including all decisions taken at different stages of the process and best practices. It consists of data pre-processing, tools, methods, evaluation, and determinants of academic performance prediction effectiveness, which are covered in the following subsections.

5.2.1 Data Preprocessing Methods

Data collected from various sources may be unsuitable for analysis and modelling in its original form due to inconsistent, duplicate missing, miscoded, and incorrect data. Hence, data in its original form requires cleaning, transformation, selection, and segmentation. In the context of EDM and its application to predict the academic performance of sub-Saharan African undergraduate STEM students, a pivotal consideration is the identification of key student involvement factors. This inquiry aligns with sub-research question two: *What are the key student involvement factors used in EDM to predict sub-Saharan African undergraduate STEM students' academic performance?* The process of data preprocessing, consisting of techniques such as feature selection and data segmentation, plays a crucial role in extracting relevant insights from the vast datasets involved. By linking the exploration of student involvement factors to data preprocessing methods, the researcher aims to determine the most influential features which can be used optimize an EDM model's capacity to predict academic outcomes accurately. Data pre-processing is a crucial step before the application of EDM, as attaining balanced data is commonly known to be problematic in EDM ([S36]). Hence, EDM researchers must reduce the number of features in their EDM models and include only the significant ones which can increase prediction accuracy.

Data cleaning

Data sources usually suffer from missing values, contain noises, and are inconsistent ([S11]). The data is considered missing when a value is not stored for it. Values at abnormal distances from others in a dataset are called outliers. Outliers and missing values are significant in EDM literature ([S21], [S35], [S36], [S39]). Hence, understanding how to handle them while preserving or improving the prediction quality is essential. As shown in Table 5.1, there are two ways of dealing with missing values. The first is a listwise deletion, consisting of either deleting a column when there are too many missing values or the row when there are few missing values ([S9], [S36]). The second is imputation, and it consists of deriving the lost value

from the remaining data using methods such as mean, median, and mode ([S21], [S22], [S26], [S32], [S35]). Outlier data, also called anomalous data, can be identified easily through visual means, such as creating a stem, box plot, leaf plot, or histogram ([S9], [S14], [S20], [S22], [S23], [S26], [S29], [S35], [S36]). Once identified, outliers can be eliminated from the modelling dataset. Another strategy involves converting dataset values from numeric to categorical variables to improve prediction accuracy ([S6], [S7], [S16], [S35]).

Table 5.1: Data Cleaning methods

Strategy	Method	Implementation	Studies
Listwise Deletion	Column deletion	When there are too many missing values	[S9], [S36]
	Row deletion	When there are few missing values	
Imputation	Derivation using the mean, median, and mode	When there are missing values, such as grades or scores	[S21], [S22], [S26], [S32], [S35]

Data Transformation

Data Transformation consists of consolidating or transforming data to allow an efficient knowledge discovery process ([S36]). As shown in Table 5.2, new features can be created or engineered during the transformation process. Smoothing and feature engineering are primary strategies implemented in the transformation process ([S16]). Smoothing removes inconsistencies, outliers, and noise from the data to enable knowledge discovery ([S30]). Feature engineering consists of deriving new features from existing values ([S25], [S36]). For instance, GPA is a commonly used variable for predicting students' academic performance, reflecting students' averages each semester. Each student's GPA over several semesters could be in various states, such as experiencing an increase or a drop in performance. Hence, calculating differences in GPA through successive semesters is essential as it will bring new knowledge. Discretization transforms continuous variables into categorical variables, leading to standardized values for more effective and accurate predictions ([S16]). Discretization methods are classified as either unsupervised or supervised ([S36]). Unsupervised discretization divides features into equally sized containers. In contrast, supervised discretization consists of equal intervals and frequency binning, where continuous variables are transformed into definite ones without physically identifying the number of intervals or bins.

Table 5.2: Data Transformation methods

Strategy	Method	Implementation	Studies
Smoothing	Removes inconsistencies, outliers, and noise	To transform data collected, such as assessment grades, attendance, and CGPA	[S16], [S30]
Feature engineering	Deriving new features from existing values	Comparing students' GPAs over different semesters	[S25], [S36]
Discretization	Transforms continuous variables into categorical variables	Converting grades to their respective letter of grade to improve efficiency	[S16], [S36]

Feature selection

EDM researchers consider feature selection as part of data pre-processing ([S36]). This process eliminates features considered to be irrelevant or noisy and determines a subset of features with equal predictive ability ([S9]). An optimum feature subset improves processing efficiency and model accuracy (Saa *et al.*, 2019). These feature subsets also interest educational stakeholders, as discovering the optimum feature subsets can improve students' academic performance predictions and help develop focused interventions. Filter and wrapper are the most commonly used feature selection methods ([S14]).

Filter methods rank features in a pre-processed dataset to identify the high-ranking features for prediction ([S16]). On the other hand, in wrapper methods, the predictors are wrapped within a search algorithm that finds a feature subset that gives the highest prediction accuracy ([S11]). Furthermore, embedded methods involve feature selection in the training process before data is segmented into two sets for training and testing ([S14]). In this study, the literature reviewed showed three filter algorithms frequently used to rank features on their importance: Information Gain, Relief-F, and Gain Ratio.

Information Gain: The change in the entropy, uncertainty, or randomness in data due to the absence or presence of a feature is measured through Information Gain ([S1]). Information Gain is a commonly used ranking method because of its efficiency and relative ease of interpretation. It measures the dependency between the attributes and the labels ([S16]). A feature is significant when it has a high information gain value and not significant when it has a low value in rank ([S9]). Following are studies that reported information gain as being effective in ranking features, as shown in Table 5.3:

Table 5.3: Studies focusing on Information Gain

ID	Findings
[S9]	Behavioural features were the most influential factors in students' academic performance prediction.
[S11]	Average SSCE score, post-UTME score, JAMB score, secondary school type, weekly study time, sponsor type, sponsor support, sponsor income, sponsor qualification, secondary school area, years before admission, work and study, university accommodation, postgraduate degree, course interest, and sports activeness were the most determining factors in student academic performance prediction.
[S25]	• The findings show that biographical and individual attributes influence student attrition. • Pre-college factors were not significant in influencing student attrition.
[S28]	The results proved that the background, individual, and pre-university features were the most influential predictors of students' academic performance.
[S29]	The results showed the seven top-ranked features: age at first year, plan description, plan code, the year started, Matric Mathematics major, school quintile, and streamline.
[S33]	The significant factors were Age, Parental factors, Satisfaction Level, and Course Applied For.

- [S9] used information gain to rank and identify significant factors in building an efficient prediction model. The features significantly influencing academic performance predictions were classified into academic, demographic, and behavioural factors.
- [S11] found sixteen top-ranked features using the Information Gain method. These features were average SSCE score, post-UTME score, JAMB score, secondary school type, weekly study time, sponsor type, sponsor support, sponsor income, sponsor qualification, secondary school area, years before admission, work and study, university accommodation, postgraduate degree, course interest, and sports activeness.
- In a study by [S25], information gain revealed that biographical and individual factors significantly influenced predicting student attrition and deducing students into correct risk profiles. However, the pre-university features showed little or no influence on student attrition.
- [S28] used information gain to measure each feature's strength concerning the risk factor. The result of the study showed that pre-college, background, and individual features were ranked the highest.

- [S29] used entropy to reveal more robust features determining students' success. The results showed the seven top-ranked features: age at first year, plan description, plan code, the year started, Matric Mathematics major, school quintile, and streamline.
- [S33] used information gain to rank features such as course applied for, satisfaction level, age, and parental factors. The highest-ranked feature was the course applied for.

Gain Ratio: The gain ratio method improves upon information gain. The gain ratio method generates knowledge by segmenting the training data into a definite number of subsets in correspondence with the number of test results ([S11]). The feature with the highest gain ratio is taken as the split variable ([S37]). The gain ratio method is commonly used in the WEKA environment with the J48 decision tree algorithm, also known as the C4.5 ([S33]). Following are studies that reported gain ratio as being effective in ranking features, as shown in Table 5.4:

Table 5.4: Studies focusing on Gain Ratio

ID	Findings
[S11]	Average SSCE score, post-UTME score, JAMB score, secondary school type, weekly study time, sponsor type, sponsor support, sponsor income, sponsor qualification, secondary school area, years before admission, work and study, university accommodation, postgraduate degree, course interest, and sports activeness were the most determining factors in student academic performance prediction.
[S37]	The most significant predictors used included preparatory school GPA, entrance Examination results, first year first-semester results, and field choice interest in predicting students' academic performance.

- [S11] used the Gain ratio filter algorithm and identified the best four attributes: average SSCE score, secondary school type, sponsor qualification, and weekly study time. At the same time, the least four features: primary school attended, sex, marital status, and family size.
- [S37] used Gain Ratio and found the Preparatory School CGPA to be a determining factor for students' academic performance.

Relief-F: Relief-F is a refinement of the Relief method. Although the Relief method can identify the most appropriate attributes, it is restricted when faced with incomplete data. It may not work with more than two classes ([S11]). The relief-F method aims to discover attributes whose values differ against instances in the same class ([S9]). Following are studies that reported information gain as being effective in ranking features, as shown in Table 5.5:

Table 5.5: Studies focusing on Relief-f

ID	Findings
[S11]	Average SSCE score, post-UTME score, JAMB score, secondary school type, weekly study time, sponsor type, sponsor support, sponsor income, sponsor qualification, secondary school area, years before admission, work and study, university accommodation, postgraduate degree, course interest, and sports activeness were the most determining factors in student academic performance prediction.

- [S11] found that the best features of the Relief-F method are weekly study time, sponsor type, sponsor qualification, and family size. In contrast, the least four attributes are course from marital status, smartphone ownership, primary school attended, and JAMB score.

Data Segmentation

EDM models can generalize training data effectively. At the same time, it can struggle to perform efficiently on unseen or new data; this is called overfitting ([S11]). EDM researchers handle overfitting through data segmentation, which involves splitting datasets into two segments, one to train the model and the other to test the model ([S9]). The splitting method used generally in literature is based on the majority-minority principle. The first segment contains the majority of data for training, and the second includes the minority data to test a model's performance in generalizing unseen data. However, there is a study that splits training and test data evenly. This is not recommended in the literature as it increases the risk of reducing the model's accuracy ([S25]). Following are studies that reported the various segmentation of data, as shown in Table 5.6:

Table 5.6: Data Segmentation

Data Segmentation Principle (Training: Testing)	Studies	Frequency
70:30	[S37], [S3], [S9], [S41]	4
66:33	[S2], [S24], [S16]	3
80:20	[S20], [S27], [S10]	3
60:20:20	[S18], [S26]	2
75:25	[S25], [S22]	2
50:50	[S38]	1
60:40	[S7]	1

- [S2] split the dataset into a 66:33 ratio, where the model was trained with sixty per cent of the data and tested using the remaining thirty-three per cent.
- [S3] used stratified sampling and segmented their dataset into 70:30 principle; the model was trained with 70% of the data and tested with the remaining 30% of data. The remaining 30% was used to test the predictive algorithms' performance.
- [S7] split their dataset on a 60:40 ratio, where the model was trained with sixty per cent of the data and tested using the remaining forty per cent.
- [S9] segmented the data into the 70:30 principle; the model was trained with 70% of the data and tested with the remaining 30%. Three sampling techniques were employed to address the imbalanced dataset class problem to attain a predictive model with optimal performance.
- [S10] segmented their data into an 80:20 ratio, where the model was trained with 80% of the data and tested with the remaining 20%. The data was trained using ten-fold cross-validation.
- [S16] segmented the data into a 66:33 ratio and conducted 30 experiments with four classification algorithms using ten-fold cross-validation to train sixty-six per cent of the data and test the remaining thirty-three per cent.
- In contrast, [S18] divided their dataset into the 60:20:20 ratio. The model was trained with sixty per cent of the data, then validated the model using twenty per cent of the data, and tested the model with the remaining twenty per cent of the data.
- [S20] segmented the data into the 80:20 ratio, where the model was trained with 80% of the data and tested with the remaining 20%.
- [S22] used the ten-fold cross-validation to train 75% of their dataset as a training set and employed 25% to test model performance.
- [S24] segmented the data into a 66:33 ratio, where the model was trained with sixty per cent of the data and tested using the remaining thirty-three per cent.
- [S25] segmented the data into two segments; the training set was 75%, whereas the testing set was 25%.
- [S26] divided their dataset into the 60:20:20 ratio. The model was trained with sixty per cent of the data, then validated the model using twenty per cent of the data, and tested the model with the remaining twenty per cent of the data.
- [S27] split the dataset into two parts, where the model was trained with 80% of the data and tested with the remaining 20%.

- [S37] classified the dataset into the 70:30 principle; the model was trained with 70% of the data and tested with the remaining 30%. Then, the model's accuracy was measured using the test data by examining how many instances were correctly predicted or classified by the model.
- In a contrary study, [S38] partitioned the sample data into a 50:50 ratio, where the model was trained with fifty per cent of the data and tested with the remaining fifty per cent.
- [S41] randomly split their dataset into two parts. The training data had 522 observations of 70 %, and the test data had 225 observations of 30 %.

5.2.2 Tools

This section compares EDM tools used for predicting students' academic performance. Many tools exist for building models to predict students' academic performance. These tools make the process of data cleansing, feature selection, data analysis, and building prediction models easy. EDM tools commonly used in literature for students' academic performance prediction are presented in this section, such as WEKA, Python, R studio, Microsoft Excel, etc.

WEKA: WEKA (Waikato Environment for Knowledge Analysis) is a collection of EDM methods that contain tools for data preparation, clustering, regression, classification, visualization, and association rule mining ([S11]). WEKA is an EDM tool for developing prediction models. WEKA can be used through a graphical or command-line interface to build EDM algorithms ([S40]). WEKA enables researchers to use the default functions or to develop custom Java algorithms ([S21]). This tool solves all data processing, feature selection, clustering, regression, and classification problems. WEKA software is freely available and open source ([S33]).

Python: The Python programming language is used by EDM researchers for implementing EDM algorithms ([S9]). Python is a freely available open-source program for personal and commercial use. Python has many modules that consist of reusable code blocks that provide various functionalities for data analytics, such as the Pandas data pre-processing library, Scikit-learn machine learning library, and Plotly data visualisation library ([S24]).

R Studio: R is a freely distributed and open-source programming tool with multiple built-in methods for computations, data manipulation, and visualisation ([S17]). R is widely accepted for data analytics and science and has overwhelming strengths in nondeterministic calculations, scoping, computational efficiency, and resource management ([S23]).

Microsoft Excel: Various researchers frequently used Excel to pre-process data collected from HEIs in different formats ([S40]). The data was converted by researchers and stored in an Excel workbook ([S20]). Researchers matched the features in the Excel workbooks to those on the various sources and defined them in the appropriate EDM format before knowledge discovery.

Some research studies used other EDM tools, such as MATrix LABoratory (MATLAB) ([S3], [S18], [S25]); Statistical Package for the Social Sciences (SPSS) ([S15], [S21], [S27], [S30]); Konstanz Information Miner (KNIME) ([S4]); Orange ([S4]); STATA ([S32]). Following are studies that reported the other EDM tools:

- [S3] used MATLAB to develop pure quadratic and linear regression algorithms to analyse students' academic performance; these models had an accuracy of 85.89%.
- [S4] used Orange software to conduct EDM analyses. The findings showed that the multiple regression method was the highest-performing model, with 53.2% accuracy. The findings also showed that the pre-admission grades alone are not significant in student academic performance prediction.
- [S15] used SPSS to determine the significant features that could be used to build the model. The results showed five critical features: lecture attendance, self-study, competent lecturer, first semester's GPA, and average matric results.
- [S18] used MATLAB to evaluate students' academic performance using performance metrics such as accuracy, receiver operating characteristics, and mean square error. The findings revealed that the Generalized Regression Neural Network method performed best, even though Multilayer Perceptron had the highest performance of 75% accuracy.

Figure 5.1 summarizes four frequent EDM techniques for students' academic performance prediction. WEKA had 16 studies (39%), Microsoft Excel had four studies (10%), KNIME had one study (2%), MATLAB had three studies (7%), Orange had one study (2%), Python had seven studies (17%), R studio had eight studies (20%), SPSS had four studies (10%), and STATA had one study (2%).

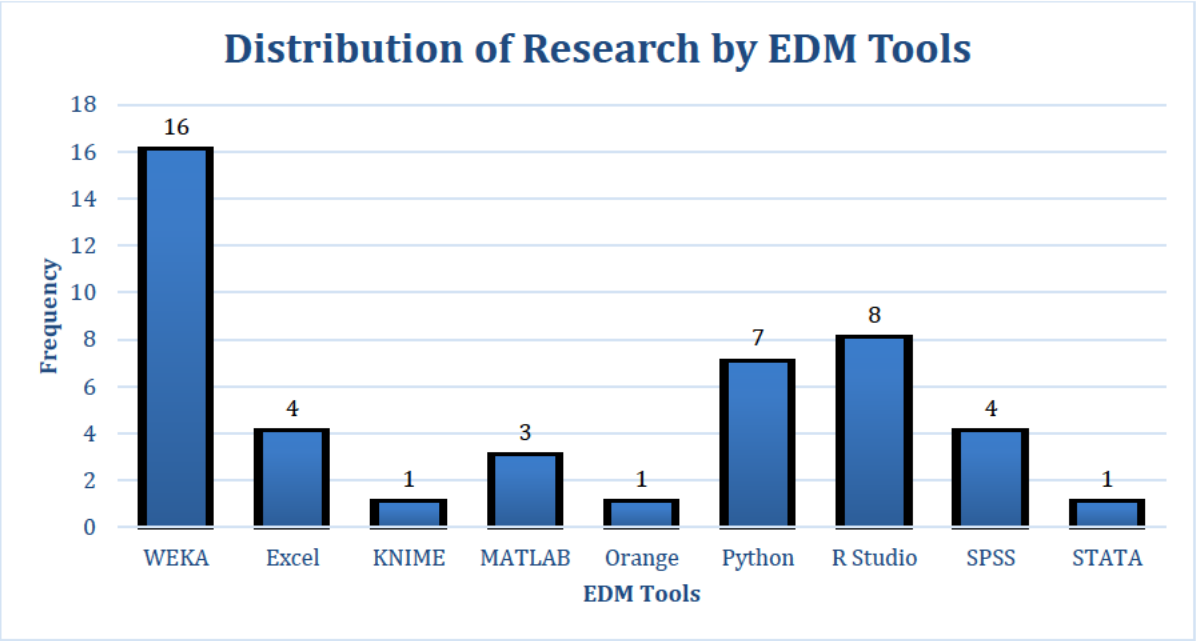

Figure 5.1: Distribution of Research by EDM tools used

WEKA is the most frequently used EDM tool, and this finding aligns with previous literature in this field (Chaka, 2021). The primary reason for the frequent use of WEKA may be the various built-in tools for data pre-processing, clustering, regression, visualisation, classification, and easy accessibility and user-friendliness.

5.2.3 EDM Methods

In the pursuit of understanding the frequently used tools and algorithms in EDM for predicting sub-Saharan African undergraduate STEM students' academic performance, the fundamental question emerges: *What are EDM's most frequently used tools and algorithms to predict sub-Saharan African undergraduate STEM students' academic performance?* This inquiry serves as the guiding force behind the exploration within the sections on tools and EDM methods, aiming to unveil the specific instruments and methods shaping the predictive landscape in the context of sub-Saharan African undergraduate STEM education.

EDM Researchers have implemented various classification, clustering, and regression approaches to extract meaningful knowledge from education datasets. The primary EDM approaches used in most literature are classification, regression, and the least used approach is clustering. Figure 5.2 illustrates the distribution of research of the three EDM approaches described above. The primary EDM approach used in the selected studies is classification. It was found that 38 studies (93%) used the classification approach for predicting students'

academic performance. Only 22 studies (54%) have used the regression approach. Lastly, only three studies (7%) have used clustering.

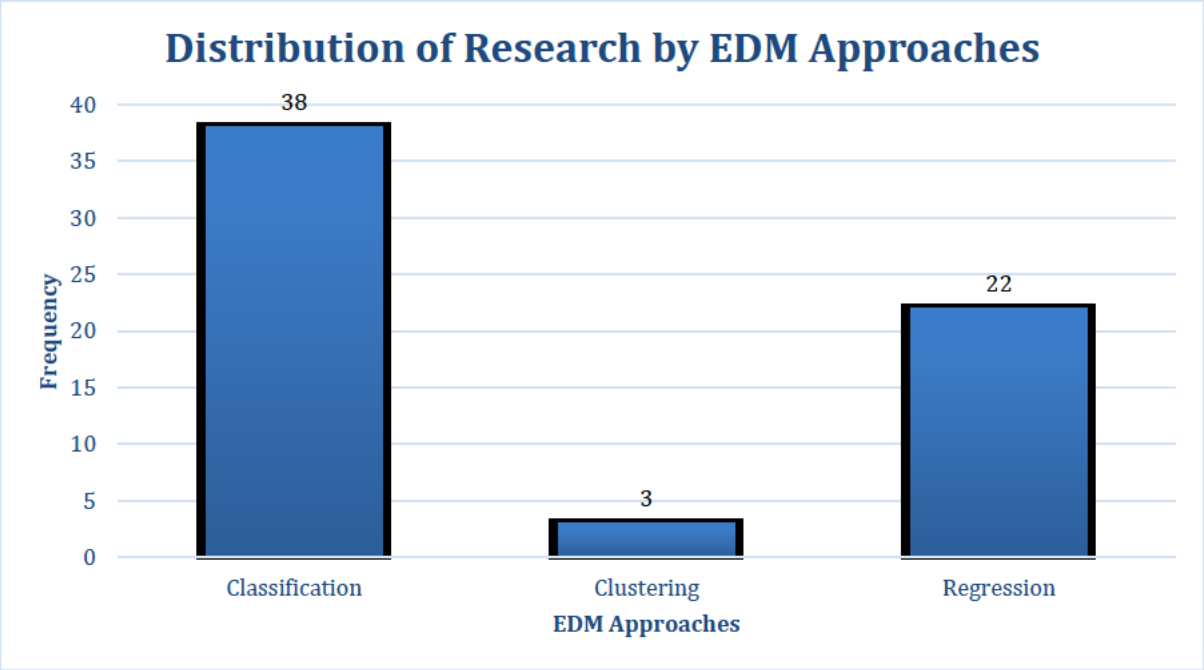

Figure 5.2: Distribution of Research by EDM approaches

Some studies used classification and clustering, which helped determine how many student groups were in the dataset and extracted particular attributes from each cluster of students ([S30]). Since some studies used more than one approach, the total percentage of EDM approaches is not equal to 100%. Classification is the most used approach in students' academic performance prediction. This finding aligns with a previous SLR by Saa *et al.* (2019), which found that classification was the most typical EDM approach used by EDM researchers.

Classification

Classification is a leading EDM approach for student academic performance prediction. EDM researchers frequently apply this approach to classify datasets into predefined classes (Khan & Ghosh, 2021). Figure 5.3 presents the frequently used methods in the selected literature. Decision Trees were used in 29 (71%) studies. Artificial Neural Networks were used in 18 (44%) studies. Naïve Bayes was used in 18 (44%) studies. Random Forest was used in 12 (29%) studies. Support Vector Machine was used in 11 (27%) studies. The instance-Based algorithm was used in seven (17%) studies. Tree Ensemble methods were used in six (15%) studies. Finally, other classification methods were used in nine (22%) studies.

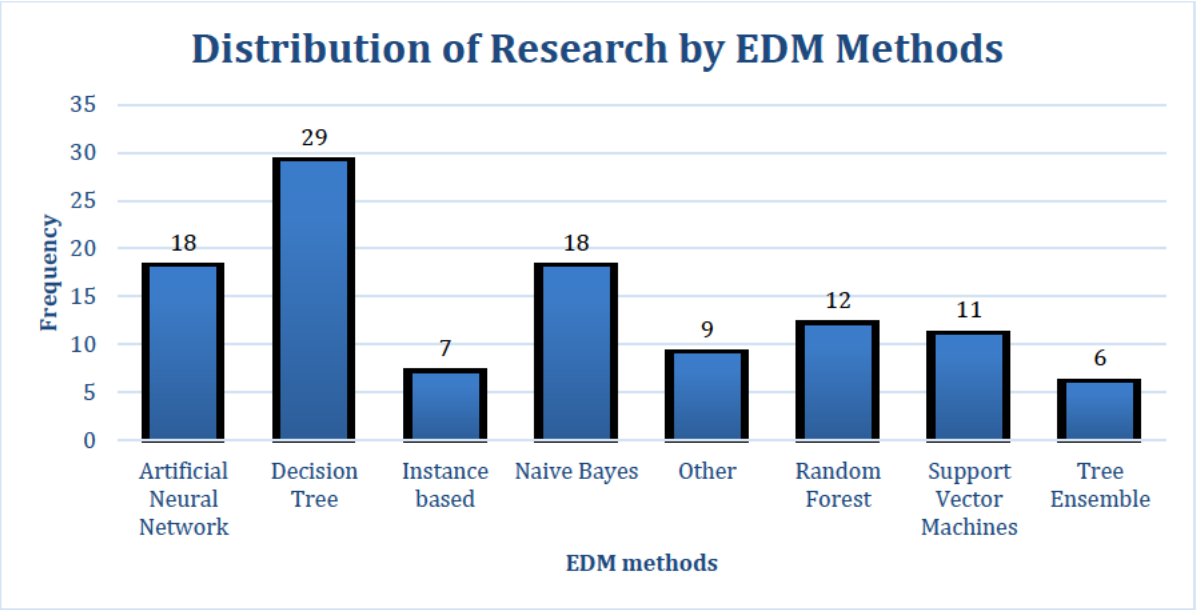

Figure 5.3: Distribution of Research by EDM Methods

As observed in Figure 5.3, the most frequent EDM methods are Decision Trees, Artificial Neural Networks, Support Vector Machines, Random Forests, and Naïve Bayes. This finding aligns with the results of a prior SLR study, where the Decision Trees model was the most frequent EDM technique used (Saa *et al.*, 2019). Although the Decision Trees method was the most frequently used EDM method, it was not most accurate EDM method. The Decision Trees method was the third best performing EDM method, as shown in Table 5.7.

The best performing EDM method was Random Forest, with an average accuracy of 93%. The high accuracy among these studies may have been caused by either the use of good feature selection to increase performance or the use of overfitted or imbalanced datasets. Imbalanced datasets are a common issue in literature which is ignored by most literature. Hence, this is a possible gap that may need to be addressed in future EDM work. Even though the Decision Trees method has been outperformed, most researchers prefer using it for their student academic performance predictions. Decision Trees are a superior choice for student performance analysis due to being highly accurate and comprehensible (Khan & Ghosh, 2021).

Table 5.7: Average accuracy of the best-performing EDM methods

EDM Method	Predictors	Studies	Average Accuracy
Random Forest	• Previous Grades and Current Class Performance • Demographics • Socio-economic data • E-Learning Behaviours	[S9], [S22], [S25], [S26], [S28], [S29], [S35]	93%
Artificial Neural Networks	• Previous Grades and Current Class Performance • Demographics • Socio-economic data	[S11], [S20], [S23], [S27]	89%
Decision Trees	• Previous Grades and Current Class Performance • Demographics • Socio-economic data • E-Learning Behaviours	[S1], [S2], [S4], [S5], [S6], [S7], [S8], [S14], [S16], [S30], [S31], [S33], [S34], [S36], [S37], [S39], [S40]	82%
Regression	• Previous Grades and Current Class Performance • Demographics • Socio-economic data	[S3], [S13], [S21], [S38], [S41]	80%
Support Vector Machine	• Previous Grades and Current Class Performance • Demographics	[S15], [S24]	80%
Naïve Bayes	• Previous Grades and Current Class Performance • Demographics • Socio-economic data	[S10]	66%

Decision Tree: Decision Trees use training data to discover the best areas for data segmentation to lower the cost metric ([S9]). The Decision Tree algorithm is frequent because it is highly interpretable and needs less data preparation, differing from other EDM models ([S14]). The Decision Tree algorithm also provides rules that are easily interpreted and function well with nominal and numeric variables ([S39]). Decision Tree models break down the decision-making process into more straightforward decisions, with classifications that are more easily understood, unlike other EDM algorithms ([S34]). There are different algorithms to construct a tree, including J48, Iterative Dichotomiser 3 (ID3), C4.5, and Random Forests, the selected

models used in the literature. The following are studies that commonly used the Decision tree method:

- [S1] employed EDM to determine the features significantly influencing science students' academic performance. The findings showed that the features that significantly affected students' academic performance were the individual factors, while the pre-university attributes had the least influence. The models applied in the study showed that these features can predict students' academic performance with 70% to 80% accuracy. The J48 Decision tree algorithm emerged as the best-performing algorithm with high precision, high recall, and high accuracy overall.
- [S2] used the J48 Decision Trees classification algorithm for student academic performance prediction. The findings revealed that the Decision Trees method accurately predicted students' academic performance with 50% precision.
- [S5] developed a predictive model to predict students' academic performance. The ID3 Decision Trees algorithm was more accurate than the C4.5 algorithm in student academic performance prediction.
- [S7] used EDM classification methods to predict students' academic performance to identify the correlation between the entry and final grades. The findings showed a weak correlation between the best six GPAs and the final CGPA. The J48 Decision Tree was the best-performing model with 98.9011% accuracy.
- [S16] developed a predictive model using four classification models: Rule Induction, Naïve Bayes, and J48 Decision Tree techniques, to determine undergraduate students' status in HEIs. The findings revealed that the J48 Decision Tree model attained the highest performance, with 91.40% accuracy.
- [S21] implemented the Decision Trees algorithm to determine students' success or failure. The results revealed that the Decision Tree algorithm had an accuracy of 90%.
- [S31] explored the impact of previous computer experience on students' academic performance using EDM methods, including Naïve Bayes, Logistic Regression, and Decision Trees. The findings showed that the Decision Trees model had the highest accuracy compared to the Naïve Bayes and Logistic Regression models.
- [S33] predicted students' academic performance using Classification and Regression Trees (CART), C4.5, and ID3. The findings revealed that C4.5 Decision Trees was the highest-performing algorithm, with 98.3% accuracy.

- [S36] implemented EDM algorithms to predict students' academic performance using physics, mathematics, age, ethnicity, and school province features. The Decision Trees was the highest-performing model, with 65.86% accuracy.
- [S37] developed an EDM model to predict first-year students' academic performance. The findings revealed that the C4.5 Decision Tree was the highest-performing method, with 81.4% accuracy.
- [S39] used the Decision Trees model for student academic performance prediction. The results revealed that the Decision Trees model had a very high accuracy of 95.24%.

Artificial Neural Networks: Artificial Neural Networks simulate the human brain's nervous system, which uses neurons to send information to other neurons through directed connections ([S7]). The most common Artificial Neural Network algorithm used in EDM is the multilayer perceptron ([S8]). The multilayer perceptron artificial neural network feeds information forward and comprises the input, hidden, and output layers ([S9]). The first layer is the input value for the features. The hidden layer contains one or more levels that specify the mathematical function. Finally, the output layer consists of the predicted outcome. The following shows studies that had Artificial Neural Networks as the best-performing model:

- [S6] uses socio-economic and student data and their final grades to build EDM algorithms such as Artificial Neural Networks, Decision Trees, and Bayesian Network classification tasks for predicting student competence in computer programming. The findings indicated that the best model was the Multilayer Perceptron, with an accuracy of 80.30%.
- [S11] used EDM algorithms to classify the low-performing students' dataset through evaluation methods such as Specificity, Recall, Root Mean Square Error (RMSE), Kappa, F-Measure, and AUC-ROC. The findings revealed that the Multilayer Perceptron was the highest-performing method, with an accuracy of 98%.
- [S18] developed Artificial Neural Network models for predicting the students' GPA. The results revealed that the Generalized Regression Neural Network model was the highest-performing algorithm with 95% accuracy.
- [S23] used EDM methods such as Linear Regression, Artificial Neural Networks, Random Forests, Support Vector Machines, Naïve Bayes, and eXtreme Gradient Boost to predict students' academic performance. The findings revealed that Artificial Neural Networks are the best-performing algorithm.

- [S27] developed predictive algorithms to predict students' academic performance and identify strategies to manage performance drivers. The findings revealed that the best-performing algorithm was Artificial Neural Networks, with a reasonable accuracy of 83%.

Naïve Bayes: The Naïve Bayes model is a probabilistic approach with solid independent assumptions between variables ([S40]). The Naïve Bayes model is effective due to its high capacity to handle missing values, resilience to noise, low variance, and quick processing times compared to most EDM models ([S31]). Although the Naïve Bayes was used in 18 studies, it did not outperform the other classification algorithms except in one study. Generally, the Naïve Bayes algorithm achieved an accuracy of above 70% in most studies it was reported. However, it was the least-performing algorithm. The following show studies that had Naïve Bayes as the best-performing model:

- [S6] built EDM models for students' academic performance prediction to examine features that mainly influenced students' proficiency in programming. The findings indicated that the Naive Bayes algorithm had a lower performance than other algorithms, whose overall accuracy was 64.09%.
- [S10] adopted an EDM model to analyse data on campus transfers to predict and automate the acceptance or rejection of transfer requests process. The findings revealed that the Multinomial Naïve Bayes was the best-performing algorithm, with an AUC performance of 74%.
- [S28] used EDM to predict student performance in a particular program. The findings indicated that the Naive Bayes algorithm performed more poorly than others, even though the model obtained an overall accuracy of 78%.

Support Vector Machine: The algorithms are supervised and implemented in classification and regression problems ([S25]). A hyperplane in Support Vector Machines is a linear line that separates various classes ([S9]). There can be more than one hyperplane on Support Vector Machines in correspondence with the attributes used in the training set. Support vectors are the areas most related to the hyperplane. The primary goal of Support Vector Machines is to identify the optimum hyperplane to maximize the gap between hyperplanes and support vectors ([S29]). Support Vector machines are well-suited for small datasets (Zohair & Mahmoud, 2019). The following show studies that had Support Vector Machines as the highest-performing algorithm:

- [S15] used Decision Trees, Support Vector Machines, and Bayesian Networks algorithms for student academic performance prediction. The best-performing algorithm was Support Vector Machines. The model accurately predicted 72.87% of students' academic performance before the year-end examinations' outcome.
- [S24] used the Support Vector Machines, Naïve Bayes, and Decision Trees methods to predict undergraduate science students' academic performance. The findings showed that Support Vector Machines were the highest-performing algorithm with 87% accuracy.

Random Forest: This model generates a variety of decision trees using a subset of the available features and returns the mode predicted class of each tree for classification or the average for regression ([S8]). Random Forests involve randomly generated Decision Trees combined to improve the reliability and effectiveness of the algorithm and avoid over-fitting issues that are more likely to occur in Decision Trees ([S9]). Random Forest algorithms are refined Decision Trees; hence, they usually perform better than Decision Trees ([S1]). In Random Forests, the training datasets are randomly selected with substitution to create various subsets to train individual decision trees ([S22]). Using training datasets with substitution refers to bagging, where each bootstrap sample has intersecting values in the original dataset with unique values, which introduces diversity, reduces overfitting in the ensemble of decision trees, and makes random forests more accurate and robust. The following show studies that had Random Forest as the model that performed higher than other models:

- [S9] used EDM models on students' log data obtained from an LMS for students' academic performance prediction. The Random Forests performance surpassed other EDM models with an accuracy value of 91%.
- [S22] applied EDM to identify students more likely to drop out. The findings revealed that Random Forests was the highest-performing algorithm, with 94.14% accuracy.
- [S25] used EDM models to discover vulnerable students early and provide them with support through effective intervention, positively influencing their academic performance. The Random Forest model was the best-performing algorithm, with an accuracy of 85%.
- [S26] used EDM to predict poor-performing mathematics students' academic performance. The findings showed that Random Forests was the highest-performing algorithm, with an accuracy of 99%.

- [S28] used predictive models such as the K-Star Algorithm, Random Forest, J-48 Decision Tree, Logistic Regression, Naive Bayes, and Multilayer Perceptron for student academic performance prediction. The findings indicated that Random Forests was the highest-performing model, obtaining an overall accuracy of 95%.
- [S29] used EDM algorithms to predict students' academic performance in a program's first three years. The findings indicated that Random Forests was the highest-performing algorithm, with an accuracy of 95.45%.

Other classification methods appeared in the analysis. However, they were not significant as they did not outperform any algorithms and were not frequently used by EDM researchers. These methods include Instance-based models such as k-Nearest Neighbour; Tree Ensemble such as Classification and Regression Trees; eXtreme Gradient Boosting Tree; Classification Tree; Discriminant Analysis; Discrete-time survival algorithm; Logistic Model Trees; Matrix Factorization; Singular Value Decomposition; Rule Induction; and ZeroR.

Regression

Regression is another common approach used in EDM for statistically analysing data, which defines the relationship between a dependent value and one or more independent values ([S15]). This approach is commonly used in EDM literature to establish or disprove the influence of teaching quality ([S41]). Various EDM research has used regression analysis for students' academic performance prediction. Figure 5.4 shows the various Regression methods used in the literature. The Logistic Regression method was the most used, and it was used in 13 (32%) studies. Linear Regression was the second most used Regression method, and it was used in four (10%) studies. Multiple Regression was used in three (7%) studies, while Multinomial Logistic Regression was used in two (5%) studies. Finally, Linear Logistic Regression and Ordinary Least Squares Linear Regression were used in only one (2%) study.

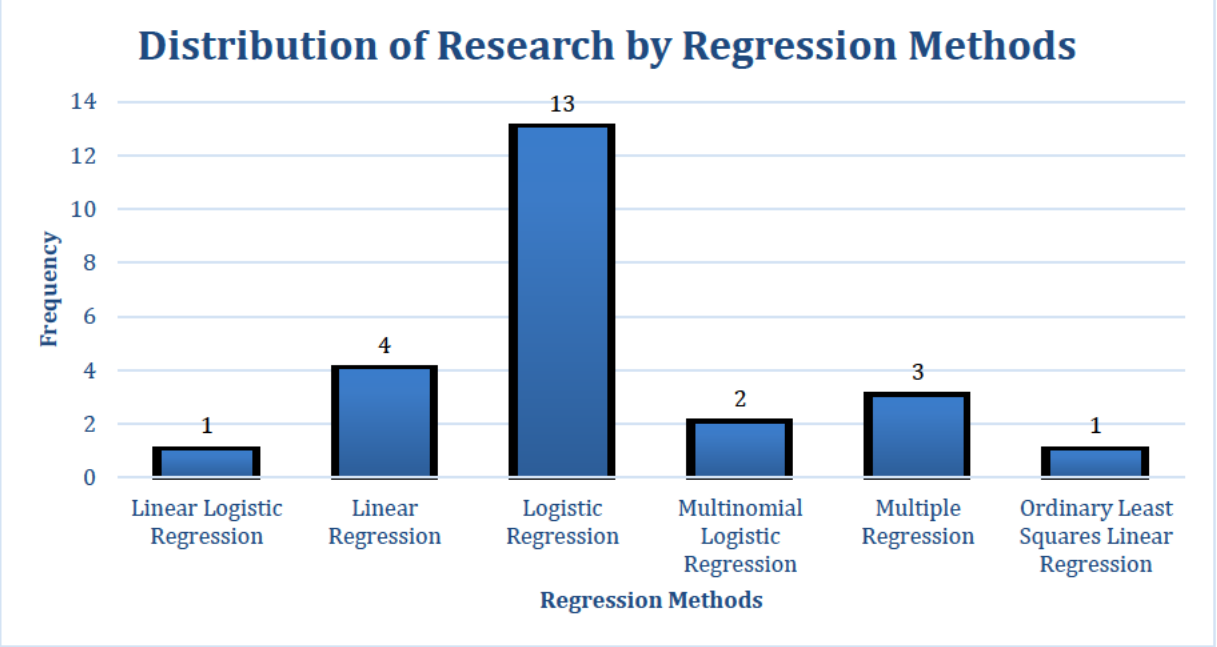

Figure 5.4: Distribution of Research by Regression Methods

Logistic regression: Logistic regression is a frequently used regression method in EDM due to its robustness in student academic performance prediction and useful in predicting model-dependent values with the help of independent values ([S9]). It consists of approximating the highest probability, and based on that probability, the observed data is the most suitable ([S14]). Logistic regression is highly accurate when independent variables are not perfectly correlated ([S8]). The flexibility and capacity to make evaluative assumptions regarding model terms are why logistic regression algorithms are commonly used ([S22]). The following show studies that had Logistic Regression as the best-performing model:

- [S3] applied EDM for student academic performance prediction in first three years of enrolment in a program. The results revealed that Logistic Regression was the best-performing method, with an accuracy of 89.15%.
- [S38] built a predictive model using Business Statistics student grades to predict at-risk students' academic performance. The findings revealed that the Logistic Regression had the best performance, with an accuracy of 80%.
- [S41] used a supervised EDM approach with enrolment data for predicting students' academic performance. The Logistics Regression algorithm was highly accurate, with an accuracy of 84.7%.

Linear Regression: The linear regression algorithm measures the correlation between two or more values based on an original dataset ([S13]). Additionally, linear regression often involves estimating and predicting unknown values from known values ([S23]). Moreover, a linear regression exists between values when the regression curve is a straight line ([S27]). The following show studies that had Linear Regression as the highest-performing algorithm:

- [S13] used the M5P Decision Trees and Linear Regression for programming students' academic performance prediction. The findings indicated that the variable-based Linear Regression model was the highest-performing algorithm. Additionally, the findings showed that the significant factors in predicting students' academic performance were student attendance, fearful perception of programming courses, the attitude of students and lecturers, student health, university facilities, and erratic power supply.

Multiple Regression: Multiple linear regression is an EDM algorithm representing the correlation between a dependent attribute and numerous independent variables ([S38]). It aims to examine how the differences in each independent variable explain differences in the dependent variable. The following show studies that had Multiple Regression as the best-performing model:

- [S4] developed predictive models to identify the effect of ethnicity on students' academic performance. The findings showed that multiple regression analysis was the best-performing method, with an accuracy of 53.2%. Additionally, over-sampling techniques were applied, and the accuracy increased to 79.8%.

Other Regression algorithms appeared in the literature. However, they were not significant and not reported regarding their impact on students' academic performance predictions. These methods included Multinomial Logistic Regression, Linear Logistic Regression, and Ordinary Least Squares Linear Regression.

Clustering

The Clustering approach divides the dataset into groups of similar things ([S12]). It is widely used in statistics and mathematics ([S5]). In EDM, clusters are related to hidden patterns, and searching for clusters results in the discovery of knowledge ([S21]). The two types of clustering commonly used are agglomerative hierarchical and k-means techniques ([S17], [S36]). Even though clustering is common in EDM literature, various researchers have experienced some

challenges. Conventional clustering methods are unsuitable when qualitative and quantitative attributes coexist ([S13]). Many educational datasets also contain quantitative and qualitative attributes, bringing further challenges. Many clustering approaches also need an input parameter to determine the resulting clusters, which is challenging for educators to specify this value ([S12]).

Regardless, EDM researchers still use this unsupervised method in students' academic performance prediction literature as it often leads to new knowledge. Association rule mining, Density-Based cluster, Hierarchical cluster, and k-means were used in research by one study each. The clustering method was the least used compared to Classification and Regression methods in this research. This presents another research gap. Future studies can focus on implementing unsupervised learning methods to discover new perspectives from different sample groups to discover the main factors that lead to low academic performance. The following are studies that used clustering methods for predicting students' academic performance:

- [S12] used a density-based clustering method to cluster the low-performing students. The result showed that the determining factors were demographic and behavioural features.
- [S17] adopted hierarchical cluster analysis to analyse students' academic performance datasets to discover an optimum amount of failed course clusters to extract interesting course-status associations. The Agglomerative Hierarchical cluster analysis had a coefficient of 92% with the five best clusters.
- [S36] used K-Means clustering to identify the relationship between student success and the given variables. The findings showed that determining factors of students' success based on the clusters were ethnicity, gender, religion, language, and province.

5.2.4 Evaluation Methods

Within the realm of evaluating EDM's effectiveness in predicting sub-Saharan African undergraduate STEM students' academic performance, the central question guiding this exploration is: *To what extent is EDM effective in predicting sub-Saharan African undergraduate STEM students' academic performance?* In this section dedicated to evaluation methods, the spotlight turns to two pivotal metrics: the confusion matrix and accuracy. These metrics are recognised for their key role in assessing prediction performance. The exploration

within this section aims to unravel the intricacies of applying these evaluation methods, shedding light on their significance in comprehensively understanding the predictive capabilities of EDM within the context of sub-Saharan African STEM education.

Various researchers in EDM literature predict students' academic performance by comparing actual and predicted variables ([S14]). The goal is the development of accurate, efficient, comprehensive algorithms that provide influential predictors at each interval in predictions. An appropriate model performance level is essential as it supports the reliability of the suggestions provided, whereas a poor model performance provides advice or outcomes that are not meaningful, which undermines the appropriateness of the suggestions ([S9]). The confusion matrix is frequently used in EDM for measuring the performance of classification methods. According to [S35], the confusion matrix is a 2x2 matrix in a binary classification problem with two class labels: students not at-risk and students at-risk. Table 5.8 shows the values of expected and actual class labels in the Confusion Matrix.

There are four crucial metrics used for evaluating EDM method performance in a confusion matrix ([S36]): True Positive (TP) refers to the number of successful students correctly predicted as successful. False Positive (FP) refers to the number of successful students predicted incorrectly as unsuccessful. False Negative (FN) refers to the number of unsuccessful students incorrectly predicted as successful. True Negative (TN) refers to the number of unsuccessful students correctly predicted as unsuccessful.

Table 5.8: Confusion Matrix

	Predicted	
Actual	TP	FN
	FP	TN

Different studies of students' academic performance prediction have used an accuracy evaluation technique to test model performance ([S25]). The results produced by prediction models are commonly evaluated and compared using accuracy, precision, recall, specificity, Area Under the Curve and Receiver Operating Characteristic (AUC-ROC) curve, F1 measure, and kappa statistics, as shown in Table 5.9.

Table 5.9: EDM evaluation methods

Evaluation Method	Studies	Frequency
Accuracy	[S1], [S3], [S4], [S5], [S7], [S8], [S9], [S10], [S11], [S14], [S15], [S16], [S18], [S20], [S21], [S22], [S24], [S25], [S28], [S29], [S31], [S33], [S35], [S37], [S38], [S39], [S40], [S41]	28
Precision	[S1], [S2], [S4], [S5], [S8], [S9], [S10], [S11], [S16], [S22], [S25], [S26], [S29], [S31], [S33], [S34], [S36]	17
Recall	[S1], [S2], [S4], [S5], [S6], [S8], [S9], [S10], [S11], [S14], [S16], [S20], [S22], [S25], [S26], [S29], [S31], [S33], [S34], [S35], [S36]	21
Specificity	[S5], [S11], [S14], [S20], [S26], [S33], [S35]	7
Kappa statistics	[S6], [S11], [S22], [S25], [S30], [S39]	6
Area Under the Curve and Receiver Operating Characteristic (AUC-ROC) curve	[S2], [S4], [S6], [S8], [S9], [S10], [S11], [S14], [S18], [S22], [S23], [S25], [S29], [S35], [S36]	15
F1 Measure	[S2], [S4], [S6], [S8], [S9], [S11], [S16], [S25], [S26], [S29], [S31], [S33], [S35], [S36], [S38]	15

Accuracy is a performance metric that can be used to measure an EDM classifier's performance or predictive capabilities. Accuracy consists of the proportion of the total number of correct predictions. According to [S35], accuracy scores are reliable when the dataset has balanced or nearly balanced classes. Accuracy as a performance measure becomes less valuable when there are highly imbalanced classes, and other performance metrics must be used for a more meaningful evaluation of the results.

The precision metric consists of the percentage of correctly classified prediction classes. The greater the precision value, the more accurate the predictions become. [S9] used the precision evaluation method to evaluate their models. They found that Multilayer perceptron, Decision Trees, and Random Forests had the highest precision values of 0.89 each.

The sensitivity or recall evaluation method consists of the number of correctly predicted classes and the average number of correctly predicted labels. [S9] evaluated their model performance using the recall method. The Random Forests and Multilayer Perceptron models had the most correctly predicted classes with 89% recall or sensitivity values.

Specificity is the proportion of negatives predicted correctly as unfavourable ([S14]). The sensitivity and specificity metrics measure the True Negatives and True Positives in datasets ([S20]). Hence, a perfect prediction model is a hundred per cent specific and sensitive, meaning the model predicted all the classes correctly ([S12]).

The Area Under the Curve and Receiver Operating Characteristic (AUC-ROC) curve is essential in students' academic performance prediction ([S8]). The AUC-ROC curve is a sensitivity function mapped against one specificity at a specific threshold ([S25]). Highly accurate prediction models have an AUC-ROC curve that passes through the top left corner, where specificity and sensitivity equal one ([S14]).

Kappa statistics consist of comparing the expected accuracy with the observed accuracy. The kappa statistic is commonly used to assess the model performance and identify class imbalances ([S25]). The kappa statistic compares predicted and actual classes' agreement ([S22]). There is no class agreement when the kappa statistic is close to zero, and the class assignment is random. However, there is total agreement when the kappa value is close to one, and the class assignment to the labels is not random.

The F1 Measure is another important metric for evaluating the efficiency of a prediction model, which is the harmonic mean of recall and precision ([S8]). [S9] and [S35] used the F1 measure to discover students' academic performance prediction model effectiveness. These studies found that the F1 score is also more valuable than accuracy when there is a class imbalance since it considers both the false positives and the false negatives. This provides for a balanced optimization between precision and recall.

5.2.5 Determinants of Academic Performance Prediction Effectiveness

In this section, which aims to uncover the determinants of EDM's effectiveness in predicting sub-Saharan African undergraduate STEM students' academic performance, a critical examination begins with the question: *How effective is EDM in predicting sub-Saharan African undergraduate STEM students' academic performance?* The focus is to determine the key factors influencing EDM's efficacy, where success prediction emerges as the more prevalent target than final grade prediction. Moreover, the inquiry extends across different levels, with degree-level prediction being the most frequent level of prediction, followed by course-level predictions, and year-level predictions being the least prominent predictions. This exploration within the section aims to dissect the complex landscape of EDM effectiveness, shedding light

on the varying degrees of success in predicting academic outcomes at different hierarchical levels within the context of sub-Saharan African STEM education.

Aim of prediction

The literature discussed in this analysis has different aims or objectives for predicting students' academic performance, classified into success or grade prediction. Some literature predicted students' academic performance in binary classes, such as success-failure or pass-fail. In contrast, others have predicted the students' academic performance to attain the actual score or final grade. Predicting students' academic performance in binary classes is more frequent than grade prediction in this analysis.

*Success prediction***:** Poor academic performance and high dropout rates are the most critical issues affecting HEIs ([S22]). Students' high dropout rates negatively affect them, the HEIs, and the economy ([S16]). Hence, preventing student dropout is essential to the success of HEIs. This prevention is done by predicting students' academic performance early. In this analysis, student success or failure predictions were found to be more frequent and effective in the literature. Most of the literature predicted students' academic performance in binary terms, including good-poor, pass-fail, and below-above ([S8], [S16], [S22], [S26]). However, some studies classify success or failure with more than two classes ([S1], [S23]). The following studies predicted students' academic performance in binary terms, as shown in Table 5.10:

Table 5.10: Studies focusing on success prediction

ID	Predictor used	Aim of prediction	Level of prediction	Best Performing EDM Algorithm	Effectiveness
[S1]	• (-) Previous Grades and Current Class Performance • (+) Demographics • (+) Socio-economic data	Pass-Fail	Degree Level	J48 Decision Tree	Accuracy: 80,4%
[S8]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Socio-economic data	Pass-Fail	Degree Level	Decision Trees	Accuracy: 97,3%
[S9]	• (+) Demographics	Pass-Fail	Degree Level	Random Forest	Accuracy: 91%

	• (+) Previous Grades and Current Class Performance • (+) E-learning Activities				
[S16]	• (+) Demographics • (+) Previous Grades and Current Class Performance	At-Risk (Pass-Fail-dropout)	Degree Level	J48 Decision Tree	Accuracy: 91,40%
[S22]	• (+) Previous Grades and Current Class Performance • (+) Demographics	At-Risk (Pass-Fail-dropout)	Degree Level	Random Forest	Accuracy: 94,14%
[S23]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Socio-economic data	Pass-Fail	Degree Level	Neural Networks	AUC (area under the curve): 86%
[S26]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Socio-economic data • (+) Instructor Attributes	Pass-Fail	Degree Level	Random Forest	Accuracy: 99%
[S32]	• (+) Previous Grades and Current Class Performance • (+) Demographics • (+) Socio-economic data	Pass-Fail	Degree Level	Discrete-time analysis	significance: < 0,05
[S35]	• (+) E-Learning Activities • (+) Previous Grades and Current Class Performance • (+) Demographics	At-Risk (Pass-Fail-dropout)	Degree Level	AdaBoost Classifier, LGBM Classifier, Random Forest Classifier, and XGB Classifier	Accuracy: 92%
[S36]	• (+) Demographics • (+) Previous Grades and Current Class Performance	Pass-Fail	Year Level	Decision Trees	Accuracy: 65,8%

	• (+) Socio-economic data				

- [S1] built a suggestion system advising students on their academic path in their courses. The target variable contained three possible values each student can take: QualMin, Qualified, and Failed. The findings showed that classification models applied were highly accurate in predicting whether a new student is likely to qualify or fail in three years or more than three years.
- [S8] developed classification models for identifying students who are more likely to drop out or fail. The F1 Score and the ROC AUC curve analysis showed that EDM can accurately predict at-risk students early.
- [S9] used EDM models with student log data collected from an LMS for predicting students' success or failure. Random Forests were the best-performing method, with an accuracy value of 91%.
- [S16] developed an EDM predictive model to predict undergraduate students' attrition or retention in HEIs. The results revealed that the Decision Trees was the highest-performing algorithm, with 91.40% accuracy.
- [S22] used classification techniques to predict whether students were likely to drop out of their courses. The results revealed that Random Forests was the highest-performing model, with an accuracy of 94.14%.
- [S23] predicted whether a student would pass or fail using EDM models. The findings revealed that Artificial Neural Networks are the best-performing algorithm.
- [S26] used EDM techniques to predict whether students were likely to pass or fail. The findings revealed that the Random Forest method is the best-performing algorithm, with an accuracy of 99%.
- [S32] used a discrete-time algorithm to predict the timing and occurrence of student dropouts in an undergraduate engineering program. The results revealed that the impact of the type of residence differed with time.
- [S35] used students' demographic and weekly VLE data for predicting students who were likely to fail or drop out through ensemble EDM techniques. The results revealed that the voting classifier ensemble method outperformed the other individual models over forty weeks and provided an improved solution to the problem of predicting students at risk.

- [S36] used EDM models to discover factors influencing a student's success or failure. The results revealed that age, ethnicity, school province, math, and physics contributed most to a student's success or failure.

Grade prediction: EDM is commonly used to predict students' overall academic performance or CGPA ([S3], [S5], [S7]). The literature has mostly used the Classification and Regression approaches for students' academic performance prediction. It is noted that the literature predicting final grades is relatively less than the success prediction literature. The following studies focus on grade prediction, as shown in Table 5.11:

Table 5.11: Studies focusing on grade prediction

ID	Predictor used	Aim of prediction	Level of prediction	Best Performing EDM Algorithm	Effectiveness
[S3]	• (-) Demographics • (+) Previous Grades and Current Class Performance	Final Grade	Year Level	Linear and pure quadratic regression models	Accuracy: 85,89%
[S5]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (-) Socio-economic data	Final Grade	Degree Level	ID3 decision trees	Accuracy: 61%
[S7]	• (+) Demographics • (-) Previous Grades and Current Class Performance	Final Grade	Year Level	J48 Decision Trees	Accuracy: 100%
[S18]	(+) Previous Grades and Current Class Performance	Final Grade	Degree Level	Generalized Regression Neural Network	Accuracy: 95%
[S30]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Socio-economic data • (+) Instructor Attributes • (+) E-learning Activities	Final Grade	Course Level	J48 decision tree	Accuracy: 98%

- [S3] used a KNIME-based EDM model to predict students' fifth year final CGPA. The findings showed that students' GPAs for the first three years significantly influenced the class grade and final-year graduation results.
- [S5] used EDM methods to predict students' final performance. The most determining factors influencing predictions were gender, age at admission, UTME score, subject grades in core subjects, and first-level grade.
- [S7] used classification techniques to predict students' final CGPA. The findings showed a weak correlation between the best six aggregates and CGPA. The Decision Tree was the best-performing algorithm.
- [S18] developed Artificial Neural Networks to predict students' CGPA using academic results. The findings revealed that the Generalized Regression Neural Network was the highest-performing method, with an accuracy of 95%.
- [S30] applied EDM on logs extracted from an LMS to predict whether students' engagement in LMS influenced their academic performance. The results indicated a significant relationship between using LMS resources and students' academic performance at a 0.01 significance level. The J48 Decision tree correctly classified 100% of the cases and was the best-performing algorithm.

Level of prediction

Predicting students' academic performance early and accurately is the primary goal of HEIs. However, the time and availability of data influence students' academic performance prediction. Hence, EDM researchers predict students' academic performance at different levels depending on the data availability. The literature classifies the different levels of students' academic performance predictions into three areas: the degree, year, and course levels.

Degree Level: Most of the EDM literature in this analysis predicted students' academic performance at the degree level. The literature commonly focuses on binary and multi-class predictions. In literature where CGPA is the target variable, the predictions are regarded as a multi-class problem ([S2]). In contrast, literature that predicts whether students will pass or fail is regarded as a binary class problem ([S28], [S29]). An interesting finding related to literature that predicted students' academic performance using enrolment data, specifically grades from the program's first two years, yielded higher performance than those that included only pre-university or demographics data.

The interest in this finding lies in the confirmation of what might seem like a straightforward assumption. While one might expect that incorporating grades from the early years of a program would improve predictive accuracy, having empirical evidence supporting this expectation is crucial. The significance of this result is in its validation of the common-sense notion, emphasizing that, within the context of sub-Saharan STEM education, using early academic indicators is not only intuitive but also empirically proven to significantly enhance the effectiveness of EDM in predicting students' academic performance. This reinforces the idea that program-specific factors, like early academic achievements, can play a more pivotal role in predictive modelling than what might be initially assumed or expected. It highlights the importance of data-driven insights in shaping our understanding of the factors influencing academic outcomes. The following literature predicted students' academic performance at the degree level, as shown in Table 5.12:

Table 5.12: Studies focusing on degree-level prediction

ID	Predictor used	Aim of prediction	Best Performing EDM Algorithm	Effectiveness
[S2]	• (+) Previous Grades and Current Class Performance • (+) Demographics	Final Grade	J48 decision tree	Accuracy: 50%
[S4]	• (-) Demographics • (+) Previous Grades and Current Class Performance	Final Grade	Decision Trees	Accuracy: 73,6%
[S11]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Socio-economic data	Final Grade	Multilayer perception network model	Accuracy: 98%
[S19]	(+) Previous Grades and Current Class Performance	Final Grade	Matrix Factorization	Root Mean Square Error: 9,7
[S24]	• (+) Demographics • (+) Previous Grades and Current Class Performance	Pass-Fail	Support Vector Machine	Accuracy: 87%
[S28]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Socio-economic data	At-Risk (Pass-Fail-dropout)	Random Forest	Accuracy: 95%
[S29]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Socio-economic data	Pass-Fail	Random Forests	Accuracy: 95,45%
[S41]	• (+) Demographics	Final Grade	Logistic Regression	Accuracy: 84,7%

	• (+) Previous Grades and Current Class Performance			

- [S2] analysed students' first two years' performances to predict their final performance. The results showed that a high GPA in discrete mathematics, programming, network, and computer architecture courses significantly predicts students' final grades.
- [S4] examined students' ethnic influence on the EDM models' predictive accuracy. The maximum classification accuracy observed from multiple regression analysis was 53.2%. The findings revealed that students' ethnicity was not significant in predicting their performance. Hence, using ethnicity alone to predict students' academic performance is unsuitable.
- [S11] used EDM models to assist with classifying students' academic performance. The results showed the multilayer perceptron was the highest-performing algorithm.
- [S19] applied Matrix Factorization using the student and course average marks for the student's academic performance prediction. The findings indicated that the Matrix Factorization performed better in the predictions.
- [S24] used the Support Vector Machines, Naïve Bayes, and Decision Trees models for predicting first-year undergraduate science students' academic performance. Support Vector Machines was the most high-performing algorithm, with an accuracy of 87%.
- [S28] used a student attrition model to predict student Earth Science program performance accurately. The results revealed that the critical attributes of a student used in the study are background, individual and pre-college information. Additionally, the Random Forests attained the maximum accuracy of 95%.
- [S29] used six classification algorithms with students' first-year grades to predict their final grades. The results showed that Random Forests attained the maximum accuracy of 95.45%.
- [S41] used EDM models to predict students' academic performance using CGPA, gender, age, region, and high school exam scores to enable efficient and timely interventions. The results showed that students with high secondary school exam grades were more likely to succeed academically.

Year Level: In this analysis, few literature articles focused on predicting students' academic performance at the year level. Identical to the previous sub-section, the literature that used only pre-enrolment and socio-economic data gave worse accuracy ([S31]). In contrast, literature containing enrolment data, such as internal assessments, had improved model performances ([S25], [S37]). It is noted that the studies that predict students' academic performance at the year level perform less than both the degree-level and course-level studies. Hence, this research gap can be explored in future studies. Researchers focus on identifying the main features that cause these models' low accuracies. Additionally, future researchers can analyse the best features and models to improve students' academic performance prediction at the year level. The following literature focused on student academic performance prediction at the year level, as presented in Table 5.13:

Table 5.13: Studies focusing on year-level prediction

ID	Predictor used	Aim of prediction	Best Performing EDM Algorithm	Effectiveness
[S15]	• (+) Previous Grades and Current Class Performance • (+) Instructor attributes	Pass-Fail	Support Vector Machine	Accuracy: 72,87%
[S25]	• (+) Demographics • (-) Previous Grades and Current Class Performance • (+) Socio-economic data	At-Risk (Pass-Fail-dropout)	Random Forest	Accuracy: 85%
[S31]	• (+) Previous Grades and Current Class Performance • (+) Prior knowledge	Pass-Fail	Decision Trees	Accuracy: 73,44%
[S34]	• (+) Previous Grades and Current Class Performance • (+) Demographics • (+) Socio-economic data • (+) E-learning Activities	At-Risk (Pass-Fail-dropout)	Decision Trees	Accuracy: 90%
[S37]	• (+) Previous Grades and Current Class Performance • (+) Demographics	Final Grade	C4.5 Decision tree	Accuracy: 81,4%

- [S15] predicted students' academic performance using Bayesian Networks, Decision Trees, and Support Vector Machines models. The findings revealed that Support Vector Machines were the best-performing algorithm for students' academic performance prediction before the year-end examinations, with an accuracy of 72.87%.

- [S25] used a learner attrition algorithm for predicting at-risk students' academic performance. The findings indicated that Random Forests was the best-performing method, with an accuracy of 85%.
- [S31] evaluated students' computer skills to predict students' failure or success in a first-year computer science program. Based on the accuracy of three classification models, the study determined that students' prior computer knowledge was the best determinant of students' academic performance, with an accuracy of more than 60% in all the models.
- [S34] used personality and e-learning behaviour features to predict students' likelihood of failing or dropping out. The results revealed that students' e-learning behaviours and personalities were more significant in predicting students' academic performance than their background.
- [S37] developed an EDM model to predict first-year students' academic performance. The findings showed that the C4.5 Decision Tree algorithm successfully predicted students' academic performance by achieving 81.4% accuracy.

Course Level: Predicting students' academic performance at the course level is this analysis's second most cited theme. The effectiveness of predicting whether students might pass or fail a course is helpful because it may assist HEIs in identifying key factors influencing students' academic performance. Hence, students' academic performance prediction at the course level is crucial as it can assist HEIs in making early, efficient, and timely decisions and improve the academic performance of students ([S14]). The following literature has predicted students' academic performance at the course level, as shown in Table 5.14:

Table 5.14: Studies focusing on course-level prediction

ID	Predictor used	Aim of prediction	Best Performing EDM Algorithm	Effectiveness
[S6]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Instructor Attributes	Pass-Fail	J48 Decision Tree	Accuracy: 71,57%
[S13]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Prior knowledge • (+) Socio-economic data	Final Grade	Regression	Accuracy: 58,5%
[S14]	• (+) Demographics • (+) E-Learning Activities • (+) Previous Grades and Current Class Performance	Pass-Fail	Decision Trees	Accuracy: 86%
[S27]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Socio-economic data	Final Grade	Artificial Neural Networks	Accuracy: 83%
[S38]	• (+) Demographics • (+) Previous Grades and Current Class Performance	At-Risk (Pass-Fail-dropout)	Multiple Regression	Accuracy: 79%
[S39]	• (+) Demographics • (+) Previous Grades and Current Class Performance	Final Grade	J48 Decision Tree	Accuracy: 95,24%
[S40]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Socio-economic data	Pass-Fail	J48 Decision Tree	Accuracy: 84%

- [S6] built EDM algorithms for predicting students' competence in computer programming. The findings indicated that the factors significantly influencing students' academic performance in programming were self-efficacy, previous knowledge of programming, lecturer's performance, motivation, and interest level.
- [S13] explored factors influencing undergraduate programming students' academic performance by developing prediction algorithms for student academic performance. The findings indicated that the determining factors of programming students' performance were fearful perceptions of programming courses, the attitude of students and lecturers, student health, student attendance, university facilities, and erratic power supply.

- [S14] developed a concept for a standalone prototype dashboard using predictive modelling for an introductory statistics course. The results showed that engagement variables and course assessments at specific intervals through a semester provide helpful insights for decision-making at those particular intervals.
- [S27] conducted predictive modelling for student academic performance prediction in computer courses. The results indicated that factors of secondary school grades and prior computer skills were not significant in predicting students' academic performance.
- [S38] built a predictive model using Business Statistics students' grades to identify students likely to fail or drop out. The Logistic Regression had the best performance, with an accuracy of 80%.
- [S39] developed a Decision Tree classification model to forecast students' academic performance in the Biological Sciences program. The results revealed that the Decision Trees method was the best-performing method, with an accuracy of 95.24%.
- [S40] used performance-weighted ensemble classifiers to mine student data enrolled in STEM courses. The results revealed that Mathematics core subject scores, high school final exam scores, belief in the ability to succeed, teacher inspiration, and expected career flexibility significantly affected the choice of students to enrol in STEM programs.

5.3 Challenges in Academic Performance Analysis and Prediction

EDM researchers have experienced various challenges in the sub-Saharan African region when predicting students' academic performance. The main challenges highlighted in the literature include imbalanced datasets, ethics, and generalizability of results.

5.3.1 Issue of Imbalanced Datasets

When the number of records from one class is considerably less than from others, the data is said to be imbalanced ([S16]). The lack of balance in datasets has negative implications for the performance of EDM models ([S25]). Data imbalances are a common issue experienced by EDM researchers. Most EDM researchers have opted for re-sampling methods, such as under-sampling or over-sampling, to address this prevalent issue ([S9], [S14], [S35]). Under-sampling involves randomly eliminating variables from significant classes or using ensemble methods such as bagging to balance the dataset.

In contrast, over-sampling involves duplicating instances randomly or synthetically generating some samples to increase the number of minor classes. In this analysis, the literature mostly used the over-sampling method, namely the Synthetic Minority Oversampling Technique (SMOTE). Even though these sampling methods try to address this issue, there are still some shortcomings. Over-sampling methods may lead to overfitting risks with increasing noise, and valuable information will be lost in the case of under-sampling methods ([S28]). Since the re-sampling methods are implemented at the algorithm level, using actual data or ensemble methods rather than synthetic data can lead to higher model performances and the discovery of new knowledge.

5.3.2 Ethical issues

A significant challenge in the analysis was that some researchers could not use particular student data for academic performance prediction due to privacy and ethical concerns, which limited the scope of some EDM research ([S9]). In this case, some studies used data from online repositories ([S35]). However, it is noted that the data from online databases are less rich than in HEIs databases. Moreover, the scarcity of robust online and free educational datasets is due to privacy and ethical issues. Hence, a call is made to HEIs to start distributing open datasets to encourage future research studies in the EDM field.

5.3.3 Generalizability of data

EDM researchers use demographic and socio-economic factors to predict students' academic performance, but with unclear success. It is clear that in this analysis, demographic and socio-economic factors are influenced significantly by the culture of the country where the EDM research is conducted. When the investigation is conducted in a collectivistic country such as South Africa, Nigeria, Tanzania, Zambia, and other sub-Saharan African countries, the factors that mainly influence predictions are family-related, which include parents' qualifications, family size, family income, and family support ([S10], [S11], [S26]). This may not be prevalent in literature conducted in individualistic countries such as the United States and Europe.

The distinction between collectivistic and individualistic cultures is a critical aspect of understanding cultural dimensions, often associated with cultural framework proposed by Hofstede (2011). In collectivistic countries the focus is on family and community goals over individual requirements, whereas in individualistic countries the focus is on personal achievements (Khan & Ghosh, 2021). This dichotomy in cultural orientations has significant implications, especially in the context of academic performance.

For instance, the observation that students from individualistic countries may exhibit a more competitive attitude compared to students from collectivistic countries aligns with expectations derived from cultural dimensions. In individualistic societies, where personal success is highly valued, students may approach academic tasks more competitively. However, acknowledging that this finding requires further investigation highlights the complexity inherent in the relationship between culture and academic performance. Exploring the differences within each cultural dimension becomes crucial to delve deeper into this phenomenon, considering factors such as educational systems, societal expectations, and the influence of cultural values on academic motivation. Additionally, Hofstede's cultural dimensions provide a valuable framework for this exploration, offering insights into power distance, uncertainty avoidance, individualism, and collectivism (Hofstede, 2011).

Moreover, it is noteworthy that the majority of studies within the selected pool in this analysis featured datasets with a sample size below 5000. While the observed findings regarding cultural influences on academic competitiveness offer valuable insights, the restricted sample sizes may limit the generalizability of these results. To bolster the robustness and applicability of future studies, it becomes imperative to advocate for the expansion of datasets. Increasing the size of datasets allows for a more diverse representation of cultural contexts, enhancing the external validity of the research findings. Therefore, future work in this domain should prioritize larger datasets to ensure that conclusions drawn from the studies are more broadly applicable and reflective of the rich diversity within collectivistic and individualistic societies. This consideration is pivotal for advancing our understanding of the intricate interplay between culture and academic performance on a global scale.

5.4 Alignment of the Thematic Analysis and HEDM Framework

The linkage of the HEDM framework with thematic analysis, as shown in Figure 5.5, can help researchers leverage the strengths of both methodologies, combining quantitative rigour with qualitative depth. This integrated approach allows a more comprehensive exploration of EDM's effectiveness in students' academic performance prediction in sub-Saharan African HEIs. The main phases of this integrated approach involve data collection, data preprocessing, quantitative analysis, identification of themes, integration of findings, and interpretation and theory building.

Data Collection: In the initial stages of HEDM, data or inputs are collected from various educational sources, including LMS, SIS, and surveys. Thematic analysis can be used to

supplement this quantitative data by collecting qualitative data through focus groups, open-ended survey questions, or interviews. Thematic analysis helps capture learners' and educators' rich, subjective experiences and perspectives.

Data Preprocessing: In HEDM, the collected data is pre-processed to remove noise, handle missing values, and transform the data into a suitable format for analysis. Similarly, thematic analysis involves transcribing and organizing qualitative data, such as open-ended survey responses or interview transcripts, into meaningful units of analysis.

Quantitative Analysis: Once the data is prepared, the HEDM framework is subjected to various EDM methods. These methods aim to uncover patterns, correlations, or predictive models that explain student performance, learning behaviours, or instructional effectiveness.

Identification of Themes: In parallel, the thematic analysis is conducted to identify recurring patterns, themes, or codes within the qualitative data. Researchers analyse the data to identify key ideas, concepts, or categories that emerge from the data. These themes provide insights into individuals' subjective experiences, beliefs, and attitudes within the educational context.

Integration of Findings: Once the thematic and quantitative analyses are completed, the findings from both approaches can be integrated. The thematic analysis provides a qualitative understanding of methods, tools, techniques, and data used to attain the predictive power of EDM. At the same time, the HEDM framework offers quantitative insights from the methods, tools, techniques, and data. By combining these findings, researchers can obtain a more comprehensive understanding of EDM's effectiveness in students' academic performance prediction in sub-Saharan African HEIs.

Interpretation and Theory Building: Integrating findings from the HEDM framework and thematic analysis allows researchers to interpret the results holistically and develop theoretical frameworks. Thematic analysis helps contextualize the quantitative findings and explains the patterns observed in the data. This iterative process of analysis and interpretation facilitates theory-building in educational research.

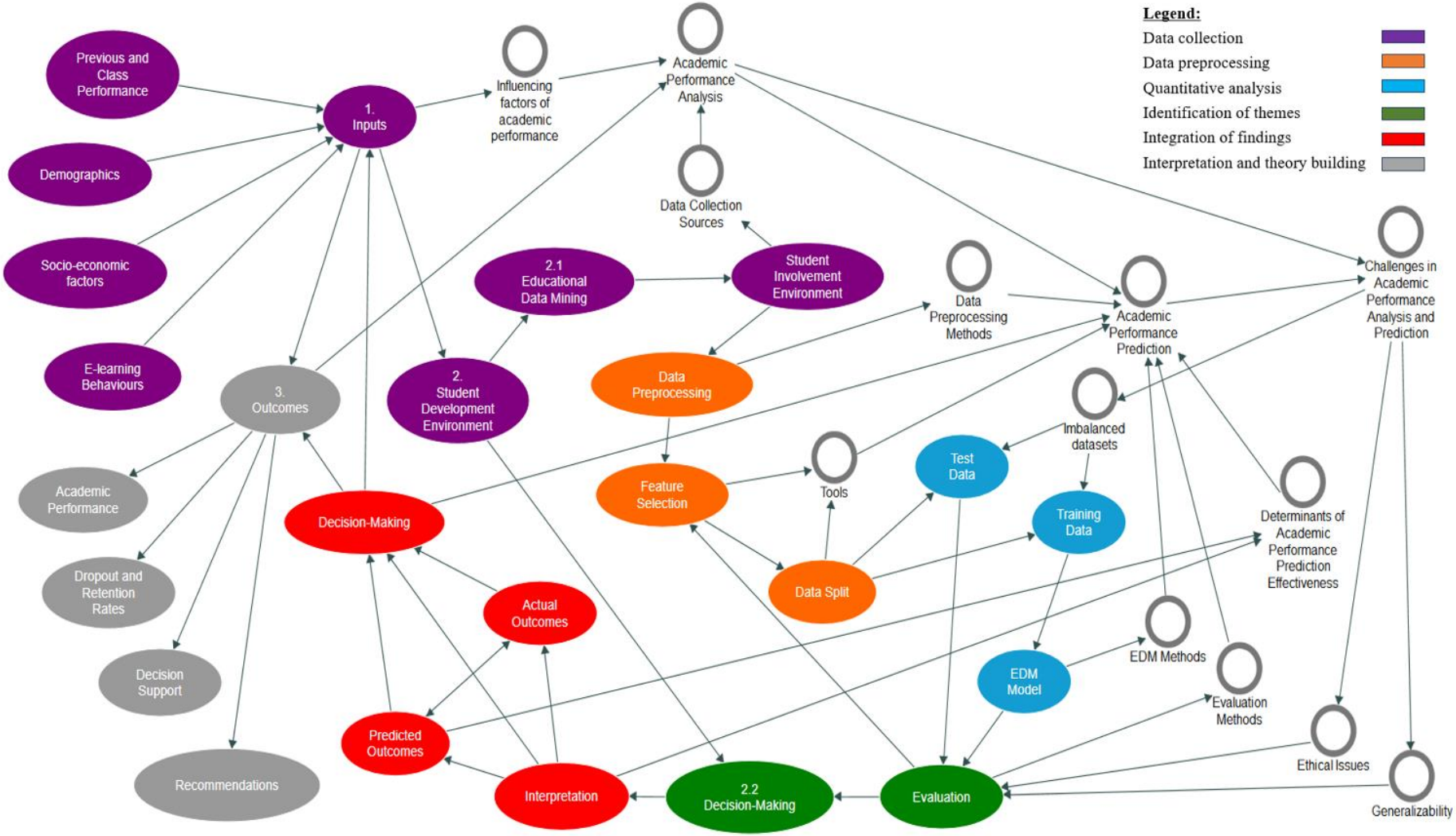

Figure 5.5: Thematic Analysis and HEDM framework alignment map

5.5 Summary

This chapter presented the second part of the thematic analysis obtained from the systematic search for literature via a research tool, NVivo version 1.7.1. This study identified the frequently used EDM methods and tools for predicting students' academic performance. The study explored the challenges of academic performance analysis and prediction to consolidate the findings from the two parts of the analysis. Finally, the study explored the alignment of the HEDM framework with the thematic analysis towards theory-building and knowledge discovery.

Chapter 6: Discussion

6.1 Introduction

This chapter synthesises the results mentioned in the previous chapters four and five and answers the research questions mentioned in section 1.3 by providing a comprehensive analysis of the potential implications of this study. This SLR focused on EDM's effectiveness in predicting students' academic performance in sub-Saharan African HEIs. The aim was to determine the effectiveness of EDM in predicting sub-Saharan African undergraduate STEM students' academic performance. This study analysed 41 selected studies that focused on predicting students' academic performance in sub-Saharan African HEIs and provided an interpretation of the results through various dimensions.

Embarking on the discussion section, the analysis will commence by dissecting the sub-research questions. Each of these smaller inquiries acts as a stepping stone, providing a more detailed view of the overall landscape. This segment aims to uncover the patterns, trends, or knowledge encapsulated within each sub-question, gradually building a clear and comprehensive picture of the current investigation. The journey involves exploring these intricacies before seamlessly connecting the dots, ultimately converging to address the main research question. It is similar to following a trail of breadcrumbs to uncover insights and revelations that will be synthesized to provide a holistic answer to the overarching inquiry.

6.2 Sub-RQ1: What key input factors are used in EDM to predict sub-Saharan African undergraduate STEM students' academic performance?

Table A.4 in Appendix B shows the most influential factors in the EDM prediction of students' academic performance. The SLR findings showed that one of the most crucial factors influencing students' academic performance was previous and current class performance. The previous and current class performance factors, such as external assessment, internal assessment, and pre-enrolment grades, were used in all the selected literature that predicted students' academic performance. External assessments refer to standardized tests, examinations, or evaluations conducted externally in the regular classroom setting. These assessments often aim to measure a student's knowledge, skills, and abilities in a standardized and impartial manner. Internal assessments encompass evaluations conducted within the classroom by instructors. These may include assignments, quizzes, projects, or any form of

assessment designed by the course instructor. Pre-enrolment grades refer to the academic performance records of students before entering the specific STEM program or course under investigation.

These factors collectively form a robust set of indicators, combining standardized measures, ongoing classroom evaluations, and historical academic records to provide comprehensive insights into students' potential academic outcomes. This finding aligns with an investigation by Hussain *et al.* (2018), who found that internal assessments and previous grades are the most frequently used factors in predicting students' academic performance. This is further supported by Saa *et al.* (2019), where most of their analysed articles considered CGPA an essential factor in students' academic performance prediction. This CGPA is the most frequently used factor because it is measurable since it is a cumulation of all the past and present performances in a degree or program.

Another most used factor in predicting students' academic performance is demographics. Students' demographic factors include background and academic factors such as age, ethnicity, year of study, school quintile, and gender. These factors contribute to a holistic understanding of the student population, considering the diverse backgrounds and contexts that may influence academic performance. This finding aligns with Alyahyan and Düştegör (2020) and Saa *et al.* (2019). Saa *et al.* (2019) found that demographic factors are frequently used for students' academic performance prediction. Furthermore, Alyahyan and Düştegör (2020) found a positive correlation between students' demographics and academic performance.

The socio-economic factor is another factor in this study that affects students' academic performance prediction. Socio-economic factors consist of elements related to students' economic and social backgrounds. The socio-economic factor includes residence, family history, and family income level. In a previous study by bin Roslan and Chen (2022), it was found that family attributes are an essential determinant of students' academic performance. Hence, they should be included as predictors of students' academic performance prediction. Some studies also used students' e-learning behavioural data obtained from LMS. E-learning behavioural data involves tracking students' interactions and engagement within online learning environments. This student behavioural data includes student engagement data, including the frequency of uploading assignments, accessing online material, and solving online quizzes. These factors were crucial determinants of students' academic performance, this also aligns with Hussain *et al.* (2018) and Chweya *et al.* (2020) findings.

Other aspects that appeared in fewer than six studies were instructor attributes and students' prior knowledge. Instructor attributes involve characteristics related to instructors, such as teaching styles or the level of experience (Khan & Ghosh, 2021). In this study, five studies examined instructor attributes that influenced students' academic performance prediction. Most literature analyses the influence of instructor attributes with indirect factors such as students' satisfaction, motivation, interest level, self-efficacy, and attendance. Only two pieces of literature explore the direct influence of instructor attributes on students' academic performance. These studies observed that effective teaching encourages students to improve their performance, whereas poor teaching degrades performance.

Prior knowledge was the least cited theme, with only four studies. Prior knowledge involves recalling or recognizing processes, ideas, theories, and terms (Khan & Ghosh, 2021). In EDM, prior knowledge is the awareness of a topic based on previous learning. EDM researchers have used prior knowledge to analyse different areas, including students' competence, enrolment choices, and performance prediction. The principal metric of prior knowledge relates to previous experience or specialization and is acquired through surveys. According to Saa *et al.* (2019), factors reported in only a few studies are likely to have little or no influence on predictions of students' academic performance. Therefore, EDM researchers may have to re-evaluate the inclusion of these aspects on such a prediction or at least re-evaluate the predictors that they are included with so that they are matched with the best predictors to increase the accuracy in the algorithms of students' academic performance prediction.

In summary, the key input factors used in EDM for predicting sub-Saharan African undergraduate STEM students' academic performance are the previous and current class performance factors consisting of external assessments, internal assessments, and pre-enrolment grades. Demographic considerations, socio-economic factors, and e-learning behavioural data complement these academic indicators, collectively providing a comprehensive understanding of students' diverse backgrounds and contextual influences. While instructor attributes and students' prior knowledge offer supplementary insights, their sporadic presence suggests a need for careful consideration. This multifaceted approach, spanning academic, contextual, and behavioural dimensions, contributes to a more holistic and accurate prediction of academic outcomes in the context of sub-Saharan African undergraduate STEM education.

6.3 Sub-RQ2: What are the key student involvement factors used in EDM to predict sub-Saharan African undergraduate STEM students' academic performance?

The typical input variables included previous educational records, demographics, e-learning behaviours, and socio-economic factors, as shown in Table A.4 in Appendix B. These input variables are also evident in previous EDM work (Hussain *et al.*, 2018). In recent years, EDM researchers have started collecting multimodal data from learning management systems, such as the frequency of uploading assignments, solving online quizzes, and accessing online material for modelling (Du *et al.*, 2020). Generally, students' demographics, previous educational records, and socio-economic factors are categorised as static data because these features are constant and do not change. In contrast, dynamic data consists of learning behaviours, student engagement, and other multimodal data (bin Roslan & Chen, 2022).

Among the 41 selected academic performance prediction articles, only six studies used e-learning behaviours with previous educational records, students' demographics, and socio-economic factors. Most of the studies used static data to predict students' academic performance. However, using static data for students' academic performance prediction has some shortcomings. One significant limitation is the inability of static data to capture the dynamic and evolving nature of students' efforts in the learning process. Static variables, including demographics, previous academic records, and socio-economic factors, offer a fixed snapshot that fails to adapt to changes in student academic performance over time. This raises concerns about the predictive model's capacity to account for improvements or declines in performance that may occur during a program.

Furthermore, using static data may oversimplify the complex interplay of factors influencing academic outcomes. Predictive models may overlook subtle shifts in student engagement, learning habits, and response to interventions by focusing on unchanging variables. This oversimplification could compromise the model's ability to provide accurate predictions. While most studies relied on static data, including e-learning behaviours introduces a dynamic aspect to the predictive modelling process. Dynamic data, comprising learning behaviours, student engagement, and multimodal data, provides a more real-time perspective on students' academic journeys. This shift from static to dynamic data acknowledges the evolving nature of students' efforts and engagement throughout their educational experiences. The selected studies in this analysis with learning behaviours as input factors found that student engagement

levels or participation frequencies positively influenced students' academic performance. This also aligns with the findings of Hussain *et al.* (2018) and Chweya *et al.* (2020).

These findings, however, are challenging to generalize for several reasons. Firstly, the researchers used absolute frequencies as input factors. Since each course or program has its set of unique demands, it is challenging to find generalizable thresholds of each learning behaviour in various programs. The unique nature of academic requirements in diverse educational settings makes pinpointing standardized benchmarks that hold true across the board challenging. Furthermore, the temporal aspect of data collection emerges as a crucial factor influencing the applicability of learning behaviours in predicting academic performance. The effectiveness of predicting outcomes appears to be dependent on the duration and timing of data collection. Notably, except for the two studies that collected learning behaviours over several weeks to identify low-performing students early, most studies that used learning behaviours for students' academic performance prediction at the end of a course were more likely to identify significant predictors than to predict anything.

The intricacies of research methods underline the challenges of integrating learning behaviours into predictive models. While some methods demonstrate effectiveness in particular situations, the difficulty lies in developing strategies that reconcile the unique characteristics of individual courses with the broader aim of creating widely applicable predictive models. Tackling these challenges is crucial for enhancing EDM's practicality and widespread applicability in predicting academic performance among sub-Saharan African STEM students.

In summary, the exploration of EDM for predicting academic performance among sub-Saharan African undergraduate STEM students reveals a dynamic interplay of key student involvement factors. These factors are a mix of traditional and evolving variables, with the conventional inputs consisting of previous educational records, demographics, and socio-economic factors. Recent advances in data collection witness the inclusion of e-learning behaviours, such as assignment uploads, online quiz participation, and course material access. Integrating dynamic data, especially e-learning behaviours, provides a comprehensive view of student engagement. However, challenges in generalizability arise due to varying behaviours across courses, and the absence of standardized thresholds complicates universal predictive measures. Temporal considerations highlight the need for a balanced approach between early identification and end-of-course predictions. These findings emphasize the necessity for innovation and refinement as EDM evolves in sub-Saharan African STEM education. Traditional factors offer a

foundational understanding, while the inclusion of dynamic data signals a progressive shift. The challenges underline the roadmap for future research, urging ongoing exploration and enhancement of these factors to enhance the effectiveness of EDM in predicting academic performance.

6.4 Sub-RQ3: What are EDM's most frequently used tools and algorithms to predict sub-Saharan African undergraduate STEM students' academic performance?

This study identified several tools for predicting students' academic performance, as presented in Figure 5.1 in section 5.2.2. The SLR findings revealed that the most frequently used EDM tool is WEKA. This is due to its flexibility and the availability of various EDM models for predictive modelling. In this study, the most applied EDM approach is Classification, which appeared in 38 (93%) of the reviewed studies, followed by Regression, which occurred in 22 (54%) reviewed studies, while only three (7%) reviewed studies employed clustering approaches, as shown in Figure 5.2 in section 5.2.3. This finding aligns with bin Roslan and Chen (2022), who suggested that the classification approach is more frequently used than other approaches in EDM research. The prominence of Classification and Regression methods highlights their effectiveness in predicting students' academic performance. However, the limited use of clustering methods signals an avenue for further exploration.

The Decision Trees model is the selected studies' most frequently used EDM classification method. The ID3, J48, and C4.5 are examples of the Decision Trees model with a structure resembling a tree with the leaf nodes at the bottom and the root node at the top bin Roslan and Chen (2022). Decision Trees are commonly used because they are easily understood and efficiently predict students' academic performance (Saa *et al.*, 2019).

Artificial Neural Networks and Bayesian Networks are the second most commonly used methods in student academic performance prediction. Artificial Neural Networks are a set of interconnected neurons or nodes that model the human brain's neurons arranged in layers (Khan & Ghosh, 2021). The most used Artificial Neural Network algorithm is the multilayer perceptron, which feeds information comprising the input, hidden, and output layers. Bayesian Networks are probabilistic graphical algorithms with directed edges and nodes (Singh & Pal, 2020). The Naïve Bayes is a simple form of Bayesian networks with all the variables being conditionally class independent and is frequently used in EDM studies (Chaka, 2021). The Naïve Bayes method consists of algorithms such as Bernoulli, Gaussian, and Multinomial. The

main advantage of using Naïve Bayes is that the predictions' outcomes can be easily translated into human language (Saa *et al.*, 2019). Although the Naïve Bayes was frequently used, it was outperformed by the other algorithms it was used with in most studies. However, it still managed to achieve an accuracy above 70% in the studies it appears in. Hence, future researchers must tune the predictor set to suit this model and achieve higher predictive accuracy.

The Random Forest algorithm is another technique used by EDM researchers. Random Forest is an improvement on Decision Trees. Random Forest produced higher accuracy in most of the analysed literature than Decision Trees and Naïve Bayes models (Hasan *et al.*, 2020). Next is the Support Vector Machine, a supervised algorithm that finds a hyperplane that divides two classes by the most significant margin between them. Support Vector Machine classifiers are not as regularly used as the previous algorithms but have their advantages. Firstly, Support Vector Machines are generally faster than other methods with a high degree of generalization (bin Roslan & Chen, 2022). Support Vector Machines are also well-suited for small datasets (Zohair & Mahmoud, 2019).

The least used EDM method by the selected articles was the k-Nearest Neighbours method. This method categorises new cases through a similarity metric and maintains all available cases. The advantage of K-Nearest Neighbours is its robustness regarding the search space, as classes do not have to be separated linearly, including its usability in both classification and regression prediction problems (Chweya *et al.*, 2020). Furthermore, a few studies used custom or ensemble methods such as bagging or gradient boosting to achieve high prediction effectiveness when the other traditional classification algorithms were not performing well.

In exploring the array of EDM tools, one notes that, despite the prevalence of popular choices, certain algorithms, notably the k-Nearest Neighbours method, see limited use, prompting reflection on their relevance in predicting academic performance. Navigating the EDM toolkit in sub-Saharan African STEM education demands a comprehensive understanding of each tool's contextual fit within the region's unique dynamics. Envisioning effective EDM in this context involves implications beyond tool selection. Researchers are prompted to delve into uncharted territories, considering the latent potential of clustering methods and critically evaluating the appropriateness of algorithmic choices. The quest for improved predictions in sub-Saharan African STEM education ultimately depends on a balanced integration of popular

tools, guided by a sharp awareness of their alignment with the intricate landscape of academic performance.

In summary, WEKA is the most favoured tool in EDM for sub-Saharan African STEM contexts, known for its flexibility. The prevalent use of Classification and Regression methods highlights their effectiveness, though the limited adoption of clustering methods suggests an area for potential exploration for future research. The Decision Trees model is the most used EDM algorithm, emphasizing its role in unravelling academic intricacies. However, the less frequent use of the k-Nearest Neighbours method raises questions about its contextual suitability. Navigating the EDM landscape in sub-Saharan African STEM education necessitates a thoughtful blend of popular tools and a keen understanding of their relevance within the unique academic dynamics.

6.5 Sub-RQ4: To what extent is EDM effective in predicting sub-Saharan African undergraduate STEM students' academic performance?

Evaluating the effectiveness of EDM in predicting the academic performance of sub-Saharan African undergraduate STEM students involves employing various key evaluation methods or metrics such as confusion matrix, accuracy, precision, recall, specificity, Area Under the Curve and Receiver Operating Characteristic (AUC-ROC) curve, F1 measure, and kappa statistics, as shown in section 5.2.4. The confusion matrix and accuracy are the most pivotal methods in this analysis. This finding aligns with the work of Khan and Ghosh (2021). The confusion matrix is a fundamental tool in EDM evaluation, providing a detailed breakdown of the model's predictions compared to the actual outcomes. It comprises four essential components: true positives (number of successful students correctly predicted as successful), true negatives (number of unsuccessful students correctly predicted as unsuccessful), false positives (number of successful students predicted incorrectly as unsuccessful), and false negatives (number of unsuccessful students incorrectly predicted as successful). This matrix enables a granular analysis of the model's performance, especially in the context of predicting academic success or failure. Accuracy is a widely used metric that gauges the overall correctness of the model's predictions. It is calculated by dividing the sum of true positives and true negatives by the total number of instances. While accuracy offers a comprehensive view of the model's overall performance, it is crucial to interpret it alongside other metrics, such as precision, recall, and F1-score, given its sensitivity to class imbalances (bin Roslan & Chen, 2022).

The evaluation of EDM models in the context of sub-Saharan African undergraduate STEM students involves critically examining the effectiveness of predictive algorithms, considering factors such as model accuracy, precision, recall, and overall performance metrics. One must also delve into the appropriateness of evaluation methods for sub-Saharan African STEM education's unique challenges and characteristics. Challenges may arise due to variations in educational systems, cultural differences, and disparities in access to resources, influencing the relevance and reliability of evaluative metrics. Moreover, the need to balance between success prediction and final grade prediction adds layers of complexity to the evaluation process. Understanding the intricacies of evaluating EDM models in sub-Saharan Africa requires a comprehensive exploration of the strengths and limitations of commonly used evaluation methods or metrics. It calls for a comprehensive approach that not only assesses the predictive power of models but also considers their practical applicability and impact on enhancing academic outcomes for STEM students in the region.

These evaluation methods go beyond mere assessment, playing a crucial role in informed decision-making. Educators and administrators can leverage insights from the confusion matrix to tailor interventions and refine the model's predictive capabilities. The continuous evaluation using these metrics forms a feedback loop, fostering iterative improvements. Importantly, in the context of sub-Saharan African STEM education, this approach acknowledges the unique challenges and intricacies of the academic environment. It contributes to the development of an EDM model that is not only accurate but also contextually sensitive, aligning with the region's specific educational landscape and paving the way for more effective student support and success.

In summary, the evaluation of EDM effectiveness in predicting sub-Saharan African STEM students' academic performance involves a comprehensive analysis, primarily through the use of the confusion matrix and accuracy metrics. These techniques provide valuable insights into the model's predictive capabilities, enabling researchers and educators to make informed decisions and enhancements to support student success in the context of sub-Saharan Africa STEM education.

6.6 Main-RQ: How effective is EDM in predicting sub-Saharan African undergraduate STEM students' academic performance?

Examining the efficacy of EDM in predicting sub-Saharan African undergraduate STEM students' academic performance requires a comprehensive examination of the predictive goals, specifically success and final grade prediction. These prediction goals are explored across different levels, encompassing degree, year, and course. Table A.7 in Appendix B lists the selected studies' details on predictors and methods used, aims, and the effectiveness observed. In this analysis, the primary emphasis lies in success prediction, particularly at the degree level, as shown in section 5.2.5. Success prediction involves anticipating positive outcomes across students' academic journeys, stressing a holistic comprehension of academic success within the selected STEM program. Significant predictors are needed to achieve higher effectiveness in prediction models. Fortunately, these predictors are readily available at the degree level, consisting of all the student data throughout their enrolled program. The focus on success prediction at the degree level corresponds with an in-depth exploration of key input factors within EDM for academic performance prediction. These predictors include overall student academic performance, previous educational records, demographics, socio-economic factors, and e-learning behaviours (Hussain *et al.*, 2018).

This study found that most studies frequently used external assessments, such as CGPA, and internal assessment grades, to achieve higher prediction accuracy. In comparison, a few have used internal assessments only to predict students' academic performance. Other studies used student demographics, socio-economic factors, and internal assessments to predict students' academic performance. Furthermore, the examination extends to student involvement factors, where a few studies used student behavioural data obtained from LMS. This student behavioural data included student engagement data, including the frequency of uploading assignments, accessing online material, and solving online quizzes. These predictors mentioned above have been found to be suitable predictors contributing to EDM's effectiveness. The integration of these diverse elements aims to capture the intricate dynamics shaping academic success, offering valuable insights into the predictive modelling landscape in sub-Saharan African STEM education.

The effectiveness of EDM in this context extends to the tools and methods employed. The most frequently used tool is WEKA. WEKA is widely used in EDM due to its versatility and user-friendly interface. EDM researchers leverage its diverse machine learning algorithms and data

preprocessing capabilities, making it a go-to tool for predictive modelling in the context of sub-Saharan African undergraduate STEM students' academic performance prediction. Furthermore, EDM researchers have also used different methods to predict students' academic performance. Most studies in this analysis found the Decision Trees to be the best-performing method for students' academic performance prediction. This is also evident in a previous review by Saa *et al.* (2019). The superiority of Artificial Neural Networks, Random Forests, Logistic Regression, k-Nearest Neighbour, Support Vector Machines, and Naïve Bayes models is also observed in the selected studies. The choice of these tools reflects the adaptability and availability of diverse EDM models, allowing for a comprehensive exploration of predictive patterns.

Moreover, many of the studies attained significant effectiveness in these algorithms. Some even reached an accuracy or effectiveness greater than 90% in classification methods, as shown in Table 5.7 in section 5.2.3. The Random Forest algorithm had the highest accuracy of 93%, signifying a robust predictive capacity within EDM. This high performance highlights the algorithm's efficacy in handling complex datasets and extracting meaningful patterns that contribute to accurate predictions. However, the high accuracy of these models was probably due to the use of imbalanced datasets. It is also evident that most studies in the literature ignore the data imbalance issues. Hence, this is a possible gap that may need to be addressed in future EDM work. Moreover, researchers can explore the best strategies that can be used to attain a balanced dataset to increase the accuracy and replicability of an EDM study.

The evaluation methods employed, such as the use of confusion matrices and accuracy measures, serve as vital components in assessing the effectiveness of the EDM models. These methods provide a quantitative lens to measure the accuracy and reliability of predictions. The linkage between success prediction, input factors, and student involvement, along with the tools and algorithms used, becomes apparent through the lens of evaluation methods. This evaluative lens allows for a detailed examination of the methodologies employed in predictive models, shedding light on the intricate relationships among these elements. By examining methodological intricacies, including specific approaches in success prediction models, researchers gain insights into how input factors, student involvement, and the selected tools and algorithms collectively contribute to the effectiveness of EDM in predicting academic success among sub-Saharan African undergraduate STEM students.

In summary, the exploration of EDM's effectiveness in predicting academic performance among sub-Saharan African undergraduate STEM students involves a multifaceted analysis. Success prediction at the degree level serves as the central theme, intertwined with diverse input factors such as previous and current class performance, demographic factors, socio-economic factors, and e-learning behavioural data. Student involvement aspects, including previous educational records and engagement levels, play a pivotal role in this predictive model. The use of EDM tools, prominently featuring the WEKA tool and algorithms like Decision Trees, Artificial Neural Networks, Random Forests, Logistic Regression, k-Nearest Neighbour, Support Vector Machines, and Naïve Bayes contribute to predictive modelling. Evaluation methods, considering methodological intricacies and the dynamic nature of learning behaviours, are crucial for assessing the reliability and applicability of these predictive models. Collectively, these components contribute to a holistic understanding of EDM's impact on enhancing the academic experience for students in this specific demographic. The intricate interplay between success prediction, input factors, student involvement factors, EDM tools, and evaluation methods forms the basis for evaluating EDM's efficacy in the context of sub-Saharan African STEM education.

6.7 Summary

This chapter discussed EDM's effectiveness in predicting academic performance among sub-Saharan African undergraduate STEM students, revealing a multifaceted exploration. The central theme revolves around success prediction at the degree level, where positive outcomes are anticipated throughout students' academic journeys within the STEM program. This focus is intricately linked to the examination of various input factors, encompassing previous and current class performance, demographic and socio-economic aspects, and dynamic e-learning behaviours. The significance of student involvement factors, including previous educational records and engagement levels, emerged as pivotal in shaping success predictions. Moreover, the study delved into the available EDM tools, highlighting the frequent use of WEKA and various algorithms such as Decision Trees, Artificial Neural Networks, Random Forests, Logistic Regression, k-Nearest Neighbour, Support Vector Machines, and Naïve Bayes. Notably, the underutilization of the k-Nearest Neighbours method prompts contemplation on its perceived fit within the predictive modelling context.

In evaluating the effectiveness of EDM, attention was directed towards methodological intricacies and the dynamic nature of learning behaviours. The assessment considered the

accuracy of predictive models. It highlighted the importance of exploring the contextual relevance of each tool and algorithm within the unique dynamics of sub-Saharan African STEM education. This chapter also discussed the evaluation methods employed, emphasizing the need to delve into specific methodologies that contribute to high accuracy in certain scenarios. The linkage between success prediction, input factors, student involvement, EDM tools, and evaluation methods is portrayed as a complex interplay that forms the basis for a comprehensive understanding of EDM's impact. This comprehensive approach contributes to the broader assessment of EDM's efficacy in enhancing the academic experience for students in the context of sub-Saharan African STEM education.

Chapter 7: Conclusion and Recommendations

7.1 Introduction

In concluding this extensive exploration, this chapter is the pivotal point where the multifaceted analyses come together. Reflecting upon the diverse dimensions of EDM effectiveness in predicting sub-Saharan African undergraduate STEM students' academic performance, this concluding segment highlights key insights, overarching themes, and implications for future research. The journey, spanning input factors, student involvement factors, EDM tools, evaluation methods, and success prediction, culminates in this synthesis, providing a comprehensive understanding of EDM's efficacy in the context of sub-Saharan African STEM education.

7.2 Summary of the Study

Using EDM to predict students' academic performance has become one of the most exciting research areas because of its significant influence on improving students' academic levels to support poorly performing students. This SLR identified the most frequently used factors that influence students' academic performance, and the most frequent EDM methods predict these factors influencing students' academic performance. This SLR used a methodology consisting of multiple steps and phases, as shown in Figure 3.2 in section 3.2.2.

This process started with the first phase, which consisted of planning the SLR by developing the research questions, eligibility criteria, and data extraction method. Moreover, the second phase involved steps when conducting the SLR, including searching and identifying the studies for further analysis, assessing their quality, and extracting and analysing the data. The 41 studies were selected and analysed to determine the gap and answer the research questions in section 1.3 of this SLR.

Furthermore, in the third phase of this research endeavour, the SLR unfolded, revealing comprehensive insights into the landscape of EDM in predicting academic performance among sub-Saharan African undergraduate STEM students. This phase sought to answer key questions that underpinned the investigation. The journey involves exploring the sub-research questions before seamlessly connecting the dots, ultimately converging to address the main research question.

Sub-Research Question 1: What key input factors are used in EDM to predict sub-Saharan African undergraduate STEM students' academic performance?

This SLR uncovered many input factors, including previous and current class performance, demographic, socio-economic, and e-learning behavioural data. These factors collectively contribute to the enhanced predictive models employed in the context of sub-Saharan African STEM education.

Sub-Research Question 2: What are the key student involvement factors used in EDM to predict sub-Saharan African undergraduate STEM students' academic performance?

The SLR revealed that student involvement factors, such as previous educational records and engagement levels, play a crucial role in success prediction. This emphasizes the dynamic interaction between students and the educational environment in the predictive modelling process.

Sub-Research Question 3: What are EDM's most frequently used tools and algorithms to predict sub-Saharan African undergraduate STEM students' academic performance?

The findings pointed to the prevalence of tools like WEKA and algorithms such as Decision trees, Artificial Neural Networks, Regression, Support Vector Machines, Naïve Bayes, and Random Forests. Understanding the toolkit is essential for researchers navigating the EDM landscape in sub-Saharan Africa.

Sub-Research Question 4: To what extent is EDM effective in predicting sub-Saharan African undergraduate STEM students' academic performance?

The findings revealed that evaluating EDM's effectiveness in predicting sub-Saharan African STEM students' academic performance involves a thorough analysis. The key methods include the confusion matrix and accuracy metrics. These assessment methods provide valuable insights into the model's predictive capabilities, enabling informed decision-making and the implementation of enhancements to support student success within the sub-Saharan African STEM education context.

Main-Research Question: How effective is EDM in predicting sub-Saharan African undergraduate STEM students' academic performance?

The findings showed that examining EDM's effectiveness in predicting academic performance among sub-Saharan African STEM students involves success prediction at the degree level, influenced by diverse input factors, student involvement factors, EDM tools like WEKA, and

various EDM algorithms such as Decision Trees, Artificial Neural Networks, Random Forests, Logistic Regression, k-Nearest Neighbour, Support Vector Machines, and Naïve Bayes. Evaluation methods addressing methodological intricacies are vital for reliability assessment. This collective analysis provides a holistic understanding of EDM's impact on enhancing the academic experience for sub-Saharan African STEM students, emphasizing the intricate interplay between success prediction, input factors, student involvement, EDM tools, and evaluation methods.

The results of this SLR not only provide answers to these questions but also offer a comprehensive narrative that illuminates the complexities, patterns, and implications embedded in using EDM for academic performance prediction among sub-Saharan African undergraduate STEM students. EDM researchers may benefit from the results of this SLR by applying it in further work, mainly the findings that show the most commonly used factors influencing students' academic performance and the most common EDM methods. Additionally, having a generic feature set that influences students' academic performances provides various possibilities for tailoring these factors to any context to discover new knowledge for improving decision-making. This study contributes to the EDM research in sub-Saharan Africa by providing a detailed report that describes the findings as qualitative values that help contextualize the quantitative findings and explain the patterns observed in the data, which may reveal new knowledge for educators, policymakers, and HEIs in sub-Saharan Africa to assist them in identifying students at risk of underperforming early and thereby implementing appropriate interventions to support their success.

7.3 Practical Implications and Recommendations

The researcher pointed out several recommendations for future research that could assist in building better EDM models for understanding students and their behaviours in courses based on the 41 selected studies on EDM. The following are the challenges of the selected studies analysed in this study that future researchers can explore:

Lack of unsupervised EDM techniques

Most researchers used supervised EDM techniques such as classification. However, there is a lack of unsupervised EDM approaches, such as clustering, that researchers use for students' academic performance prediction. This is a critical issue that EDM researchers can address in

the future. Future research can group students by their similar and observable patterns to provide tailored feedback for each cluster of students.

Data imbalance and overfitting problems

The imbalances in datasets are among the most frequently experienced issues in the selected literature. Most EDM methods perform better if the data is distributed appropriately. Hence, if the minor class is higher than the majority class or vice-versa, it will lead to an imbalanced dataset and decrease the model's performance. While some researchers have attempted to address this challenge through re-sampling techniques such as under-sampling and over-sampling, some shortcomings remain, such as overfitting risks due to increased noise in over-sampling cases. At the same time, valuable information will be lost in under-sampling cases. The re-sampling techniques are implemented at the algorithm level, so using actual data or ensemble methods rather than synthetic data can lead to higher model performances and the discovery of new knowledge. Hence, future EDM work needs to explore this issue in more detail to uncover the best strategies researchers can apply to achieve a balance in data and increase prediction accuracy.

Ethical issues

Undoubtedly, HEIs have continuously gathered and analysed vast amounts of data about students. However, ethical and privacy concerns have arisen concerning the transparency of the collection of this data. From the findings of this SLR, it is evident that few researchers focused on addressing these issues. They used strategies such as anonymizing students' identifying information and signing agreements on disclosing sensitive data to address these issues. Future research must recognize the ethical aspects by mentioning useful measures to ensure data privacy. Furthermore, privacy-related issues in ethics have resulted in the limited availability of robust online and free educational datasets, which has limited some studies. Hence, a call is also made to HEIs to start distributing open datasets to encourage future research studies in the EDM field.

Generalizability of Results

Various factors can reduce the generalizability of the EDM methods, especially when concentrating on a single program in HEIs. As mentioned before, EDM requires vast data for training prediction models to achieve high accuracy. Most of the selected studies in this analysis had dataset sizes of below 5000 records, which is not enough to draw generalizable

conclusions from the predictions. Hence, the more data available for predictions, the more accurate and precise generalizations are from the predicted outcomes. Future studies need to investigate how sample sizes can be increased in data collection so that better results can be obtained and for more generalisability. Additionally, every program varies in terms of how the assessments are conducted and marked. Thus, developing an efficient EDM model that considers factors that may be shared among students across various programs in HEIs is essential. Researchers should focus on designing EDM models that demonstrate adaptability to different assessment scenarios. Ensuring a balance between program-specific insights and generalizable findings can contribute to the broader applicability of predictive models.

7.4 Limitations of the Study

This research had some limitations. The study's scope was restricted to the sub-Saharan African region because there was insufficient EDM literature to conduct an adequate review in the South African context. Initial searches for studies were conducted using online databases, but this was not enough. Hence, other sources of evidence, such as manual searching, were also needed. While the papers in the literature search were thoroughly screened using the abstracts and conclusions, this may have resulted in rejected studies with good content due to poor abstracts and conclusions. It was noticeable throughout this study that EDM was identified as a method, technique, or framework. However, some papers selected in the literature search may have no proposed terminology for the method or framework used; hence, they were excluded. Furthermore, this study was mainly tailored to STEM fields. However, it was broad enough to be valuable and applicable to scholars from other research fields outside STEM to discover new knowledge.

7.5 Threats to Validity

In EDM research, the assessment of validity consists of internal and external validity (Namoun & Alshanqiti, 2021). In this study, the researcher followed the SLR protocols recommended by Moher *et al.* (2009) to increase the quality of the conclusions and lower the threats to the validity of this study. However, the validity of the results of this study was impacted mainly by the quality of the selected literature for analysis. The researcher found that the majority of the selected literature concentrated on the factors and models that only predicted students' academic performance successfully, which potentially introduced bias in the publications.

Studies with negative findings were few in the selected studies, which may have influenced the results of this SLR.

Another threat of validity that may have influenced the results of this study was missing important literature during the literature search process. The researcher followed the best practices for conducting SLR in EDM to minimize validity threats (Moher *et al.*, 2009). The researcher also used varied search terms or phrases to attain as many relevant studies as possible for each electronic bibliographic database.

Concerning external validity, assuming the same observation for different disciplines is dangerous since the majority of selected studies predicted students' academic performance in only STEM disciplines. Moreover, the findings should be treated cautiously, especially concerning generalization to other HEIs worldwide. Most studies were conducted in South Africa and Nigeria, amongst other sub-Saharan African countries in the selected studies, and the results of this SLR may not apply to developed countries.

7.6 Future Directions

This study strongly encourages EDM researchers to conduct future work in predicting students' academic performance, which is still in its infancy, specifically at the course and year level. EDM researchers should test the effectiveness of current EDM models on a vast number of datasets to judge their generalizability and validity accordingly. EDM researchers need to focus more on understanding how various factors influence students' academic performance prediction and how these factors improve the course and year levels predictions and decisions made to intervene early. There is a critical need for EDM researchers to explain the correlation between significant predicted factors and observed ones and how they influence students' academic performance predictions. Moreover, predicting students' academic performance should be extended to more disciplines, including humanities. Further studies need to also report negative findings besides publishing positive results, as this may open opportunities for new research directions.

7.7 Conclusion

In conclusion, this SLR delved into the intricate landscape of EDM with a specific focus on determining EDM effectiveness in predicting sub-Saharan African undergraduate STEM students' academic performance. The SLR followed a three-phase approach, commencing with

a detailed formulation of research questions, proceeding to the systematic search and selection of relevant studies, and culminating in a detailed analysis and synthesis of findings.

The overarching objective was to uncover the effectiveness of EDM in providing valuable insights into the academic trajectories of sub-Saharan African STEM students. Throughout the SLR, the core theme revolved around success prediction at the degree level, offering a comprehensive lens through which the multifaceted aspects of EDM's impact could be assessed. The exploration encompassed various input factors, including previous and current class performance, demographic, socio-economic, and e-learning behavioural dimensions. Student involvement factors, such as previous educational records and engagement levels, were identified as crucial elements shaping success predictions. The deployment of EDM tools, with a notable emphasis on WEKA and various algorithms such as Decision Trees, Artificial Neural Networks, Random Forests, Logistic Regression, k-Nearest Neighbour, Support Vector Machines, and Naïve Bayes, provided the technological backbone for predictive modelling. Evaluation methods were examined to uncover the intricacies contributing to predictive models' accuracy. The discussion on methodological intricacies highlighted the need for a comprehensive approach to assessing these models' broader applicability and reliability.

In summary, this SLR navigated the complex terrain of EDM, uncovering insights, limitations, and potential avenues for future research. The synthesis of findings contributes to a deeper understanding of how EDM can be effectively harnessed to enhance academic predictions for sub-Saharan African STEM students. This journey through the literature offers a foundation for continued exploration, refinement, and advancement in EDM within sub-Saharan African STEM education.

7.8 Data Availability

The tables of included studies and raw data are available in Appendix B.

7.9 Conflicting Interests

There were no conflicts of interest in this research.

7.10 Funding

No non-profit or commercial organisation funded this study.

References

Abdous, M. h., Wu, H., & Yen, C.-J. (2012). Using data mining for predicting relationships between online question theme and final grade. *Journal of Educational Technology & Society, 15*(3), 77.

Abe, E. N., & Chikoko, V. (2020). Exploring the factors that influence the career decision of STEM students at a university in South Africa. *International Journal of STEM Education, 7*(1), 1-14.

Abed, T., Ajoodha, R., & Jadhav, A. (2019). *A Programme Recommendation Engine to Improve Student Placement at a South African Higher-Education.* Paper presented at the 2020 International SAUPEC/RobMech/PRASA Conference.

Adak, M. F., Yumusak, N., & Taskin, H. (2016). *An elective course suggestion system developed in computer engineering department using fuzzy logic.* Paper presented at the 2016 International Conference on Industrial Informatics and Computer Systems (CIICS).

Adam, I. Y., Bello, H., Abdullahi, A. A., Dan-Azumi, M., & Abdullahi, N. (2020). Using Formative and Summative Assessments in Data Mining to Predict Students‟ Final Grades. *International Research Journal of Innovations in Engineering and Technology, 04*(11), 43-49. doi:10.47001/irjiet/2020.411006

Adams, J., Hillier-Brown, F. C., Moore, H. J., Lake, A. A., Araujo-Soares, V., White, M., & Summerbell, C. (2016). Searching and synthesising 'grey literature' and 'grey information' in public health: critical reflections on three case studies. *Systematic reviews, 5*(1), 1-11.

Adams, R. J., Smart, P., & Huff, A. S. (2017). Shades of grey: guidelines for working with the grey literature in systematic reviews for management and organizational studies. *International Journal of Management Reviews, 19*(4), 432-454.

Adejo, O., & Connolly, T. (2017). An integrated system framework for predicting students' academic performance in higher educational institutions. *International Journal of Computer Science and Information Technology (IJCSIT), 9*(3), 149-157.

Adekitan, A. I., Adewale, A. A., & Olaitan, A. (2019). Determining the operational status of a Three Phase Induction Motor using a predictive data mining model. *International Journal of Power Electronics and Drive System, 10*(1), 91-103.

Adekitan, A. I., & Noma-Osaghae, E. (2019). Data mining approach to predicting the performance of first year student in a university using the admission requirements. *Education and Information Technologies, 24*(2), 1527-1543.

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

Adekitan, A. I., & Salau, O. (2019). Toward an improved learning process: the relevance of ethnicity to data mining prediction of students' performance. *SN Applied Sciences, 2*(1), 1-15. doi:10.1007/s42452-019-1752-1

Afeni, B. O., Oloyede, I. A., & Okurinboye, D. (2019). Students' Performance Prediction Using Classsification Algorithms. *Journal of Advances in Mathematics and Computer Science, 30*(2), 1-9. doi:10.9734/jamcs/2019/45438

Agarwal, S., Pandey, G., & Tiwari, M. (2012). Data mining in education: data classification and decision tree approach. *International Journal of e-Education, e-Business, e-Management and e-Learning, 2*(2), 140.

Akinola, S., & Abraham, B. (2017). A Comparative Analysis of Classification Techniques in Educational Data Mining Using Computer Programming Proficiency Indicators. *Nigerian Journal of Science Vol, 51*(2), 77-91.

Al-Shabandar, R., Hussain, A., Laws, A., Keight, R., Lunn, J., & Radi, N. (2017). *Machine learning approaches to predict learning outcomes in Massive open online courses.* Paper presented at the 2017 International joint conference on neural networks (IJCNN).

Al Luhaybi, M. (2021). *Explainable machine learning for educational data.* (Doctoral Dissertation). Brunel University London, United Kingdom.

Al Riyami, T. (2015). Main approaches to educational research. *International Journal of Innovation and Research in Educational Sciences, 2*(5), 412-416.

Alban, M., & Mauricio, D. (2019). Predicting university dropout through data mining: A Systematic Literature. *Indian Journal of Science and Technology, 12*(4), 1-12.

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques. *Education Sciences, 11*(9), 552. Retrieved from https://www.mdpi.com/2227-7102/11/9/552

Alharahsheh, H. H., & Pius, A. (2020). A review of key paradigms: Positivism VS interpretivism. *Global Academic Journal of Humanities and Social Sciences, 2*(3), 39-43.

Alom, B. M., & Courtney, M. (2018). Educational data mining: a case study perspectives from primary to university education in australia. *International Journal of Information Technology and Computer Science, 10*(2), 1-9.

Alturki, S., Hulpuș, I., & Stuckenschmidt, H. (2022). Predicting academic outcomes: A survey from 2007 till 2018. *Technology, Knowledge and Learning*, 1-33.

Alwarthan, S. A., Aslam, N., & Khan, I. U. (2022). Predicting Student Academic Performance at Higher Education Using Data Mining: A Systematic Review. *Applied Computational Intelligence & Soft Computing, 2022*.

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education, 17*(1), 1-21.

Ashenafi, M. M. (2017). A Comparative Analysis of Selected Studies in Student Performance Prediction. *International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol, 7*.

Astin, A. W. (1984). Student involvement: A developmental theory for higher education. *Journal of college student personnel, 25*(4), 297-308.

Astin, A. W. (2012). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education*: Rowman & Littlefield Publishers.

Aulck, L., Aras, R., Li, L., L'Heureux, C., Lu, P., & West, J. (2017). STEM-ming the Tide: Predicting STEM attrition using student transcript data. *arXiv preprint arXiv:1708.09344*.

Bard, K. (2016). *Successful rural community college students: examining the association of student demographics, high school environmental variables, and high school outcome variables on community college degree attainment*: University of Missouri-Columbia.

Bassi, J. S., Dada, E. G., Hamidu, A. A., & Elijah, M. D. (2019). Students Graduation on Time Prediction Model Using Artificial Neural Network. *IOSR J. Comput. Eng, 21*(3), 28-35.

Bawack, R. E., & Kamdjoug, J. R. K. (2020). The role of digital information use on student performance and collaboration in marginal universities. *International Journal of Information Management, 54*, 102179.

Bawah, F. U., & Ussiph, N. (2018). Appraisal of the Classification Technique in Data Mining of Student Performance using J48 Decision Tree, K-Nearest Neighbor and Multilayer Perceptron Algorithms. *International Journal of Computer Applications, 179*(33), 39-46.

Bellefontaine, S. P., & Lee, C. M. (2014). Between black and white: Examining grey literature in meta-analyses of psychological research. *Journal of Child and Family Studies, 23*(8), 1378-1388.

Bengesai, A. V., & Pocock, J. (2021). Patterns of persistence among engineering students at a South African university: A decision tree analysis. *South African Journal of Science, 117*(3-4), 1-9.

bin Roslan, M. H., & Chen, C. J. (2022). Educational Data Mining for Student Performance Prediction: A Systematic Literature Review (2015-2021). *International journal of emerging technologies in learning, 17*(5).

Bokana, K., & Tewari, D. (2014). Determinants of student success at a South African university: An econometric analysis. *The Anthropologist, 17*(1), 259-277.

Bokgoshi, L., Jadhav, A., & Ajoodha, R. (2021). *Predicting Students That Are At Risk Of Not Graduating In Record Time.* The University of the Witwatersrand, Johannesburg, South Africa.

Booth, A. M., Wright, K. E., & Outhwaite, H. (2010). Centre for Reviews and Dissemination databases: value, content, and developments. *International journal of technology assessment in health care, 26*(4), 470-472.

Bornmann, L. (2014). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of informetrics, 8*(4), 895-903.

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology, 66*(11), 2215-2222.

Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., & Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of systems and software, 80*(4), 571-583.

Buraimoh, E. F., Ajoodha, R., & Padayachee, K. (2021). *Predicting Student Success Using Student Engagement in the Online Component of a Blended-Learning Course.* (MS Thesis). Univ. Witwatersrand, Johannesburg, South Africa.

Bussu, A., Detotto, C., & Serra, L. (2019). Indicators to prevent university drop-out and delayed graduation: an Italian case. *Journal of Applied Research in Higher Education*.

Cantabella, M., Martínez-España, R., Ayuso, B., Yáñez, J. A., & Muñoz, A. (2019). Analysis of student behavior in learning management systems through a Big Data framework. *Future Generation Computer Systems, 90*, 262-272.

Caruth, G. (2018). 'Student engagement, retention, and motivation: Assessing academic success in today's college students,''Participatory Educ. In: Res.

Chacon, F., Spicer, D., & Valbuena, A. (2012). Analytics in support of student retention and success. *Research Bulletin, 3*, 1-9.

Chaka, C. (2021). Educational Data Mining, Student Academic Performance Prediction, Prediction Methods, Algorithms and Tools: An Overview of Reviews. *Journal of e-Learning and Knowledge Society, 18(2)*, 58-69.

Chang, P.-C., Lin, C.-H., & Chen, M.-H. (2016). A hybrid course recommendation system by integrating collaborative filtering and artificial immune systems. *Algorithms, 9*(3), 47.

Chanthiran, M., Hishamuddin, M., Ibrahim, A. B., & Mariappan, P. (2022). *A Systematic Literature Review with Bibliometric Meta-Analysis of Text Visualization in Education.*

Paper presented at the Proceedings of the UR International Conference on Educational Sciences.
Chapman, P., Clinton, J., Ncr, R. K., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999). CRISP-DM 1.0.
Charmaz, K. (2014). *Constructing grounded theory*: sage.
Chigbu, B. I., & Nekhwevha, F. H. (2021). High school training outcome and academic performance of first-year tertiary institution learners-Taking'Input-Environment-Outcomes model'into account. *Heliyon, 7*(7), e07700.
Christy, C., Parimala, M. M., & Prema, M. (2018). The Review on Data Mining Techniques and its Applications. *Database, 7*(01).
Chuddher, B. A. (2015). *A novel knowledge discovery based approach for supplier risk scoring with application in the HVAC industry.* Brunel University London, United Kingdom.
Chweya, R., Shamsuddin, S. M., Ajibade, S.-S. M., & Moveh, S. (2020). A literature review of student performance prediction in E-learning environment. *Journal of Science, Engineering, Technology, and Management ISSN*, 9989-7858.
Cios, K. J., & Kurgan, L. A. (2005). Trends in data mining and knowledge discovery. In *Advanced techniques in knowledge discovery and data mining* (pp. 1-26): Springer.
Coronel, C., & Morris, S. (2016). *Database systems: design, implementation, & management*: Cengage Learning.
Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*: Sage publications.
Daramola, O., Emebo, O., Afolabi, I., & Ayo, C. (2014). Implementation of an intelligent course advisory expert system. *International Journal of Advanced Research in Artificial Intelligence, 3*(5), 6-12.
Darvas, P., Gao, S., Shen, Y., & Bawany, B. (2017). *Sharing higher education's promise beyond the few in sub-Saharan Africa*: World Bank Publications.
Davidson, T. (2019). Black-box models and sociological explanations: Predicting high school grade point average using neural networks. *Socius, 5*, 2378023118817702.
Dlamini, N., Marebane, S., & Makhubela, J. (2020). *Mining Campus Transfer Request Data.* Paper presented at the 2020 7th International Conference on Soft Computing & Machine Intelligence (ISCMI).
Driscoll, D. L. (2011). Introduction to primary research: Observations, surveys, and interviews. *Writing spaces: Readings on writing, 2*(2011), 153-174.
Du, X., Yang, J., Hung, J.-L., & Shelton, B. (2020). Educational data mining: a systematic review of research and emerging trends. *Information Discovery and Delivery*.
Dumond, E. J., & Johnson, T. W. (2013). Managing university business educational quality: ISO or AACSB? *Quality Assurance in Education*.
Durning, S. J., & Carline, J. D. (2015). *Review Criteria for Research Manuscripts: A Joint Project of Academic Medicine and the Group on Educational Affairs' Medical Education Scholarship Research and Evaluation Section*: Association of American Medical Colleges.
Dutt, A., Aghabozrgi, S., Ismail, M. A. B., & Mahroeian, H. (2015). Clustering algorithms applied in educational data mining. *International Journal of Information and Electronics Engineering, 5*(2), 112.
Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access, 5*, 15991-16005.
Ekubo, E. (2020). *Predictive system for characterizing low performance of Undergraduate students using machine learning techniques.* North-West University (South Africa),

Ekubo, E. A., & Esiefarienrhe, M. B. (2019). *Attributes of Low Performing Students in E-Learning System Using Clustering Technique*. Paper presented at the 2019 International Conference on Computational Science and Computational Intelligence (CSCI).

Fagbola, T. M., Adeyanju, I. A., Olaniyan, O., Esan, A., Omodunbi, B., Oloyede, A., & Egbetola, F. (2019). Development of mobile-interfaced machine learning-based predictive models for improving students performance in programming courses. *arXiv preprint arXiv:1901.06252*.

Fan, X., & Thompson, B. (2001). Confidence intervals for effect sizes: Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial. *Educational and Psychological Measurement, 61*(4), 517-531.

Farhan, R. (2019). Understanding postmodernism: Philosophy and culture of postmodern. *Journal International Social Sciences and Education,(October)*, 1-11.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM, 39*(11), 27-34.

Gajadhur, S. (2021). *Prototype learning analytics dashboard (LAD) for an introductory statistics course at UCT.* UCT, Cape Town, South Africa.

Gatsheni, B. N., & Katambwa, O. N. (2018). The Design of Predictive Model for the Academic Performance of Students at University Based on Machine Learning. *J. of Electrical Engineering, 6*(4). doi:10.17265/2328-2223/2018.04.006

Girma, S. (2019). *Developing a Predictive Model to Determine Higher Education Students' Academic Status Using Data Mining Technology.* (Doctoral dissertation). St. Mary's University, Ethiopia.

Given, L. M. (2008). *The Sage encyclopedia of qualitative research methods*: Sage publications.

Gray, D. E. (2016). Doing research in the real world. In: Los Angeles: SAGE.

Guzmán-Valenzuela, C., Gómez-González, C., Tagle, A. R.-M., & Lorca-Vyhmeister, A. (2021). Learning analytics in higher education: a preponderance of analytics but very little learning? *International Journal of Educational Technology in Higher Education, 18*(1), 1-19.

Hasan, R., Palaniappan, S., Mahmood, S., Abbas, A., Sarker, K. U., & Sattar, M. U. (2020). Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. *Applied Sciences, 10*(11), 3894.

Hasan, R., Palaniappan, S., Mahmood, S., Shah, B., Abbas, A., & Sarker, K. U. (2019). Enhancing the teaching and learning process using video streaming servers and forecasting techniques. *Sustainability, 11*(7), 2049.

Hashima, A. S., Hamoud, A. K., & Awadh, W. A. (2018). Analyzing students' answers using association rule mining based on feature selection. *Journal of Southwest Jiaotong University, 53*(5).

Haverila, M. J., Haverila, K., & McLaughlin, C. (2020). Variables affecting the retention intentions of students in higher education institutions: A comparison between international and domestic students. *Journal of International Students, 10*(2), 358-382.

Hegazi, M. O., & Abugroon, M. A. (2016). The state of the art on educational data mining in higher education. *International Journal of Computer Trends and Technology, 31*(1), 46-56.

Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., Murray, D. J., & Long, Q. (2018). Predicting academic performance by considering student heterogeneity. *Knowledge-Based Systems, 161*, 134-146.

Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (2019). *Cochrane handbook for systematic reviews of interventions*: John Wiley & Sons.

Hofstede, G. (2011). Dimensionalizing cultures: The Hofstede model in context. *Online readings in psychology and culture, 2*(1), 8.

Huebner, R. A. (2013). A Survey of Educational Data-Mining Research. *Research in higher education journal, 19*.

Hussain, M., Zhu, W., Zhang, W., & Abidi, S. M. R. (2018). Student engagement predictions in an e-learning system and their impact on student course assessment scores. *Computational intelligence and neuroscience, 2018*.

Imtiaz, S., Bano, M., Ikram, N., & Niazi, M. (2013). *A tertiary study: experiences of conducting systematic literature reviews in software engineering.* Paper presented at the Proceedings of the 17th international conference on evaluation and assessment in software engineering.

Inyang, U. G., Eyoh, I. J., Robinson, S. A., & Udo, E. N. (2019). Visual association analytics approach to predictive modelling of students' academic performance. *International Journal of Modern Education and Computer Science, 11*(12), 1.

Iqbal, Z., Qadir, J., Mian, A. N., & Kamiran, F. (2017). Machine learning based student grade prediction: A case study. *arXiv preprint arXiv:1708.08744*.

Iyanda, A. R., Ninan, O. D., Ajayi, A. O., & Anyabolu, O. G. (2018). Predicting Student Academic Performance in Computer Science Courses: A Comparison of Neural Network Models. *International Journal of Modern Education and Computer Science, 10*(6), 1-9. doi:10.5815/ijmecs.2018.06.01

Jaafar, F. M., Hashim, R. A., & Ariffin, T. F. T. (2012). Malaysian university student learning involvement scale (MUSLIS): Validation of a student engagement model. *Malaysian Journal of Learning and Instruction, 9*, 15-30.

Jacob, J., Jha, K., Kotak, P., & Puthran, S. (2015). *Educational data mining techniques and their applications.* Paper presented at the 2015 International Conference on Green Computing and Internet of Things (ICGCIoT).

Jembere, E., Rawatlal, R., & Pillay, A. W. (2017). *Matrix factorisation for predicting student performance.* Paper presented at the 2017 7th World Engineering Education Forum (WEEF).

Kabakchieva, D. (2012). Student performance prediction by using data mining classification algorithms. *International journal of computer science and management research, 1*(4), 686-690.

Kanetaki, Z., Stergiou, C., Bekas, G., Troussas, C., & Sgouropoulou, C. (2022). Evaluating Remote Task Assignment of an Online Engineering Module through Data Mining in a Virtual Communication Platform Environment. *Electronics, 11*(1), 158.

Kasisi, R. (2019). *An Artificial Neural Network Decision Support Model For University Students Progression.* (Doctoral dissertation). Kca University, Kenya.

Keele, S. (2007). Guidelines for performing systematic literature reviews in software engineering. In: Technical report, ver. 2.3 ebse technical report. ebse.

Kelly, L. M., & Cordeiro, M. (2020). Three principles of pragmatism for research on organizational processes. *Methodological innovations, 13*(2), 2059799120937242.

Khan, A., & Ghosh, S. K. (2021). Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Education and Information Technologies, 26*(1), 205-240.

Khan, A., Ghosh, S. K., Ghosh, D., & Chattopadhyay, S. (2021). Random wheel: An algorithm for early classification of student performance with confidence. *Engineering Applications of Artificial Intelligence, 102*, 104270.

Khan, S. N. (2014). Qualitative research method: Grounded theory. *International journal of business and management, 9*(11), 224-233.

Khedr, A. E., & El Seddawy, A. I. (2015). A proposed data mining framework for higher education system. *International Journal of Computer Applications, 113*(7).

Kim, Yoon, M., Jo, I.-H., & Branch, R. M. (2018). Learning analytics to support self-regulated learning in asynchronous online courses: A case study at a women's university in South Korea. *Computers & Education, 127*, 233-251.

Kim, D. (2017). The impact of learning management systems on academic performance: Virtual Competency and student Involvement. *Journal of Higher Education Theory and Practice, 17*(2), 23-35.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology, 51*(1), 7-15.

Kovač, R., & Oreški, D. (2018). *Educational data driven decision making: early identification of students at risk by means of machine learning.* Paper presented at the Central European Conference on Information and Intelligent Systems.

Kroeze, J. H. (2012). Postmodernism, interpretivism, and formal ontologies. In *Research methodologies, innovations and philosophies in software systems engineering and information systems* (pp. 43-62): IGI Global.

Kumar, V., & Chadha, A. (2012). Mining association rules in student's assessment data. *International Journal of Computer Science Issues (IJCSI), 9*(5), 211.

Langa, H. M. (2018). *Modelling Student Success or Failure in Electrical Engineering at the Vaal University of Technology Using Machine Learning Techniques.* (Doctorate Dissertation). University of Johannesburg South Africa.

Lee, M. J., & McLoughlin, C. (2010). *Web 2.0-based e-learning: Applying social informatics for tertiary teaching: Applying social informatics for tertiary teaching*: IGI Global.

Lefebvre, C., Glanville, J., Briscoe, S., Littlewood, A., Marshall, C., Metzendorf, M. I., . . . Thomas, J. (2019). Searching for and selecting studies. *Cochrane Handbook for systematic reviews of interventions*, 67-107.

Lei, Yang, M., & Cai, Y. (2017). Educational data mining for decision-making: A framework based on student development theory. *Advances in Engineering Research, 117*, 628-641.

Li, G., & Xu, J. (2023). *The Explore of Virtual Learning Environment: A Study of Higher Education.* Paper presented at the 2023 11th International Conference on Information and Education Technology (ICIET).

Liang, J., Li, C., & Zheng, L. (2016). *Machine learning application in MOOCs: Dropout prediction.* Paper presented at the 2016 11th International Conference on Computer Science & Education (ICCSE).

Lottering, R., Hans, R., & Lall, M. (2020). A Machine Learning Approach to Identifying Students at Risk of Dropout: A Case Study. *International Journal of Advanced Computer Science and Applications, 11*(10), 417-422.

Magini, E. B., Matos, L. d. O., Curtarelli, R. B., Sordi, M. B., Magrin, G. L., Flores-Mir, C., . . . Cruz, A. C. C. (2022). Simvastatin Embedded into Poly (Lactic-Co-Glycolic Acid)-Based Scaffolds in Promoting Preclinical Bone Regeneration: A Systematic Review. *Applied Sciences, 12*(22), 11623.

Mahood, Q., Van Eerd, D., & Irvin, E. (2014). Searching for grey literature for systematic reviews: challenges and benefits. *Research synthesis methods, 5*(3), 221-234.

Makombe, F., & Lall, M. (2020). A predictive model for the determination of academic performance in private higher education institutions. *International Journal of Advanced Computer Science and Applications (IJACSA), 11*(9).

Makombe, F., & Lall, M. (2020). A predictive model for the determination of academic performance in private higher education institutions. *International Journal of Advanced Computer Science and Applications, 11*(9).

Mallett, R., Hagen-Zanker, J., Slater, R., & Duvendack, M. (2012). The benefits and challenges of using systematic reviews in international development research. *Journal of development effectiveness, 4*(3), 445-455.

Maphosa, M., & Maphosa, V. (2020). *Educational data mining in higher education in sub-saharan africa: a systematic literature review and research agenda.* Paper presented at the Proceedings of the 2nd International Conference on Intelligent and Innovative Computing Applications.

Mashiloane, L. (2016). *Educational data mining (EDM) in a South African University: a longitudinal study of factors that affect the academic performance of computer science I students.* (MSC). University of Witwatersrand, Johannesburg.

Matafeni, G. (2017). *Predicting the Completion of a Students Science Degree based only on their First-year Marks.* (Honours Dissertation). Univeristy of the Witwatersrand, Johannesburg, South Africa.

Matsebula, F., & Mnkandla, E. (2016). *Information systems innovation adoption in higher education: Big data and analytics.* Paper presented at the 2016 International Conference on Advances in Computing and Communication Engineering (ICACCE).

Mauthner, N. S. (2020). Research philosophies and why they matter. In *How to Keep your Doctorate on Track*: Edward Elgar Publishing.

Mayilvaganan, M., & Kalpanadevi, D. (2014). *Comparison of classification techniques for predicting the performance of students academic environment.* Paper presented at the 2014 International Conference on Communication and Network Technologies.

McCollum, G. (2018). *Collegiate Recreation Participation and Student Retention, Progression, and Graduation.* (Doctoral dissertation). Georgia Southern University, United States.

Meghji, A. F., Mahoto, N. A., Unar, M. A., & Shaikh, M. A. (2018). *Analysis of student performance using EDM methods.* Paper presented at the 2018 5th International Multi-Topic ICT Conference (IMTIC).

Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access, 8*, 55462-55470.

Mgala, M. (2016). *Investigating prediction modelling of academic performance for students in rural schools in Kenya.* (Doctoral Dissertation). UCT, South Africa.

Mgala, M., & Mbogho, A. (2015). *Data-driven intervention-level prediction modeling for academic performance.* Paper presented at the Proceedings of the Seventh International Conference on Information and Communication Technologies and Development.

Miguéis, V. L., Freitas, A., Garcia, P. J., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems, 115*, 36-51.

Mirza, S., Mittal, S., & Zaman, M. (2016). A review of data mining literature. *International Journal of Computer Science and Information Security (IJCSIS), 14*(11), 437-442.

Mngadi, N. (2020). *A theoretical model to predict undergraduate learner attrition using background, individual, and schooling attributes.* (PhD diss). University of the Witwatersrand, Johannesburg.

Modi, K., & Oza, B. (2016). Outlier analysis approaches in data mining. *Ijirt, 3*(7), 6-12.

Moher, D. (2018). Reporting guidelines: doing better for readers. In (Vol. 16, pp. 1-3): Springer.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group*, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine, 151*(4), 264-269.

Molina-Azorin, J. F. (2012). Mixed methods research in strategic management: Impact and applications. *Organizational Research Methods, 15*(1), 33-56.

Mushi, P. K., & Ngondya, D. (2021). Prediction of mathematics performance using educational data mining techniques. *International Journal of Advanced Computer Research, 11*(56), 83.

Mutanu, L., & Machoka, P. (2019). *Enhancing Computer Students' Academic Performance through Predictive Modelling-A Proactive Approach.* Paper presented at the 2019 14th International Conference on Computer Science & Education (ICCSE).

Muthukrishnan, S., Govindasamy, M., & Mustapha, M. (2017). Systematic mapping review on student's performance analysis using big data predictive model. *Journal of fundamental and applied sciences, 9*(4S), 730-758.

Naidoo, J. T. (2019). *Using Background, Individual and Pre-College Attributes for Student Placement in the Earth Sciences.* (Honours Dissertation). University of the Witwatersrand, Johannesburg.

Namoun, A., & Alshanqiti, A. (2021). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences, 11*(1), 237.

Naseem, M. (2021). *Analysis of discipline level university student attrition using data mining.* (Masters Dissertation). University of the South Pacific, South Pacific.

Ndou, N., Ajoodha, R., & Jadhav, A. (2020). A Case Study to Enhance Student Support Initiatives Through Forecasting Student Success in Higher-Education. *Advances in Science, Technology and Engineering Systems Journal, 6*(1), 230-241. doi:10.25046/aj060126

Ndou, N., Ajoodha, R., & Jadhav, A. (2020). *Educational data-mining to determine student success at higher education institutions.* Paper presented at the 2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC).

Nguyen, V. A., Nguyen, H.-H., Nguyen, D.-L., & Le, M.-D. (2021). A course recommendation model for students based on learning outcome. *Education and Information Technologies*, 1-27.

Niazi, M. (2015). Do systematic literature reviews outperform informal literature reviews in the software engineering domain? An initial case study. *Arabian Journal for Science and Engineering, 40*(3), 845-855.

Nudelman, Z., Moodley, D., & Berman, S. (2018). *Using bayesian networks and machine learning to predict computer science success.* Paper presented at the Annual Conference of the Southern African Computer Lecturers' Association.

Nwosu, J., John, H., Izang, A., & Akorede, O. (2018). Assessment of information and communication technology (ICT) competence and literacy skills among undergraduates as a determinant factor of academic achievement. *Educational Research and Reviews, 13*(15), 582-589.

Ofori, F., Maina, E., & Gitonga, R. (2020). Using Machine Learning Algorithms to Predict Studentsâ€™ Performance and Improve Learning Outcome: A Literature Based Review. *Journal of Information and Technology, 4*(1), 33-55.

Okike, E. U., & Mogorosi, M. (2020). Educational data mining for monitoring and improving academic performance at university levels. *International Journal of Advanced Computer Science and Applications, 11*(11).

Okoli, C., & Schabram, K. (2010). A guide to conducting a systematic literature review of information systems research. *Sprouts: Working Papers on Information Systems, 10(26)*. Retrieved from http://sprouts.aisnet.org/10-26

Olaniyi, A. S., Kayode, S. Y., Abiola, H. M., Tosin, S.-I. T., & Babatunde, A. N. (2017). STUDENT'S PERFORMANCE ANALYSIS USING DECISION TREE ALGORITHMS. *Annals. Computer Science Series, 15*(1), 55-62.

Olaya, D., Vásquez, J., Maldonado, S., Miranda, J., & Verbeke, W. (2020). Uplift Modeling for preventing student dropout in higher education. *Decision Support Systems, 134*, 113320.

Osmanbegovic, E., & Suljic, M. (2012). Data mining approach for predicting student performance. *Economic Review: Journal of Economics and Business, 10*(1), 3-12.

Paez, A. (2017). Gray literature: An important resource in systematic reviews. *Journal of Evidence-Based Medicine, 10*(3), 233-240.

Page, M. J., Shamseer, L., Altman, D. G., Tetzlaff, J., Sampson, M., Tricco, A. C., . . . Sarkis-Onofre, R. (2016). Epidemiology and reporting characteristics of systematic reviews of biomedical research: a cross-sectional study. *PLoS medicine, 13*(5), e1002028.

Pallant, J. (2020). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS*. UK: Routledge.

Patil, A. P., Ganesan, K., & Kanavalli, A. (2017). *Effective deep learning model to predict student grade point averages.* Paper presented at the 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC).

Patton, K., Coleman, D., & Kay, L. W. (2019). High-impact honors practices: Success outcomes among honors and comparable high-achieving non-honors students at Eastern Kentucky University. *Chapters from NCHC Monographs Series, 51*.

Pawson, R. (2013). *The science of evaluation: a realist manifesto*: sage.

Pelletier, K., Brown, M., Brooks, D. C., McCormack, M., Reeves, J., Arbino, N., . . . Gibson, R. (2021). 2021 EDUCAUSE Horizon Report Teaching and Learning Edition.

Pham, L. T. M. (2018). Qualitative approach to research a review of advantages and disadvantages of three paradigms: Positivism, interpretivism and critical inquiry. *University of Adelaide*.

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. *Knowledge discovery in databases*, 229-238.

Poh, N., & Smythe, I. (2014). *To what extend can we predict students' performance? A case study in colleges in South Africa.* Paper presented at the 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM).

Prabowo, H., Hidayat, A. A., Cenggoro, T. W., Rahutomo, R., Purwandari, K., & Pardamean, B. (2021). Aggregating time series and tabular data in deep learning model for university students' gpa prediction. *IEEE Access, 9*, 87370-87377.

Pratheebha, M. T., Indhumathi, M. V., & Megala, S. (2021). An empirical study on data mining techniques and its applications. *Int J Softw Hardware Res Eng, 9(4)*, 23-31.

Pratt, I. S., Harwood, H. B., Cavazos, J. T., & Ditzfeld, C. P. (2019). Should I stay or should I go? Retention in first-generation college students. *Journal of College Student Retention: Research, Theory & Practice, 21*(1), 105-118.

Pressman, R. S. (2015). “Software Engineering–A Practitioner‟ s Approach”, Mc Graw-Hill International Edition, 2010. In: Bharathidasan Engineering College.

Raghavjee, R., Subramaniam, P. R., & Govender, I. (2021). Learning Analytics in Higher Education. In *Perspectives on ICT4D and Socio-Economic Growth Opportunities in Developing Countries* (pp. 398-431): IGI Global.

Ramaano, T., Ajoodha, R., & Jadhav, A. (2021). Different models relating prior computer experience with performance in first year computer science. In *Interdisciplinary Research in Technology and Management* (pp. 652-659): CRC Press.

Ramesh, V., Parkavi, P., & Ramar, K. (2013). Predicting student performance: a statistical and data mining approach. *International Journal of Computer Applications, 63*(8).

Ramokolo, P. L. (2021). *Application of discrete-time survival analysis techniques in modelling student dropout: a case of engineering students at Tshwane University of Technology, South Africa.* (Masters Dissertation). Tshwane University of Technology, South Africa.

Ray, S., & Saeed, M. (2018). Applications of educational data mining and learning analytics tools in handling big data in higher education. In *Applications of big data analytics* (pp. 135-160): Springer.

Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., & Koffel, J. B. (2021). PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. *Systematic reviews, 10*(1), 1-19.

Rodríguez-Triana, M. J., Prieto, L. P., Vozniuk, A., Boroujeni, M. S., Schwendimann, B. A., Holzer, A., & Gillet, D. (2017). Monitoring, awareness and reflection in blended technology enhanced learning: a systematic review. *International Journal of Technology Enhanced Learning, 9*(2-3), 126-150.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3*(1), 12-27.

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), e1355.

Roy, S., & Garg, A. (2017). *Predicting academic performance of student using classification techniques.* Paper presented at the 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON).

Roy, S., & Singh, S. N. (2017). *Emerging trends in applications of big data in educational data mining and learning analytics.* Paper presented at the 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence.

Ryan, G. (2018). Introduction to positivism, interpretivism and critical theory. *Nurse researcher, 25*(4), 41-49.

Saa, A. A., Al-Emran, M., & Shaalan, K. (2019). Factors affecting students’ performance in higher education: a systematic review of predictive data mining techniques. *Technology, Knowledge and Learning, 24*, 567-598.

Saa, A. A., Al-Emran, M., & Shaalan, K. (2019). *Mining student information system records to predict students’ academic performance.* Paper presented at the International conference on advanced machine learning technologies and applications.

Saheed, Y. K., Oladele, T. O., Akanni, A. O., & Ibrahim, W. M. (2018). Student performance prediction based on data mining classification techniques. *Nigerian Journal of Technology, 37*(4). doi:10.4314/njt.v37i4.31

Sani, N. S., Nafuri, A. F. M., Othman, Z. A., Nazri, M. Z. A., & Mohamad, K. N. (2020). Drop-out prediction in higher education among B40 students. *International Journal of Advanced Computer Science and Applications, 11*(11).

Saraswat, D. (2016). *Knowledge Discovery With Hybrid Data Mining Approach.* Dayalbagh Educational Institute,

Saunders, M. N., Lewis, P., Thornhill, A., & Bristow, A. (2015). Understanding research philosophy and approaches to theory development. Retrieved from https://www.researchgate.net/publication/330760964

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science, 181*, 526-534.

Sekaran, U., & Bougie, R. (2016). *Research methods for business: A skill building approach*: john wiley & sons.

Seota, S., Klein, R., & van Zyl, T. (2021). Modeling E-Behaviour, Personality and Academic Performance with Machine Learning. *Applied Sciences, 11*(22). doi:10.3390/app112210546

Setia, M. S. (2016). Methodology series module 3: Cross-sectional studies. *Indian journal of dermatology, 61*(3), 261.

Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science, 72*, 414-422.

Shingari, I., Kumar, D., & Khetan, M. (2017). A review of applications of data mining techniques for prediction of students' performance in higher education. *Journal of Statistics and Management Systems, 20*(4), 713-722.

Shuqfa, Z., & Harous, S. (2019). *Data mining techniques used in predicting student retention in higher education: A survey.* Paper presented at the 2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA).

Siemens, G., & Baker, R. S. d. (2012). *Learning analytics and educational data mining: towards communication and collaboration.* Paper presented at the Proceedings of the 2nd international conference on learning analytics and knowledge.

Singh, D. (2019). Understanding philosophical underpinnings of research with respect to various paradigms: Perspective of a research scholar. *Institute of Management, NIRMA University*.

Singh, H. P., & Alhulail, H. N. (2022). Predicting Student-Teachers Dropout Risk and Early Identification: A Four-Step Logistic Regression Approach. *IEEE Access, 10*, 6470-6482.

Singh, R., & Pal, S. (2020). Application of machine LearningAlgorithms to predict students performance. *International Journal of Advanced Science andTechnology, 29*(5), 7249-7261.

Slimani, T., & Lazzez, A. (2013). Sequential mining: patterns and algorithms analysis. *arXiv preprint arXiv:1311.0350*.

Smita, P. S., & Sharma, P. (2014). Use of data mining in various field: A survey paper. *IOSR Journal of Computer Engineering, 16*(3), 18-21.

Sokkhey, P., Navy, S., Tong, L., & Okazaki, T. (2020). Multi-models of educational data mining for predicting student performance in mathematics: a case study on high schools in Cambodia. *IEIE Transactions on Smart Processing and Computing, 9*(3), 217-229.

Soobramoney, R. (2021). *Early prediction of students at risk in a virtual learning environment using ensemble machine learning techniques.* (Masters Dissertation). Durban University of Technology, South Africa.

Sorour, S. E., Mine, T., Goda, K., & Hirokawa, S. (2014). *Predicting students' grades based on free style comments data by artificial neural network.* Paper presented at the 2014 IEEE Frontiers in Education Conference (FIE) Proceedings.

StatsSA. (2023). Beyond unemployment – Time-Related Underemployment in the SA labour market. Retrieved from https://www.statssa.gov.za/?p=16312

Stewart, L. A., Clarke, M., Rovers, M., Riley, R. D., Simmonds, M., Stewart, G., & Tierney, J. F. (2015). Preferred reporting items for a systematic review and meta-analysis of individual participant data: the PRISMA-IPD statement. *Jama, 313*(16), 1657-1665.

Sultan, M. E., Abd Rani, M. N., Radzuan, N. F. M., & Huay, L. (2021). Predictive Analytics on University Student Dropouts from Online Learning due to MCO. *Proc. Knowl. Manage. Int. Conf.(KMICe)*, 117-123.

Taodzera, T. (2018). *Predicting Student Performance Using Machine Learning Analytics.* (Masters Dissertation). University of Johannesburg, South Africa.

Taodzera, T., Twala, B., & Carroll, J. (2017). *Predicting engineering student success using machine learning.* Paper presented at the 4th Biennial Conference of the South African Society for Engineering Education.

Tegegne, A. K., & Alemu, T. A. (2018). Educational data mining for students' academic performance analysis in selected Ethiopian universities. *Information Impact: Journal of Information and Knowledge Management, 9*(2), 1-15. doi:10.4314/iijikm.v9i2.1

Thakar, P. (2015). Performance analysis and prediction in educational data mining: A research travelogue. *arXiv preprint arXiv:1509.05176*.

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., . . . Weeks, L. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Annals of internal medicine, 169*(7), 467-473.

Ugalde, B., & Venkateswaran, R. (2018). A Research Travelogue Towards Educational Data Mining. *International Journal of Computer Applications, 975*, 8887.

Uzoka, F.-M. E., Connolly, R., Schroeder, M., Khemka, N., & Miller, J. (2013). *Computing is not a rock band: student understanding of the computing disciplines.* Paper presented at the Proceedings of the 14th annual ACM SIGITE conference on Information technology education.

Vambe, W. T., & Sibanda, K. (2016). *Using Data Mining Techniques for the Prediction of Student Dropouts from University Science Programs.* University of Fort Hare,

Van Appel, V., & Durandt, R. (2019). Investigating possibilities of predictive mathematical models to identify at-risk students in the South African higher education context. *Perspectives in Education, 37*(2), 1-15.

Verner, J. M., Brereton, O. P., Kitchenham, B. A., Turner, M., & Niazi, M. (2014). Risks and risk mitigation in global software development: A tertiary study. *Information and software technology, 56*(1), 54-78.

Vo, T. N. C., & Nguyen, H. P. (2012). *A knowledge-driven educational decision support system.* Paper presented at the 2012 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future.

Waidor, T., & Akpojaro, J. (2019). The Use of Classification Algorithm for Forecasting the Academic Performance of Students of Biological Sciences, University of Africa, Toru-Orua. *African Scientist, 20*(2).

Wang, Ji, W., Liu, M., Wang, X., Weng, J., Deng, S., . . . Yuan, C.-a. (2018). Review on mining data from multiple data sources. *Pattern Recognition Letters, 109*, 120-128.

Wang, X. (2013). Modeling entrance into STEM fields of study among students beginning at community colleges and four-year institutions. *Research in Higher Education, 54*(6), 664-692.

Wanjau, S. K., & Muketha, G. M. (2018). Improving student enrollment prediction using ensemble classifiers. *International Journal of Computer Applications Technology and Research*.

Xiao, W., Ji, P., & Hu, J. (2021). RnkHEU: A Hybrid Feature Selection Method for Predicting Students' Performance. *Scientific Programming, 2021*.

Yaacob, W. W., Sobri, N. M., Nasir, S. M., Norshahidi, N., & Husin, W. W. (2020). *Predicting student drop-out in higher institution using data mining techniques.* Paper presented at the Journal of Physics: Conference Series.

Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Data mining applications: A comparative study for predicting student's performance. *arXiv preprint arXiv:1202.4815*.

Yadav, S. K., & Pal, S. (2012). Data mining: A prediction for performance improvement of engineering students using classification. *arXiv preprint arXiv:1203.3832*.

Yakubu, M. N., & Abubakar, A. M. (2021). Applying machine learning approach to predict students' performance in higher educational institutions. *Kybernetes, 51*(2), 916-934. doi:10.1108/k-12-2020-0865

Yang, F., & Li, F. W. (2018). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Computers & Education, 123*, 97-108.

Zacharis, N. Z. (2015). A multivariate approach to predicting student outcomes in web-enabled blended learning courses. *The Internet and Higher Education, 27*, 44-53.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1), 1-27.

Zohair, A., & Mahmoud, L. (2019). Prediction of Student's performance by modelling small dataset size. *International Journal of Educational Technology in Higher Education, 16*(1), 1-18.

Zollanvari, A., Kizilirmak, R. C., Kho, Y. H., & Hernández-Torrano, D. (2017). Predicting students' GPA and developing intervention strategies based on self-regulatory learning behaviors. *IEEE Access, 5*, 23792-23802.

Appendix A: Ethical clearance

Mr Langelihle Lucky Mhlongo (217071140)
School Of Man Info Tech &Gov
Pietermaritzburg

Dear Mr Langelihle Lucky Mhlongo,

Original application number: 00015252
Project title: Using educational data mining to predict sub-Saharan African science, technology, engineering, and mathematics students' academic performance: A systematic review

Exemption from Ethics Review

In response to your application received on 28 March 2022 , your school has indicated that the protocol has been granted **EXEMPTION FROM ETHICS REVIEW.**

Any alteration/s to the exempted research protocol, e.g., Title of the Project, Location of the Study, Research Approach and Methods must be reviewed and approved through an amendment/modification prior to its implementation. The original exemption number must be cited.

For any changes that could result in potential risk, an ethics application including the proposed amendments must be submitted to the relevant UKZN Research Ethics Committee. The original exemption number must be cited.

In case you have further queries, please quote the above reference number.

PLEASE NOTE:
Research data should be securely stored in the discipline/department for a period of 5 years.

I take this opportunity of wishing you everything of the best with your study.

Yours sincerely,

[redacted] 26/04/2022

Dr Vangeli Wiseman Gamede
Academic Leader Research
School Of Man Info Tech &Gov

Appendix B: Research Materials

Table A.1: PRISMA checklist

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	Cover page
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	Iv
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	Pg 3
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	Pg 5
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Pg 50
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	Pg 48
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Pg 48
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	Pg 52
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	Pg 53
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	Pg 51
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Pg 51
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked	Pg 51

Section and Topic	Item #	Checklist item	Location where item is reported
		independently, and if applicable, details of automation tools used in the process.	
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	Pg 51
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	Pg 51
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	Pg 55
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Pg 56
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	Pg 57
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	N/A
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	N/A
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	Pg 56
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	Pg 56
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	Pg 52
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	Pg 55
Study characteristics	17	Cite each included study and present its characteristics.	Pg 56
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	Pg 56
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	N/A
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	Pg 56

Section and Topic	Item #	Checklist item	Location where item is reported
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	N/A
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	N/A
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	N/A
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	Pg 56
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	Pg 56
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	Pg 122
	23b	Discuss any limitations of the evidence included in the review.	Pg 137
	23c	Discuss any limitations of the review processes used.	Pg 139
	23d	Discuss implications of the results for practice, policy, and future research.	Pg 140
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	Pg 48
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	Pg 48
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	Pg 48
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	Pg 141
Competing interests	26	Declare any competing interests of review authors.	Pg 141
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	Pg 141

Table A.2: Selected Studies

ID	Author and Year	Title	Country	Publication Type
[S1]	Abed *et al.* (2019)	A Programme Recommendation Engine to Improve Student Placement at a South African Higher-Education Institution	South Africa	Thesis/Dissertation
[S2]	Adam *et al.* (2020)	Using Formative and Summative Assessments in Data Mining to Predict Students' Final Grades	Nigeria	International Research Journal of Innovations in Engineering and Technology (IRJIET)
[S3]	Adekitan and Salau (2019)	The impact of engineering students' performance in the first three years on their graduation result using educational data mining	Nigeria	Heliyon
[S4]	Adekitan and Salau (2019)	Toward an improved learning process: the relevance of ethnicity to data mining prediction of students' performance	Nigeria	SN Applied Sciences
[S5]	Afeni *et al.* (2019)	Students' Performance Prediction Using Classification Algorithms	Nigeria	International Journal of Advanced Computer Science and Applications (IJACSA)
[S6]	Akinola and Abraham (2017)	A Comparative Analysis of Classification Techniques in Educational Data Mining Using Computer Programming Proficiency Indicators	Nigeria	Nigerian Journal of Science
[S7]	Bawah and Ussiph (2018)	Appraisal of the Classification Technique in Data Mining of Student Performance using J48 Decision Tree, k-Nearest Neighbor and Multilayer Perceptron Algorithms	Ghana	International Journal of Computer Applications
[S8]	Bokgoshi *et al.* (2021)	Predicting Students That Are At Risk Of Not Graduating In Record Time	South Africa	Thesis/Dissertation
[S9]	Buraimoh *et al.* (2021)	Predicting Student Success Using Student Engagement in the Online Component of a Blended-Learning Course	South Africa	Thesis/Dissertation
[S10]	Dlamini *et al.* (2020)	Mining Campus Transfer Request Data	South Africa	International Conference on Soft Computing & Machine Intelligence (ISCMI)
[S11]	Ekubo (2020)	Predictive system for characterizing low performance of Undergraduate students using machine learning techniques	South Africa	Thesis/Dissertation
[S12]	Ekubo and Esiefarienrhe (2019)	Attributes of low performing students in e-learning system using clustering technique	South Africa	International Conference on Computational Science and Computational Intelligence (CSCI)
[S13]	Fagbola *et al.* (2019)	Development of Mobile-Interfaced Machine Learning-Based Predictive Models for Improving Students' Performance in Programming Courses	Nigeria	International Journal of Advanced Computer Science and Applications (IJACSA)
[S14]	Gajadhur (2021)	Prototype Learning Analytics Dashboard (LAD) for an Introductory Statistics Course at UCT	South Africa	Thesis/Dissertation

[S15]	Gatsheni and Katambwa (2018)	The Design of Predictive Model for the Academic Performance of Students at University Based on Machine Learning	South Africa	Journal of Electrical Engineering
[S16]	Girma (2019)	Developing a Predictive Model to Determine Higher Education Students' Academic Status Using Data Mining Technology	Ethiopia	Thesis/Dissertation
[S17]	Inyang *et al.* (2019)	Visual Association Analytics Approach to Predictive Modelling of Students' Academic Performance	Nigeria	International Journal of Modern Education and Computer Science
[S18]	Iyanda *et al.* (2018)	Predicting Student Academic Performance in Computer Science Courses: A Comparison of Neural Network Models	Nigeria	International Journal of Modern Education & Computer Science
[S19]	Jembere *et al.* (2017)	Matrix Factorisation for Predicting Student Performance	South Africa	World Engineering Education Forum (WEEF)
[S20]	Kasisi (2019)	An artificial neural network decision support model for university students progression	Kenya	Thesis/Dissertation
[S21]	Langa (2018)	Modelling student success or failure in electrical engineering at the vaal university of technology using machine learning techniques	South Africa	Thesis/Dissertation
[S22]	Lottering *et al.* (2020)	A Machine Learning Approach to Identifying Students at Risk of Dropout: A Case Study	South Africa	International Journal of Advanced Computer Science and Applications (IJACSA)
[S23]	Makombe and Lall (2020)	A Predictive Model for the Determination of Academic Performance in Private Higher Education Institutions	South Africa	International Journal of Advanced Computer Science and Applications (IJACSA)
[S24]	Matafeni (2017)	Predicting the Completion of a Student's Science Degree based only on their First-year Marks	South Africa	Thesis/Dissertation
[S25]	Mngadi (2020)	A Theoretical Model to Predict Undergraduate Learner Attrition using Background, Individual, and Schooling Attributes	South Africa	Thesis/Dissertation
[S26]	Mushi and Ngondya (2021)	Prediction of mathematics performance using educational data mining techniques	Tanzania	International Journal of Advanced Computer Research
[S27]	Mutanu and Machoka (2019)	Enhancing Computer Students' Academic Performance through Predictive Modelling – A Proactive Approach	Kenya	International Conference on Computer Science & Education (ICCSE)
[S28]	Naidoo (2019)	Using Background, Individual and Pre-College Attributes for Student Placement in the Earth Sciences	South Africa	Thesis/Dissertation
[S29]	Ndou *et al.* (2020)	A Case Study to Enhance Student Support Initiatives Through Forecasting Student Success in Higher-Education	South Africa	Advances in Science, Technology and Engineering Systems Journal
[S30]	Okike and Mogorosi (2020)	Educational Data Mining for Monitoring and Improving Academic Performance at University Levels	South Africa	International Journal of Advanced Computer Science and Applications (IJACSA)

[S31]	Ramaano *et al.* (2021)	Different Models Relating Prior Computer Experience with Performance in First Year Computer Science	South Africa	Thesis/Dissertation
[S32]	Ramokolo (2021)	Application of discrete-time survival analysis techniques in modelling student dropout: A case of engineering students at Tshwane University of Technology, South Africa	South Africa	Thesis/Dissertation
[S33]	Saheed *et al.* (2018)	Student performance prediction based on data mining classification techniques	Nigeria	Nigerian Journal of Technology
[S34]	Seota *et al.* (2021)	Modeling E-Behaviour, Personality and Academic Performance with Machine Learning	South Africa	Applied Sciences
[S35]	Soobramoney (2021)	Early prediction of students at risk in a virtual learning environment using ensemble machine learning techniques	South Africa	Thesis/Dissertation
[S36]	Taodzera (2018)	Predicting Student Performance Using Machine Learning Analytics	South Africa	Thesis/Dissertation
[S37]	Tegegne and Alemu (2018)	Educational Data Mining for Students' Academic Performance Analysis in Selected Ethiopian Universities	Ethiopia	Journal of Information and Knowledge Management
[S38]	Van Appel and Durandt (2019)	Investigating possibilities of predictive mathematical models to identify at-risk students in the south african higher education context	South Africa	Perspectives in Education
[S39]	Waidor and Akpojaro (2019)	The Use of Classification Algorithm for Forecasting the Academic Performance of Students of Biological Sciences, University of Africa, Toru-Orua	Nigeria	African Scientist
[S40]	Wanjau and Muketha (2018)	Improving Student Enrollment Prediction Using Ensemble Classifiers	Kenya	International Journal of Computer Applications Technology and Research
[S41]	Yakubu and Abubakar (2021)	Applying machine learning approach to predict students' performance in higher educational institutions	Nigeria	International Journal of Systems & Cybernetics

Table A.3: Quality Assessment of Selected Studies

ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Quality score	Percentage
[S1]	1	1	1	1	1	0,5	1	0,5	1	8	88,89
[S2]	1	1	1	1	1	0,5	1	1	1	8,5	94,44
[S3]	0,5	1	1	1	1	0,5	1	1	1	8	88,89
[S4]	1	1	1	1	1	0,5	1	1	1	8,5	94,44
[S5]	0	0	1	1	1	1	0,5	1	1	6,5	72,22
[S6]	0	0	1	1	1	0,5	1	1	1	6,5	72,22
[S7]	1	1	1	1	1	0,5	1	1	1	8,5	94,44
[S8]	1	1	1	1	0,5	0,5	0,5	0,5	0,5	6,5	72,22
[S9]	1	1	1	1	1	0,5	1	1	1	8,5	94,44
[S10]	0	0	1	0,5	0,5	0,5	1	1	1	5,5	61,11
[S11]	1	1	1	1	1	0,5	1	1	1	8,5	94,44
[S12]	1	1	1	1	1	0,5	0,5	1	0,5	7,5	83,33
[S13]	1	0,5	1	1	1	0,5	1	1	0,5	7,5	83,33
[S14]	1	1	1	1	1	0,5	1	1	0,5	8	88,89
[S15]	0,5	0,5	1	1	1	0,5	1	0,5	0,5	6,5	72,22
[S16]	1	1	1	1	1	0,5	1	1	1	8,5	94,44
[S17]	1	0,5	1	1	1	1	0,5	1	0,5	7,5	83,33
[S18]	0,5	0,5	1	1	1	1	0,5	0	0	5,5	61,11
[S19]	0	0	1	1	1	0,5	0,5	0,5	0,5	5	55,56
[S20]	1	1	1	1	1	0,5	1	1	0,5	8	88,89
[S21]	1	1	1	1	1	0,5	1	1	1	8,5	94,44
[S22]	1	1	1	1		0,5	0,5	1	1	7	77,78
[S23]	0,5	0,5	0,5	1	0,5	0	0,5	1	1	5,5	61,11
[S24]	1	1	0,5	1	1	0,5	1	0,5	1	7,5	83,33
[S25]	1	1	1	1	1	0,5	1	1	1	8,5	94,44
[S26]	1	1	1	1	1	0,5	1	1	1	8,5	94,44
[S27]	1	0,5	1	1	1	0,5	1	1	1	8	88,89
[S28]	1	1	1	1	1	0,5	1	1	1	8,5	94,44
[S29]	0,5	0,5	1	1	1	0,5	1	1	1	7,5	83,33
[S30]	1	1	1	1	1	1	1	1	1	9	100,00
[S31]	1	1	1	1	0,5	0,5	1	1	1	8	88,89
[S32]	1	1	1	1	1	0,5	1	0,5	1	8	88,89
[S33]	0,5	0,5	1	1	1	1	1	1	1	8	88,89
[S34]	1	1	1	1	1	0,5	1	1	1	8,5	94,44
[S35]	1	1	1	1	1	1	1	1	1	9	100,00
[S36]	1	1	1	1	1	1	1	1	1	9	100,00
[S37]	1	1	0,5	1	1	0,5	1	0,5	0	6,5	72,22
[S38]	1	1	1	1	1	0,5	1	1	1	8,5	94,44
[S39]	0	0	1	1	1	0,5	0,5	0,5	0,5	5	55,56
[S40]	0,5	0,5	1	1	1	0,5	1	1	1	7,5	83,33
[S41]	0,5	0,5	1	1	1	0,5	1	1	1	7,5	83,33

Table A.4: Factor Categories with Descriptions

Factor Categories	Individual Factors	Studies	Frequency
Previous Grades and Current Class Performance	• student's grade point average (GPA) at secondary school • placement tests score on entry to university • current GPA for courses taken • raw scores • cumulative grade point average (CGPA) • Group assignment marks • Tutorial marks • Test marks • Bonus marks • Individual assignment marks • First to third-year GPA • Final CGPA • grades obtained in prerequisite module • Online cumulative average quiz mark • Matric results • Admission Point Score (APS) • National Benchmark Tests (NBT) • Outcome • Admission Matriculation Board (JAMB) score • Unified Tertiary Matriculation Examination (UTME) score • Senior School Certificate Examination (SSCE) score	[S1], [S2], [S3], [S4], [S5], [S6], [S7], [S8], [S9], [S10], [S11], [S12], [S13], [S14], [S15], [S16], [S17], [S18], [S19], [S20], [S21], [S22], [S23], [S24], [S25], [S26], [S27], [S28], [S29], [S30], [S31], [S32], [S33], [S34], [S35], [S36], [S37], [S38], [S39], [S40], [S41]	41

	• Class average • Cumulative weighted Average (CWA)		
Demographics	• student's major • current year of study • year of admission • number of modules registered • bursary • Full-time or part-time study • Number of years in degree • Year of graduation • Number of employed parents or guardians • Age • Date of birth • Gender • Race • Marital status • High school quintile ranking • Admission date • Email address • Telephone number • Home address • Name of previous School • Level of study • Country of origin • Home language • Additional language • Geopolitical zone • Nationality • Place of birth • Field of study • Disability indicator	[S1], [S2], [S3], [S4], [S5], [S6], [S7], [S8], [S9], [S10], [S11], [S12], [S13], [S14], [S16], [S17], [S20], [S21], [S22], [S23], [S24], [S25], [S26], [S27], [S28], [S29], [S30], [S32], [S33], [S34], [S35], [S36], [S37], [S38], [S39], [S40], [S41]	37

	• Financial sources • Financial aid indicator • Credits obtained • Years in the system • Persistence (a count of modules attempted) • Number of modules completed • Number of modules passed the first time • Employment status • Student status (attrition or graduated) • Progression outcome type		
Socio-economic data	• Residence • School quintile • Rural/Urban School • Class communication • Proper guidance • Study habit • Student Attitude • Teaching Aids • Family Stress • Parent education • Family size • Family income • Family support structure • Motivation • Academic self-efficacy • Stress and time pressure • Parents survey answers • parent sponsor • Parent school satisfaction • Study hours	[S5], [S7], [S8], [S9], [S11], [S14], [S15], [S17], [S22], [S25], [S27], [S29], [S31], [S34], [S36], [S37], [S41]	24

	• Class attendance • course or program • Event context • Event name • Origin • IP address • Course code • Course name • Course likes		
E-learning Activities	• Discussion • Time of discussion • Viewing announcements • Students virtual learning environment (VLE) • Description • User full name • Components • Time • Affected user • Event name • Event context • IP address • Origin • Visual resources	[S17], [S22], [S23], [S26], [S39]	6
Instructor Attributes	• Full name • Instructor Position • Instructor Gender • Instructor Nationality • Description • Number • Likes • Lecture competency • Lecturer Availability • Lecturer Dedication	[S10], [S13], [S25], [S41]	5

	• Lecturer Attitude		
Prior knowledge	• Prior knowledge • Pre-university awareness • Self-efficacy	[S11], [S14], [S19], [S20], [S22], [S34]	4

Table A.5: Data Collection techniques and Dataset size

ID	Predictors used	Data Collection Technique(s)	Dataset Size
[S1]	(+) School quintile, (+) Age at first year, (-) high school grades, and (+) province.	Students' Information Systems (SIS)	1069
[S2]	(+) Class grade, (+) summative assessment, (+) formative assessment, and (+) socio-demographics.	Students' Information Systems (SIS)	197
[S3]	(-) Year of entry, (+) course of study, (+) first three years GPA, and (+) Final CGPA.	Students' Information Systems (SIS)	1841
[S4]	(+) Preadmission scores, (+) college, (-) geopolitical zone, (+) CGPA, and (+) year of graduation.	Students' Information Systems (SIS)	2413
[S5]	(+) Gender, (+) UTME score, (+) age, (+) SSCE score, (+) 100 level grade.	Students' Information Systems (SIS)	100
[S6]	(+) motivation, (+) self-efficacy, (+) interest level, (+) previous programming knowledge, and (+) final grades.	• Structured Questionnaire/Survey • Students' Information Systems (SIS)	401
[S7]	(+) High school average, (-) end-of-semester examination marks, and (-) CGPA.	Students' Information Systems (SIS)	1425
[S8]	(+) Rural/Urban School, (+) School Quintile, (+) Home language, (+) Class communication, (+) Home Country, (+) APS, (+) Age, (+) year of study, (+) Time management, (+) Home Province, (+) English, (+) Plan Description, (+) Student Absent Days, (+) Financial support, (+) Accounting, (+) Science, (+) Interest in sports, (+) Visited Resources, (+) International, (+) Mathematics, (+) Cognitive Difficulties.	Students' Information Systems (SIS)	50000
[S9]	(+) Discussion, (+) Gender, (+) Nationality, (+) Relation, (+) Semester, (+) Topic, (+) Section ID, (+) Grade ID, (+) raise hands, (+) Visual resources, (+) Announcement views, (+) Parent School satisfaction, (+) Parent Answering Survey, (+) Student Absent Days.	• Students' Information Systems (SIS) • Student online learning logs	500
[S10]	(+) Motivation, (+) Family support structure, (+) specialization, and (+) employment status.	Students' Information Systems (SIS)	1391
[S11]	(+) Age, (+) Sex, (+) Marital status, (+),(+) Secondary school area, (+) Secondary school type, Attended primary school (+) Years before admission, (+) Sports activeness, (+) Weekly study time, (+) Sponsor type, (+) Sponsor income, (+) Sponsor support, (+)	Students' Information Systems (SIS)	3481

	Sponsor qualification, (+) University accommodation, (+) Work and Study, (+) Family size, (+) Course interest, (+) post-UTME score, (+) JAMB score, (+) Average SSCE score, (+) CGPA, and (+) Postgraduate degree.		
[S12]	(+) Viewing Announcements, (+) Visited Resources, (+) Topic, (+) section ID, (+) raised hand in Class, (+) Discussion groups, (+) Place of Birth, (+) Gender, (+) Nationality, (+) Parent School Satisfaction, (+) Parent responsible, (+) Student Absence Days, (+) Educational Stage, (+) Parent Participation, (+) Semester, (+) Parent Answering Survey, and (+) class grade.	Moodle LMS Logs	480
[S13]	(+) Health, (+) Background knowledge, (+) Student Attendance, (+) Lecture time, (+) Family stress, (+) Parent education, (+) Family income, (+) Student Attitude, (+) Lecturer Attitude, (+) Student Fear and Perception, (+) Student Study Habit, (+) Tutorials, (+) Extra Classes, (+) Lecturer Dedication, (+) Lecturer Availability, (+) Teaching aids, (+) Teaching Style, (+) Communication Skills, (+) Class population, (+) Electricity, and (+) Facilities.	Survey	Not specified
[S14]	(+) Final Mark, (+) Assignment Mark, (+) Tutorial Mark, (+) Test Mark, (+) Bonus Mark, (+) NBT Math, (+) Math Prerequisites, and (+) Math High School.	• Students' Information Systems (SIS) • Student online learning logs	1970
[S15]	(+) Average matriculation results, (+) type of high school the student attended, (+) first semester first-year university results, (+) attendance, (+) self-study hours, and (+) lecturers' competency.	Students' Information Systems (SIS)	247
[S16]	(+) Age, (+) Sex, (+) Employment status, (+) Type of school attended, (+) Preparatory completion year, (+) Preparatory attended region, (+) Field of study, (+) Financial sources, (+) Admission scores, (+) Year of admission, and (+) students' status.	Students' Information Systems (SIS)	7361
[S17]	(+) CGPA, (+) course GPA, and (+) raw matric grades.	Students' Information Systems (SIS)	846
[S18]	(+) GPA of students throughout six semesters.	Students' Information Systems (SIS)	100
[S19]	(+) students average mark and the (+) course average mark.	Students' Information Systems (SIS)	501

[S20]	(+) Enrolment, (+) Drop out, (+) deferment, (+) progression rates.	Students' Information Systems (SIS)	2976
[S21]	(+) Lecturer (likes, name, number), (+) subject (likes, name, code), (+) grade (average).	Students' Information Systems (SIS)	Not specified
[S22]	(+) Gender, (+) age, (+) disability, (+) home language, (+) financial aid, (+) accommodation, previous year's grades, (+) modules, (+) qualification, and (+) CGPA.	Students' Information Systems (SIS)	4419
[S23]	(+) Study hours per week, (+) bursary, (+) class attendance, (+) full-time or part-time classes, (+) number of modules registered, (+) language, (+) test marks, (+) group assignment marks, and (+) number of employed parents or guardians.	Survey	Not specified
[S24]	(+) Year of study, (+) student subjects, (+) student marks, (+) Progression Outcome type.	Students' Information Systems (SIS)	8557
[S25]	(-) NBT, (+) School Quintile, (+) Number of years in degree, (+) Gender, (+) Age at first year, (+) From Rural / Urban, (+) Home Country, (+) Home Province, (+) Year Started, (+) Race, (+) Mathematics Major, (+) Plan Description, (+) Home Language, (-) Matric Maths grades.	Students' Information Systems (SIS)	2000
[S26]	Students' (+) loan allocation data, (+) living locations, (+) coursework, (+) remarks, (+) final examination results, (+) mathematics grades, (+) gender, (+) age, and (+) placement scores.	Students' Information Systems (SIS)	3347
[S27]	(+) Prior computer skills, (+) High school GPA, (+) placement test scores, (+) course GPA, (+) year of admission, (+) year of study, and (+) majors.	Students' Information Systems (SIS)	6000
[S28]	(+) Home Country, (+) Career Choice, (+) High school marks for a host of courses, (+) Additional Language, (+) Year Started, (+) School Quintile, (+) Home Province, and (+) National Benchmark Tests.	Students' Information Systems (SIS)	768
[S29]	(+) Admission point score, (+) home province, (+) home language, (+) home country, (+) stress and time pressure, (+) age at first year, (+) dwelling value, (+) academic self-efficacy, (+) work status, (+) financial support, (+) parents' education, (+) family income, (+) parents' occupation, (+) college activity participation, (+) sense of loneliness, (+) class communication, (+) organization and attention to study, (+) interest in sports, and (+) statistics major.	Students' Information Systems (SIS)	2000
[S30]	Course, staff, and student data in the format: (+) username, (+) time, (+) description, (+)	Moodle LMS Logs	712

	IP address, (+) origin, (+) event name, (+) event context, (+) affected user, and (+) components.		
[S31]	(+) Computer studies high school grades, (+) Pure mathematics high school grades, (+) English first language high school grades, and (+) prior computer experience.	Students' Information Systems (SIS)	428
[S32]	(+) Gender, (+) Race, (+) Residence, (+) Home language, (+) Second language, (+) APS score, (+) Mathematics score, and (+) CGPA.	Students' Information Systems (SIS)	502
[S33]	(+) Attendance of the student in a lecture room, (+) end semester exam mark, (+) friends and family support, (+) economic status, (+) living location, and (+) parents' qualification.	Students' Information Systems (SIS)	234
[S34]	(+) Quintile, (+) province, (+) gender, (+) financial assistance, (+) township school, (+) grade, (+) outcome, (+) discussion, (+) message, and (+) time.	Moodle LMS Logs	4748
[S35]	(+) Student Info, (+) courses, (+) student Registration, (+) assessments, (+) student Assessments, and (+) student VLE.	• Students' Information Systems (SIS) • Student online learning logs	32593
[S36]	(+) Age at first year, (+) socio-economic information, (+) demographic information, (+) Matric results, and (+) first-year results.	Students' Information Systems (SIS)	2203
[S37]	(+) Field Choice Interest, (+) Ethiopian University Entrance Examination result, (+) High school GPA, (+) First-Semester First-Year grades, and (+) CGPA.	Students' Information Systems (SIS)	5729
[S38]	(+) Two formal semester test marks, (+) online cumulative average quiz mark, (+) final period mark, (+) grade obtained in a prerequisite module, (+) repeating the module, (+) gender, and (+) high school quintile ranking.	Students' Information Systems (SIS)	395
[S39]	(+) Gender, (+) admission grades, (+) first-year GPA, (+) students' registration numbers, (+) department, and (+) academic standing.	Students' Information Systems (SIS)	84
[S40]	(+) Career Flexibility, (+) High School Final Grade, (+) Math Grade, (+) Pre-University awareness, (+) Teacher Inspiration, (+) Financial Aid, (+) Extracurricular, (+) Parent Career, (+) Societal Expectation, (+) Career Earning, (+) Self-Efficacy, (+) Gender, (+) Age, and (+) Family Income.	Student Questionnaire/Survey	209
[S41]	(+) Gender, (+) marital status, (+) country of origin, (+) date of birth, (+) level of study, (+) telephone number, (+) home address, (+)	Students' Information Systems (SIS)	978

	email address, (+) name of previous school, (+) admission date, (+) JAMB score, and (+) CGPA.		

Table A.6: Data extraction of selected studies

ID	Findings	Data Mining Approach(es)	Data Mining Algorithms	Tools used
[S1]	• The school quintile, age at first year, province, NBT marks, and Plan description greatly influenced students' academic performance predictions. • However, the high school grades were less influential.	Classification	• Decision Trees • Naïve Bayes • Logistic Regression • Multilayer Perceptron • Random Forests • Sequential Minimal Optimization	Not specified
[S2]	The results show that timely completion of discrete mathematics and programming courses significantly predicts students' final grades.	Classification	Decision Tree	WEKA (Waikato Environment for Knowledge Analysis)
[S3]	• The most influential factor is students' third-year GPA. • While the year of entry is the less significant in students' academic performance prediction.	Classification	• Naïve Bayes • Decision Trees • Probabilistic Neural Network • Logistic Regression • Random Forests • Tree Ensemble	KNIME (Konstanz Information Miner)
[S4]	The findings revealed that students' ethnicity was not significant in predicting their performance.	Classification	• Naïve Bayes • Artificial Neural Network • Random Forests • Classification Trees	Orange Software
[S5]	The significant factors influencing students' academic performance prediction were 100 level grade, UTME score, Age at admission, gender, and Subject grades in core subjects.	Classification	• ID3 Decision Tree • C4.5 Decision Tree	WEKA tool
[S6]	The findings showed that the key factors influencing students' academic performance were motivation, interest level, previous programming knowledge, and self-efficacy.	Classification	• Naive Bayes • Multilayer perceptron • J48 Decision Tree	WEKA tool

[S7]	The study found a weak correlation between students' best six GPAs and CGPA.	Classification	• Artificial Neural Network • Multilayer perceptron • J48 decision tree • k- Nearest Neighbour	Waikato Environment for Knowledge Analysis (WEKA)
[S8]	Background Factors, Pre & Intra-College Scores, Individual Features, and Psycho-Social Factors were determining factors to predict students at-risk in each year of study.	Classification	• Random Forests • Logistic Regression • K-Nearest Neighbour • Naïve Bayes • Multi-Layer Perceptron • Decision Trees	Not specified
[S9]	Behavioural features were the most crucial factors in students' academic performance prediction.	Classification	• Logistic Regression • Naïve Bayes • Multilayer Perceptron Neural Network • Support Vector Machines • Decision Tree • Gradient • Boosting Tree • Linear Discriminant • Analysis • Random Forests	Python anaconda on Jupyter notebook
[S10]	Family support structure, change in program offering, and employment status were the most crucial factors in students' academic performance prediction.	Classification	Multinomial Naïve Bayes	Not specified
[S11]	Age, Sex, Marital status, attended primary school, Secondary school area, Secondary school type, Years before admission, Sports activeness, Weekly study time, Sponsor type, Sponsor income, Sponsor support, Sponsor qualification, University accommodation, Work and Study, Family size, Course interest, post-UTME score,	Classification	• J48 Decision Tree • Multilayer Perceptron • Logistic Regression • Sequential Minimal Optimization • Naïve Bayes	WEKA tool

	Jamb score, Average score, CGPA, and Postgraduate degree were the most determining factors in predicting students' academic performance.			
[S12]	Low-level absentee students with parents who do not participate enough in their studies are more likely to perform poorer than those with parents who participate.	Clustering	MakeDensityBasedClusterer	WEKA (Waikato Environment for Knowledge Analysis)
[S13]	The significant factors contributing to students' academic performance in programming courses were lecturer attitudes, student attitudes, fearful perception of programming courses, student attendance, student health, university facilities, and erratic power supply.	Classification	• M5P Decision Tree • Linear Regression	WEKA (Waikato Environment for Knowledge Analysis)
[S14]	Final Marks, Assignment Marks, Tutorial Marks, Test Marks, Bonus Marks, NBT Math Marks, Math Prerequisites, and Math High School were significant in predicting students' academic performance.	Classification	• Decision Trees • Logistic Regression	Not specified
[S15]	The first-semester university results, average matriculation, lecturer competency, self-study hours, and attendance were significant factors for predicting students' academic performance before their final exams.	Classification	• Bayesian Networks • Support Vector Machines • Decision Trees	WEKA tool
[S16]	The most significant factors influencing student retention and attrition were the financial sources, division, types of preparatory attended school, department, the background of the study, and preparatory completion year.	Classification	• Rule induction (PART and JRIP) • J48 Decision Tree • Naïve Bayes	WEKA tool
[S17]	Of the 21 first-year courses offered, only seven, such as mathematics, physics, and	Clustering	Association Rule Mining	R Studio

	chemistry, frequently occur as failed courses, significantly influencing students' academic performance.			
[S18]	The grade point average factor significantly predicts students' academic performance.	Classification	• Multilayer Perceptron • Generalized Regression Neural Network	MATLAB
[S19]	The internal assessments and course average does well in predicting students' performance enrolled in the BSCSIT programme.	Classification	• Matrix Factorization • Singular Value Decomposition	Java
[S20]	• The number of deferments of students highly determined the progression rate of students. • More students were deferring than those dropping out of their courses altogether.	Classification	Artificial Neural Network	• Excel • R software
[S21]	The study found that the lecturer's name, subject code, class average, and the percentage of students that like the lecturer's teaching style can potentially determine students' academic performance.	Classification	• Fuzzy Logic • Artificial Neural Networks • Support Vector Machines • k-Nearest Neighbour • Linear and multiple regression • Decision Trees • Naïve Bayes • Multilayer Perceptron	WEKA (Waikato Environment for Knowledge Analysis)
[S22]	The factors significantly influencing students' academic performance were gender, age, financial aid, accommodation, home language, disability, previous year activity, qualification, modules, and CGPA.	Classification	• Random Forests • Naïve Bayes • Decision Trees • K-Nearest Neighbour • Support Vector Machines • Logistic Regression	Not specified
[S23]	Group assignments and financial aid availability had a positive correlation with students' academic performance.	Classification	• Random Forests • Neural Network-Multilayer Perceptron	R programming language

			• Support Vector Machines • Linear Regression • Naïve Bayes • eXtreme Gradient Boosting	
[S24]	The first-year grade point average factor significantly predicts students' academic performance.	Classification	• Support Vector Machines • Naïve Bayes • Decision Trees	WEKA tool
[S25]	• The findings showed that biographical and individual attributes influence student attrition. • Pre-college factors were not significant in influencing student attrition.	Classification	• Support Vector Machines • Linear Logistic Regression • Decision Trees • Bayesian Network • eXtreme Gradient Boosting Trees • Random Forest	• MATLAB • R studio
[S26]	Students' loan allocation data, living locations, coursework, remarks, final examination results, gender, age, mathematics grades, and placement scores were significant in predicting students' academic performance.	Classification	• K-Nearest Neighbour • Random Forests • Decision Tree • Support Vector Classification • Multilayer Perceptron	• Microsoft Excel • Jupyter Notebook
[S27]	The significant factors influencing students' academic performance were the year of admission year of study, Mathematics placement score, and Introduction to Computers grade.	Classification	• Linear and Multiple Regression • Artificial Neural Networks	• Microsoft Excel Spreadsheet • R Studio • SPSS
[S28]	The results proved that the background, individual, and pre-university features were the most influential predictors of students' academic performance.	Classification	• Random Forests • Multi-Layer Perceptron • Decision Trees • Logistic Regression • Naive Bayes • K-Star	Not specified
[S29]	The results showed the seven top-ranked features: age at first year, the year started, Matric Mathematics	Classification	• Naïve Bayes • C4.5 Decision Trees • Random Forests	Not specified

	major, plan description, plan code, school quintile, and streamline.		• Multinomial Logistic Regression • Sequential Minimal Optimization • Logistic Model Trees	
[S30]	A strong correlation exists between using LMS resources and students' academic performance at a 0.01 significance level.	Classification, Clustering	• J48 Decision Tree • ZeroR	WEKA (Waikato Environment for Knowledge Analysis)
[S31]	Students with previous computer experience had better performances than those that did not have experience.	Classification	• Decision Tree • Naïve Bayes • Logistic Regression • Linear Regression	Not specified
[S32]	• The outcomes of the type of residence differed with time. First-year students in private residences were more likely to drop out than on-campus students. • In contrast, third-year students living in private-based accommodations were less likely to drop out than on-campus students.	Classification	Discrete-Time Survival Analysis	STATA
[S33]	The significant factors were Age, family factors, Parents' Satisfaction Level, and Course Applied For.	Classification	• ID3 Decision Tree • C4.5 Decision Tree • Classification and Regression tree	WEKA tool
[S34]	Economic and cultural capital combined with the gender factor provided better performances in students' academic performance prediction.	Classification	• Decision Tree • K-Nearest Neighbour • Ordinary Least Squares Linear Regression	Not specified
[S35]	• The findings show that identifying at-risk students can be done using demographic and VLE interaction data	Classification	• Logistic Regression • k-Nearest Neighbour • Naïve Bayes	Python

	instead of just assessment results.		• Support Vector Machines • Multilayer perceptron • Random Forests • Decision Tree • Gradient boosting models such as XGBoost, LightGBM, and AdaBoost	
[S36]	Ethnicity, age, school province, physics grades, and mathematics grades were the main factors of greater significance in predicting students' academic performance.	Classification	• Support Vector Machines • Decision Tree • Naïve Bayes • Artificial Neural Networks • Logistic regression	Python
[S37]	The most significant predictors used included preparatory school GPA, entrance Examination results, first-year first-semester results, and field choice interest in predicting students' academic performance.	Classification	Decision tree	WEKA tool
[S38]	The significant factors in students' academic performance prediction were prerequisite course marks, semester test one marks, and average quiz marks.	Classification	• Multiple Regression • Logistic Regression • Decision Trees • Classification Tree • Regression Tree	Not specified
[S39]	The significant factors included the level of matric achieved, gender, programme, first semester GPA, second semester GPA, CGPA, and current standing in predicting students' academic performance.	Classification	Decision Tree C4.5 (J48)	WEKA tool
[S40]	The significant factors in predicting students' academic performance were the high school final exam score, beliefs in competency for succeeding in a STEM-related career,	Classification	• Decision Tree J48 • Naïve Bayes • CART • Bagging	WEKA tool

	inspiration from the high school teachers, and expected career flexibility.			
[S41]	• Students with high Joint Admissions and Matriculation Board (JAMB) grades are more likely to attain a high CGPA. • Also, third- and fourth-year students were more likely to attain high CGPA.	Classification	Logistic regression	Scikit-learn (Sklearn)

Table A.7: Effectiveness of EDM methods

ID	Predictor used	Aim of prediction	Level of prediction	Best Performing EDM Algorithm	Effectiveness
[S1]	• (-) Previous Grades and Current Class Performance • (+) Demographics • (+) Socio-economic data	Pass-Fail	Degree Level	J48 Decision Tree	Accuracy: 80,4%
[S2]	• (+) Previous Grades and Current Class Performance • (+) Demographics	Final Grade	Degree Level	J48 decision tree	Accuracy: 50%
[S3]	• (-) Demographics • (+) Previous Grades and Current Class Performance	Final Grade	Year Level	Linear and pure quadratic regression models	Accuracy: 85,89%
[S4]	• (-) Demographics • (+) Previous Grades and Current Class Performance	Final Grade	Degree Level	Decision Trees	Accuracy: 73,6%
[S5]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (-) Socio-economic data	Final Grade	Degree Level	ID3 decision trees	Accuracy: 61%
[S6]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Instructor Attributes	Pass-Fail	Course Level	J48 Decision Tree	Accuracy: 71,57%
[S7]	• (+) Demographics • (-) Previous Grades and Current Class Performance	Final Grade	Year Level	J48 Decision Trees	Accuracy: 100%
[S8]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Socio-economic data	Pass-Fail	Degree Level	Decision Trees	Accuracy: 97,3%
[S9]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) E-learning Activities	Pass-Fail	Degree Level	Random Forest	Accuracy: 91%

[S10]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Socio-economic data	At-Risk (Pass-Fail-dropout)	Course Level	Multinomial Naïve Bayes	Accuracy: 66%
[S11]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Socio-economic data	Final Grade	Degree Level	Multilayer perception network model	Accuracy: 98%
[S12]	• (+) Previous Grades and Current Class Performance • (+) Demographics • (+) Socio-economic data • (+) E-learning Activities	Pass-Fail	Degree Level	MakeDensityBasedClusterer	Not Specified
[S13]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Prior knowledge • (+) Socio-economic data	Final Grade	Course Level	Regression	Accuracy: 58,5%
[S14]	• (+) Demographics • (+) E-Learning Activities • (+) Previous Grades and Current Class Performance	Pass-Fail	Course Level	Decision Trees	Accuracy: 86%
[S15]	• (+) Previous Grades and Current Class Performance • (+) Instructor attributes	Pass-Fail	Year Level	Support Vector Machine	Accuracy: 72,87%
[S16]	• (+) Demographics • (+) Previous Grades and Current Class Performance	At-Risk (Pass-Fail-dropout)	Degree Level	J48 Decision Tree	Accuracy: 91,40%
[S17]	• (+) Demographics • (+) Previous Grades and Current Class Performance	Pass-Fail	Year Level	Hierarchical Cluster Analysis, Association Rule Mining	coefficient: 92%
[S18]	(+) Previous Grades and Current Class Performance	Final Grade	Degree Level	Generalized Regression Neural Network	Accuracy: 95%
[S19]	(+) Previous Grades and Current Class Performance	Final Grade	Degree Level	Matrix Factorization	Root Mean Square Error: 9,7

[S20]	• (+) Demographics • (+) Previous Grades and Current Class Performance	Pass-Fail	Degree Level	Artificial Neural Network	Accuracy: 78,5%
[S21]	• (+) Instructor Attributes • (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Socio-economic data	Final Grade	Degree Level	Regression	Accuracy: 93%
[S22]	• (+) Previous Grades and Current Class Performance • (+) Demographics	At-Risk (Pass-Fail-dropout)	Degree Level	Random Forest	Accuracy: 94,14%
[S23]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Socio-economic data	Pass-Fail	Degree Level	Neural Networks	AUC (area under the curve): 86%
[S24]	• (+) Demographics • (+) Previous Grades and Current Class Performance	Pass-Fail	Degree Level	Support Vector Machine	Accuracy: 87%
[S25]	• (+) Demographics • (-) Previous Grades and Current Class Performance • (+) Socio-economic data	At-Risk (Pass-Fail-dropout)	Year Level	Random Forest	Accuracy: 85%
[S26]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Socio-economic data • (+) Instructor Attributes	Pass-Fail	Degree Level	Random Forest	Accuracy: 99%
[S27]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Socio-economic data	Final Grade	Course Level	Artificial Neural Networks	Accuracy: 83%
[S28]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Socio-economic data	At-Risk (Pass-Fail-dropout)	Degree Level	Random Forest	Accuracy: 95%

[S29]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Socio-economic data	Pass-Fail	Degree Level	Random Forests	Accuracy: 95,45%
[S30]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Socio-economic data • (+) Instructor Attributes • (+) E-learning Activities	Final Grade	Course Level	J48 decision tree	Accuracy: 98%
[S31]	• (+) Previous Grades and Current Class Performance • (+) Prior knowledge	Pass-Fail	Year Level	Decision Trees	Accuracy: 73,44%
[S32]	• (+) Previous Grades and Current Class Performance • (+) Demographics • (+) Socio-economic data	Pass-Fail	Degree Level	Discrete-time analysis	significance: < 0,05
[S33]	• (+) Previous Grades and Current Class Performance • (+) Demographics • (+) Socio-economic data	Pass-Fail	Course Level	C4.5 Decision Trees, and CART	Accuracy: 98,3%
[S34]	• (+) Previous Grades and Current Class Performance • (+) Demographics • (+) Socio-economic data • (+) E-learning Activities	At-Risk (Pass-Fail-dropout)	Year Level	Decision Trees	Accuracy: 90%
[S35]	• (+) E-Learning Activities • (+) Previous Grades and Current Class Performance • (+) Demographics	At-Risk (Pass-Fail-dropout)	Degree Level	AdaBoost Classifier, LGBM Classifier, Random Forest Classifier, and XGB Classifier	Accuracy: 92%
[S36]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Socio-economic data	Pass-Fail	Year Level	Decision Trees	Accuracy: 65,8%

[S37]	• (+) Previous Grades and Current Class Performance • (+) Demographics	Final Grade	Year Level	C4.5 Decision tree	Accuracy: 81,4%
[S38]	• (+) Demographics • (+) Previous Grades and Current Class Performance	At-Risk (Pass-Fail-dropout)	Course Level	Multiple Regression	Accuracy: 79%
[S39]	• (+) Demographics • (+) Previous Grades and Current Class Performance	Final Grade	Course Level	J48 Decision Tree	Accuracy: 95,24%
[S40]	• (+) Demographics • (+) Previous Grades and Current Class Performance • (+) Socio-economic data	Pass-Fail	Course Level	J48 Decision Tree	Accuracy: 84%
[S41]	• (+) Demographics • (+) Previous Grades and Current Class Performance	Final Grade	Degree Level	Logistic Regression	Accuracy: 84,7%