Predicting at-Risk Students in a Higher Educational Institution in Ghana for Early Intervention Using Machine Learning

A thesis submitted in fulfilment of the requirement for the degree of Doctor of Philosophy (PhD) in Information Technology (IT) at Durban University of Technology, South Africa.

Student Name: Fati Tahiru

DATE SUBMITTED

Supervisor: Dr Steven Parbanath (PhD) Date: July 1, 2023

ABSTRACT

Learning analytics (LA) uses data and evidence to suggest a better learning approach that suits a particular student. This data and evidence are gathered from students' online engagement with systems such as Blackboard, Moodle, Sakai, eLibrary platforms, and other e-learning platforms. LA continues to gain much attention as digitization of the learning environment is advancing. It allows educators to analyze and interpret data correctly, setting in motion strategies that offer points of leverage and performance for and among students**.** The use of predictive systems and Early Warning Systems (EWS) in education addressed the issue of student dropouts and suggested interventions for improving students' performance. High dropout rates in education continue to be a global challenge; however, EWS provide a solution to curb the menace in education in various developed nations, such as the United States, Australia, and the United Kingdom. Developing countries face similar problems of dropouts in the educational sector, but not much research has been undertaken in LA to address the intervention needed to leverage the situation. Some studies have designed models predicting student failure and success, student attrition, student performance and final grades. Most of these studies have focused on only virtual learning environments (VLE) datasets. Nonetheless, this study uses student "activity logs", "student courses", "demographics", and "student assessments" to design a predictive model to identify at-risk students (ARS) from not graduating. The purpose of this study is to use LA and Machine Learning (ML) to analyse the characteristics and behaviours of students in order to identify those who may need support to improve their academic performance. The study adopted the systematic literature review (SLR) approach to determine which emerging ML tools/techniques have been applied successfully in designing predictive systems in education. The SLR enabled the study to identify ML methods and the features that have been used in the domain of predictive systems in education. The study used an integrated 5-step LA process and ML workflow to predict which students are likely to dropout. Using the OULAD dataset, the findings indicated that non-graduated students had habits of not revising the learning materials early before the final exams. Although it was noted that both graduated and non-graduated students access the learning materials simultaneously, variations were recorded in the habits of assignment submission and revision patterns. Graduated students recorded higher clicks for accessing VLE activities than non-graduated students, which signifies that the graduated students

interacted more with course activities than non-graduated students. The study also compared different ML algorithms and determined the method that achieved the best predictive accuracy that could be adapted in higher educational institutions. The evaluation of the models concluded that the ensemble machine-learning methods outperformed the traditional methods. The Random Forest ensemble learning algorithms outperformed the GB, Catboost, KNN, LG and NB on the accuracy, precision, recall and f-1 score. The study identified important features such as “date-of-assignment-submission”, “sum_clicks-of-activities”, “score on the assessment”,”date-of-registration”, “date-of-assignment-submission”, “studied-credits”, and “date-the-student-unregistered” for predicting students dropout in higher educational institution (HEI). The model was trained with the important features to predict ARS and achieved an accuracy of 92% in less time than using all the features. The research indicated that implementing LA and ML techniques can effectively identify students at risk of withdrawing from higher education. In view of this, the study concluded that targeted interventions can be developed to mitigate the risk of students dropping out of school through improved learning outcomes.

DECLARATION

Name: FATI TAHIRU

Student Number 22176488

PhD. (Information Technology)

I hereby declare that this research project is the result of my own work, except for quotations and summaries that have been duly acknowledged.

Signature …… Date: July 1, 2023

(Author)

Approval for examination

Signature Date: 04/12/2023

Dr Steven Parbanath (Supervisor)

Department of Information Technology,

Durban University of Technology,

Durban, South Africa

DEDICATION

This work is dedicated to my parents, Amidu Tahiru and Rose Birago, and my brothers' Kevin Amidu Tahiru, Basiru Tahiru and Abdul Razak Tahiru, for their immense support throughout my education. To all family and friends, I appreciate your support and encouragement.

LIST OF ABBREVIATION

LA	Learning Analytics
ML	Machine Learning
AI	Artificial Intelligence
EWS	Early Warning System
VLE	Virtual Learning Environment
ARS	At-risk student
SLR	Systematic Literature Review
RF	Radom Forest
GB	Gradient Boosting
CB	CatBoost
KNN	K-nearest Neighbour
LG	Logistic Regression
NB	Naïve Bayes
SVM	Support Vector Machine
ANN	Artificial Neural Network
BN	Bayesian Network
DT	Decision Tree
LMS	Learning Management System

MOOC	Massive Open Online Courses
SIS	Student Information System
SOLAR	Society of Learning Analytics Research
EDM	Educational Data Mining
HEI	Higher Educational Institution
OULAD	Open University Learning Analytics dataset

ACKNOWLEDGEMENTS

I would like to extend my heartfelt gratitude to God and everyone who played a part in this research. I would like to express my special thanks to the following individuals:

- Dr Steven Parbanath, DUT, South Africa, my supervisor, for his professional guidance and contribution to my thesis.
- Prof. Valentine Cardenoso-Payo, University of Valladolid (Uva), Spain, for his mentorship and technical contributions to my thesis during my Erasmus mobility in Uva, Spain.
- Dr Samuel Agbesi, IT University, Denmark, for his professional advice and support in conducting this study.
- To Dr Cyril Latzoo and Christopher J.K. Korang for proofreading and editing the thesis.

I want to thank my family and friends for their tremendous support and encouragement during this academic journey.

Fati Tahiru

Durban

RESEARCH PUBLICATIONS EMANATING FROM THIS THESIS

- Tahiru, F., Parbanath, S. and Agbesi, S (2023). Machine Learning-based predictive systems in higher education: A bibliometric analysis. Journal of Scientometric Research (JSCIRES).

- Tahiru, F. and Parbanath, S. (2023). Using an exploratory analytical approach to distinguish the habits of graduating and non-graduating students in a virtual learning environment. In Pudaruth, S., ed. International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD). Virtual conference, South Africa and Mauritius. 4-5 August 2023.

- Tahiru, F. and Agbesi, S. (2021). The Future of Artificial Intelligence in Education. In Digital Technology Advancements in Knowledge Management (pp. 187-194). IGI Global.

Contents

LIST OF FIGURES

LIST OF TABLES

1 Chapter One– Background and Context

This section presents how data available through the emergence of Massive Open learning platforms (MOOC) and using the Learning Management system (LMS) in education contributed to study success. It further defined LA and how its implementations in the educational sector have contributed to study success. The section also highlights the need for ensemble learning techniques for developing predictive models. The section further discussed the issue of high dropout rates among tertiary students and the necessity of establishing a predictive model to identify students who may be at risk of withdrawing from higher educational institutions in Ghana. Additionally, the section presents the research objectives, questions, rationale, privacy and ethics consideration of the study.

1.1 Introduction

The increasing use of Massive Open Online Courses (MOOC), Learning Management systems (LMS) and Student Information Systems (SIS) has led to the generation of educational data for academic and research purposes (Mothukuri et al. 2017). Dyckhoff (2014) remarked that the massive availability of educational data and increasing computer processing power have contributed to a sizeable increase in interest in Learning Analytics (LA). According to the Society of Learning Analytics Research (SOLAR), LA is defined as the "measurement, collection, analysing and reporting of data about learners and their contexts, for understanding and optimizing learning and the environment in which it occurs" (Romero & Ventura 2020:2).

LA uses data and evidence to suggest a better learning approach that suits a particular student. This data and evidence are gathered from students' online engagement with systems such as Blackboard, Moodle, Sakai, eLibrary platforms and other e-learning platforms (Kurilovas 2019; Siemens & Latour 2013). Research in LA continues to gain much attention as digitization of the learning environment is advancing (Flanagan & Flanagan 2018). LA allows educators to analyze and interpret data correctly, setting in motion strategies that offer points of leverage and performance for and among students (see Lewis *et al.*, 2010). Despite the importance of LA and

how it can contribute to positive shifts in learning among at-risk student (ARS) in higher educational institutions (HEI), there is little exploration of its techniques and algorithms in previous research (See Kika, 2018). The deployment of LA in developed nations like the United States, the United Kingdom, and Australia has not yet achieved a substantial level of implementation. According to a 2017 report by the Learning Analytics Forum, the utilisation of LA within higher education institutions in the UK has shown improvement in recent years. However, the adoption rate still falls below the 50% mark (Newland & Trueman 2017). In the case of economically developing nations of Africa, Asia, and Latin America LA is its nascent stage. However, Teachers and researchers in contemporary systems of education need to track the performance and progress of students (Costa, Souza, Salvador & Amorim 2019).

A fundamental responsibility of educators is to atttain study success, which is understood as the attainment of "satisfaction and improvement in the learning process" (Ifenthaler and Yau, 2020, p.1963). In order to achieve success in the pursuit of education, it is imperative to complete individual learning tasks or activities diligently. In order to optimize one's satisfaction and progress in learning, it is essential to consider personal characteristics such as motivation, age, gender, and prior academic achievements. Such is the reason Ifenthaler and Yau (2020) argued that students could maximise study success in higher education by improving and motivating students in their learning activities as well as improving the learning process to incorporate adaptive learning pathways. This contention rests at the heart of the "Organisation for Economic Cooperation and Development's" (OECD) call for paradigmatic shifts in educational policy formulation and implementation among member nations (OECD 2019). OECD connected educational policy formulation and implementation within the framework of success that engages the necessity of attaining satisfaction and improvement in the learning process among ARS in higher educational institutions. This work highlights an important issue, thus the need to ensure that ARS in higher education are able to improve and find satisfaction in their learning process. In the pursuit of enhancing the educational experience and promoting student contentment, scholars are directing their attention towards LA and ML algorithms. These tools are employed to forecast and backstop students who may be at risk in higher education.

Using ML algorithms in educational activities has received much attention in scholarly works (Mothukuri et al. 2017; Er 2012; Zafra & Ventura 2009). Many intervention systems in education have been designed and deployed with the help of ML models (Kokoç & Altun 2021; Gardner & Brooks 2018; Bainbridge et al. 2015; Jayaprakash, Moody, Lauría, Regan & Baron 2014; Marques, Hobbs & Graf 2014). For example ML algorithms such as "Bayesian Network", "Decision Tree", "Artificial Neural Networks" (ANN) and "Support Vector Machine" (SVM) algorithms have been used to design EWS that brought students at risk of dropping out back on track in the educational sector (Tahiru and Agbesi, 2021).

Additionally, numerous institutions have created their own predictive or EWS using LA and ML to identify students who may be at-risk of dropping out (Liu, Atif, Froissard & Richards 2019). For example, Autoscholar is a full web suite system that provides services to enable students to achieve success in their studies at the Durban University of Technology (DUT) in South Africa. The system uses coursework metadata to facilitate the concept. It comprises other cohorts such as student counselling services, predicting ARS, and career mapping, and it also serves as a research portal. The system predicts student performance based on student assessment, course work and attendance ('Technical Specification Document', no date).

Besides the institutional-acquired proprietary systems, the use of LMS has proven to support LA in most institutions. Moodle and Blackboard Learn are the two leading LMS adopted by institutions. Moodle is an LMS that operates on an open-source platform and is equipped with a diverse range of LA plugins. Moodle, as an LMS, gathers substantial student data, encompassing their engagement with course materials, evaluations, and communication. This data is leveraged to anticipate student performance and develop features that foster deeper engagement. Other plugins available in Moodle include GISMO, MACLog, Analytics Graphs, SmartKlass, and Engagement Analytics plugins. The system is basically used for "delivering course content", "course progression plans", "grading", "creating activities", "collecting course feedback", and "communication". However, quizzes, assignment feedback and workshop modules are the few most used and essential features of Moodle (Deepak 2017). Notwithstanding the benefits of Moodle systems, research posit that data are difficult to query and even more challenging to act upon from within Moodle (Liu et al. 2019).

Blackboard LMS, used as a repository for course materials and course information, also extends as a tool for communication via emails, announcements, discussion boards, podcasts and blackboard analytics. The LMS has gained popularity in teaching and learning in higher education (Matarirano, Jere, Sibanda & Panicker 2020). Blackboard is notably used for administrative purposes as opposed to pedagogical purposes. However, the blackboard implementation is considered expensive due to hardware, licence subscription and maintenance fees, and additional training costs (Alokluk 2018).

The predictive systems described above make use of academic data stored in the LMS without considering student-specific characteristics or features such as age, socio-economic factors, and distance to school, to mention a few. Also, these systems have limitations such as difficulty in acting upon data in Moodles, use of only academic data to make predictions in AutoScholar and high cost of license subscription in Blackboard. Moreover, ML models are designed and selected to match the educational priorities of an institution. In this regard, this study sought to determine ARS in a higher educational institution in Ghana using ML approaches. This study considered additional student context features such as age, gender, and other behavioural features such as students' learning habits in developing the model for predicting ARS.

There is little explicit exploration of LA and ML potential for study success among higher education students in Ghana. The current study aims to dawn upon the potential of learning analytics and ML to identify students at risk of dropping out of higher education. The results intend to offer pragmatic strategies such as improved information activities, study assistance, and mentorship programs that set-in motion leverage change for ARS in higher education environments. This research will develop and compare different ML algorithms using traditional ML and ensemble earning methods. Each model will be optimised to fit the ARS data and evaluated based on precision, accuracy and sensitivity to determine the best ML algorithm for the predictive model. This study will also distinguish the characteristics of a successful student from non-performing students. Furthermore, the study will contribute to the use of Artificial Intelligence (AI) in higher education as well as assist educational stakeholders to better utilize learning analytics to inform study success in Ghana.

In this regard, this study sought to determine ARS at the selected University in Ghana going beyond the academic features such as students' course content, grade and attendance recorded in the LMS. Additionally, student context features such as age, gender, region, and online logs were utilized in developing the model for predicting ARS. Also, this study sought to improve an EWS in higher education by developing and comparing the performance of different ensemble learning to traditional ML predictive models and evaluating the models for accuracy, precision, recall and F1-score.

1.2 The need for ensemble methods in the design of a Predictive System

A single algorithm can be used in modelling a predictive system or several algorithms may be combined to produce a better outcome. Ensemble methods train multiple learners and combine them for use in order to achieve an optimal predictive system (Stapel, Zheng & Pinkwart 2016; Opitz & Maclin 1999). Study success has been predicted using individual techniques such as decision trees, random forest, neural networks and many more for training the dataset (Bainbridge et al. 2015; Arnold et al. 2012). Different authors have alluded to one algorithm outperforming the other based on different criteria of evaluation and the strength of the algorithm (Bravo-Agapito, Romero & Pamplona 2021; Lakkaraju et al. 2015; Raju 2012). Individual learning methods perform better in some configurations and worse in others; single procedures are inconsistent across datasets (Marapelli 2020). A Study by Dawson *et al.*, (2017) used single methods in the design of predictive models. The authors experienced high variability in data when a more advanced algorithm was applied to the data (Dawson et al. 2017). Studies have revealed that when compared to individual models, ensemble models perform better in prediction and have a higher accuracy score (Marapelli 2020; Kim, Heider & Meystre 2018). The following are the most common Ensembled Techniques discussed in the literature:

Voting

When dealing with classification and regression problems, the voting ensemble algorithm is a useful strategy to consider. This involves generating multiple sub-models, typically two or more, in order to attain improved performance. The sub-models use a majority vote to make predictions. Prediction is made by taking the mean or the mode of the aggregated predictions,

and the voting algorithms vote for the best outcome (Marapelli 2020; Kabari & Onwuka 2019; Yu, Lee, Lara & Gan 2018).

Bagging

Bagging or Boostrap Aggregation can also be used for both classification and regression problems. The technique works by taking the mean of random samples from data. This technique is useful when data is limited. This technique trains multiple ML models with random samples selected from the training data with replacement. Then the predicted results are averaged to offer the best prediction in this technique. The decision tree is the most used algorithm for bagging (Marapelli 2020; Yu et al. 2018; Berk 2005).

Boosting

Boosting makes predictions using the number of weighted individuals' accuracy determined by the votes of the weak learners. By this technique, the weak learners are trained to become better predictors. AdaBoost is the most used Boosting technique(Marapelli 2020; Yu et al. 2018; Berk 2005).

Ensemble learning techniques are utilised to tackle the issue of class imbalance in the design of EWS (Knowles 2014). Class imbalance arises when data is skewed to a particular instance. Studies have shown that combined models significantly increased the results' accuracy (Adnan et al. 2021; Zemel 2014). The goal of any ML model is to produce an optimum outcome, and ensemble methods consider that by combining different models and averaging them to produce one final model.

This study's data gathered from students' behaviour and characteristics may include biases that can lead to class imbalance. For instance, the data obtained from an institution's Learning Management System (LMS) may contain 80 % examples of students who do not need intervention and 20% examples of students who need intervention. This can create a problem of class imbalance as the data contains more data from one instance than another. The class imbalance issue in training data is proven to be minimixed using ensemble methods. Also, ensemble algorithms have been employed to develop models that generate useful models for the

previously unseen environment. The objective of the training is to empower the model to retrieve important facts from the database in order to identify future input patterns or to generalize the model (Farrahi *et al.*, 2019). The ultimate goal of the ML model is to achieve generalization, which means being able to apply what it has learned to new situations. This study makes use of an ensemble learning technique to produce models that predict accurate outcomes for unseen data. This study used the bagging ensemble learning methods to design an EWS to predict students at risk of drop-out by averaging the results of different models to achieve an accurate and generalized system.

1.3 Problem Statement

The successful completion of students enrolled in higher education is a continuous global concern. The OECD 2019 report on education indicates that 12% of students in developed countries who enter programmes of higher education on a full-time basis leave the tertiary system before the beginning of the second year (OECD 2019). Although students enrolled in higher education in African countries have appreciated, the Minister of State in charge of Education in Ghana confirmed that the estimated rate of student dropout stands at almost 50% in several African countries. According to the United Nations Educational Scientific and Cultural Organization (UNESCO) report, Sub-Saharan Africa has an estimated 42% school dropout rate estimated to be the highest global school dropout rate (UNESCO 2012).

Due to the high cost of dropout to society and the economy (Lee & Chung 2019; Knowles 2014), EWS has been developed as an intervention to curb the menace of student dropout in different areas across developed countries such as the USA, Australia and the UK (Christie, Jarratt, Olson & Taijala 2019; Ferguson, Clow, Griffiths & Brasher 2019; Lee & Chung 2019). In the United States of America, about 730,000 students do not complete high school on time every year (Lakkaraju et al. 2015). For instance, the state of Wisconsin has developed the Dropout Early Warning System (DEWS) to predict student dropouts. A good number of public schools in the USA implemented similar interventions between 2014 and 2015 (Lee & Chung 2019).

Developing countries face similar problems of dropouts in the educational sector, but such inventions are not present to leverage the situation. This is evidenced in a report by the Ghana

Fact Sheet (2020), stating that about 71% of children complete primary education. However, completion rates declined to 35% at the upper secondary level (UNICEF 2020). The report further indicated that the steep decline in the completion rate at the upper secondary is a result of dropout, repetition or delayed completion. Studies have shown that those who dropout of school experience higher unemployment rates, negatively impacting their life earnings and life expectancy (Islam Sarker, Wu & Hossin 2019; Knowles 2014). These negative impacts poor life earnings and short life expectancy—translate into long-term debilitating consequences for countries. For example, a country may experience a chronic skilled labour shortage, unleashing economic havoc on its productive capacity. Additionally, increases in school dropouts undermine a nation's productive capacity and drain resources that could be used for development projects. In other words, monies intended for developmental projects would have to be tailored to subsidise the unemployed community (Lee & Chung 2019).

Ordinarily, human judgement and heuristics are usually used as standards to identify problems such as absenteeism, tidiness and would-be dropouts in some educational institutions. However, these standards are susceptible to errors due to their subjective nature. Since no one-size-fits-all standard exists for all institutions, just a rule-based algorithm may not be a solution. Thus, the ML approach would be considered in the design of an EWS to objectively identify ARS earlier and more accurately in the Ghanaian context.

1.4 Research objectives and research questions

This study aims at investigating and utilizing Learning Analytics and ML to analyze characteristics and behaviours that best provide an early indication of students that need intervention to improve academic performance. This study will focus on the following specific objectives.

Specific Objectives

1. To identify the characteristics that distinguish students who did not graduate from those who did.

2. To develop and compare various ML algorithms to determine the best predictive model for higher education.
3. To build ML-based predictive model for predicting student dropout.

This study will answer the following research questions:

1. What are the most prominent features/characteristics of students who graduated and those who did not graduate from higher education?
2. Which emerging ML tools/techniques have been applied successfully in the design of predictive systems in education?
3. Which emerging ML model can be adapted to predict student dropout in higher institutions in Ghana?

1.5 The rationale for the study

The study focuses on higher educational institutions in Ghana. Available records from Ghana Communication Technology University (GCTU) indicate the continuous decline of students who graduate from programs they registered for each academic year. In 2018 and 2019, the number of students admitted to undergraduate programs was 1,588 and 1,881, respectively. However, the graduation figures did not commensurate with the number of enrollments. The student records indicate that 78.63% and 35.04% did not graduate in 2018 and 2019, respectively, and this results from dropouts or students not completing their studies on time.

It was observed that individual staff members have different approaches to determining students at risk of leaving school. These approaches include using computerised data on students and manual data recorded by course teachers. Since there is no “single technique” for determining ARS, most faculty members rely on different standards and methods to determine students’ outcomes in a course. Some standards and methods commonly used are students’ past performance in the course, attendance, classroom participation, and assignment completion. The

different standards and methods, however, do not predict student performance early before it happens. The researcher observed that the standards and methods used do not provide accurate results as they are primarily subjective, undermining their efficiency and reliability in predicting students at risk of leaving school before completion. As a result, the researcher explored other productive ways to predict students at risk of dropping out of school.

After an in-depth literature review on designing an effective predictive system for students at risk of dropping out of school, it was decided that a ML approach would be the best solution to predict ARS. Using the “bagging ensemble learning” methods to design a predictive system for higher institutions in Ghana would solve the problem of a high rate of dropouts in higher institutions in Ghana. The rationale for this study is to predict ARS using the ML method based on “ensemble learning” techniques.

1.6 Privacy and Ethics

The Data Protection Act of 2012 in Ghana stipulates definitive regulations and guidelines concerning collecting, handling, using, and sharing personal data and information. These provisions aimed to ensure that all individuals' privacy rights are respected and safeguarded in accordance with legal requirements (Act 843, sections 22, 23, and 24 set out the rules for gathering, using and disposing of data). The current research adhered to the guidelines and ethical considerations authorized by the Ghana Communication and Technology University (GCTU) as outlined in Ghana's data protection Act 2012. The guidelines ensure that the data of all humans (students) involved in the current research are safeguarded. As such, every piece of data gathered in this study strictly adhered to the privacy and security rules of GCTU.

1.7 Outline of chapters

The study was organised under the following chapters:

Chapter One dealt with the background of the study, the research problem, the need for an ensembled learning-based solution to the problem, the objectives, and the rationale for carrying out this study.

Chapter Two looked at available and relevant literature supporting the research and discussed the research gap to motivate its relevancy.

Chapter Three discussed how the predictive model was developed, using the 5-stepped process of the learning analytics model and the ML workflow.

Chapter Four compared different approaches to developing a predictive system using traditional and ensemble techniques for the OULAD dataset and reporting the results.

Chapter Five presented a discussion of the results and the research implications.

Chapter Six concluded the study with limitations and recommendations for future research.

1.8 Summary

This chapter put the entire study into context by focusing on the problems relating to high rates of dropouts and the designing of systems that aid in predicting ARS in higher institutions in Ghana. An ensemble learning-based solution was suggested to solve these problems. The research objectives and the rationale for conducting this study were also focused upon, this was followed by an in-depth literature review that justifies conducting this study.

2 Chapter Two - Literature Review

2.1 Introduction

This section addresses literature relevant to the study. The first section presents a review of the Tertiary education landscape in Ghana from the era of pre-independence and post-independence. The next discusses the need for predictive systems (an EWS) in the context of Ghana. Subsequently, a review of Learning Analytics (LA) and the most commonly used tools in Higher Education is presented. Followed by discussions on the various ML methods and prominent features used in the design of predictive systems in Education. Additional ensemble methods currently used in developing predictive systems are discussed, followed by the feature selection method and its relevance. Finally, the summary of the chapter and the research gap are presented.

2.2 A review of the Tertiary education landscape in Ghana

This section focuses on the systems of tertiary education in the pre-independence era and the post-independence educational system in Ghana. Discussions on the structure, challenges and existing state of Tertiary education are also presented.

Education in Ghana is divided into three phases: basic education (kindergarten, primary school, lower secondary school), secondary education (upper secondary school, technical and vocational education) and tertiary education (universities, polytechnics and colleges). Education is compulsory between the ages of four (4) and fifteen (15) (Girdwood 1999). However, this study focused on tertiary education, specifically university education; henceforth, all discussions in this section are limited to tertiary education at the university level.

Tertiary education or higher education institutions span institutions that are beyond secondary education, such as colleges, universities, and post-secondary specialized educational institutions. These include teacher training colleges, nurses training colleges, agricultural training colleges, polytechnic and technical education training centres, labour colleges, police and army staff

training colleges, and vocational training colleges (Ghana National Council for Tertiary Education 2018).

The significance of tertiary education in advancing Ghana's socio-economic growth and development cannot be overemphasised. Ghana aims to ensure equitable access to tertiary education (Ayisi 2018). According to the World Bank report on education, tertiary enrolment in Ghana has steadily increased, reflecting the country's commitment to reducing disparities in access to higher education (Blunch 2020). This is in line with education's goal of promoting social mobility and inclusion. Tertiary education is a crucial aspect of equipping individuals with specialised knowledge and skills in their chosen fields. Ghana has demonstrated a profound understanding of the significance of human capital in driving economic growth and enhancing competitiveness (Acquah 2021; Cisneros 2020).

2.2.1 Pre-Independence education

During the pre-independence era, few Ghanaians enrolled in tertiary education due to the fewer established educational institutions in Ghana. Most people were granted scholarships to study abroad in institutions of higher education at the time. These periods saw few people enrolled in educational institutions in Ghana, and therefore problems of large classes and student dropout were of less concern to the institutions available. In the 1940s, efforts to develop more educational institutions led to the establishment of the University of Ghana and its auxiliary institutions (Apusigah 2009). The main purpose was to promote university education, learning and research among citizens and member states (Joe Adu-Agyem and Patrick Osei-Poku 2012). The mode of instructional delivery at that time was face-to-face, requiring teachers and students to be present in class for teaching and learning. Due to the smaller size of classes, students' academic performance could be easily monitored without using any technological tools.

2.2.2 Post-Independence – Today

Ghana has stood out among several Sub-Saharan African (SSA) countries regarding educational advancements since its independence (Atuahene & Owusu-Ansah 2013). In the post-independence era, the quest for tertiary education for citizens became even higher, and several universities, polytechnic colleges, training colleges for teachers and nurses, as well as agriculture training, were established by the Government (Bawakyillenuo, Osei-Akoto, Ahiadeke, Aryeetey & Agbe 2013). With time and as the country's population increased, these schools began to experience an increase in enrollment against fewer teachers coupled with limited available infrastructure (Apusigah 2009). The issue of the teacher-student ratio remains a challenge in institutions of higher education. The student-teacher ratio for the 2016/2017 academic year of public universities for the Applied science, Technology and Health Sciences programme was 27:1 (Ghana National Council for Tertiary Education 2018). Monitoring student performance in courses enables teachers to track the performance of students in a class to identify and address students with challenges in education. However, due to large classes, teachers are unable to monitor the progress of students, which has led to the student failing in courses and some cases, dropping out of school. In other words, teachers are unable to monitor the performance of students to identify and provide assistance to those who might need help. Enrolment for the 2012/2013 academic year for public universities in Ghana indicates that 128,116 students were enrolled into tertiary education in Ghana, however, a total of 50,957 students were recorded to have graduated (Ghana National Council for Tertiary Education 2018).

The problems with large class sizes and limited infrastructure informed higher institutions to find alternative instructional modes that could create a balance in class. In addressing the problem of infrastructure, large class sizes and student attrition, institutions adopted distance learning programmes (Awiagah, Kang & Lim 2016; Kumi-Yeboah 2010; Osei 2010) and the use of technology. This mode of teaching and learning does not require a student to travel from one location to another to study (Kuranchie, Okyere & Larbi 2021; Djan & George 2016). Collins (2010) studied student perception of distance learning education by the KNUST. The study observed that distance education was highly patronized by students, especially students who are already employed and the married class since this mode gave them the flexibility and comfort to

study without being present in the classroom. With the emergence of the internet and advancement in technology, the use of online studies is adopted in education. The use of online studies has been proven to provide more benefits for students in terms of monitoring and providing feedback interventions. However, the outbreak of the COVID-19 pandemic has necessitated higher institutions in Ghana to embrace the use of Technology in teaching and learning.

In light of the pandemic, almost all institutions in Ghana resorted to the use of online education as an additional mode of instruction (Edumadze, Barfi, Arkorful & Baffour Jnr 2022; Adarkwah 2021; Ali 2020; Upoalkpajor & Upoalkpajor 2020). This period saw a massive increase in the use of various Learning Management Systems (LMS) in teaching and learning in Ghana. The opportunity of gathering student footprints based on their interactions with various systems of learning is enormous. To most educators, this change has been a swift one. The most interesting aspect is that the recent online learning systems come with the combined fields of data analytics and ML tools used in business and industries (Webster, Andre & Giang 2019). Therefore, obtaining data for analysis has become handy for the design of predictive systems that can inform teachers about the performance of their students and also facilitate the development of interventions to prevent student dropout in the context of Ghana.

2.3 The need for context-based predictive systems to address student dropout in Ghana

The challenge of student dropout continues to be a concern in Ghana. However, research in systems that can predict students with academic difficulties in higher institutions is limited (Wandera, Marivate & Sengeh 2019). Predictive systems, like EWS, have effectively addressed student dropout in countries such as the US, UK, and Australia. EWS solutions implemented in developed countries require the use of data, statistical tools and ML algorithms for efficient early-warning intervention systems (Mothukuri et al. 2017). Moreover, since most ML systems perform effectively based on their trained environment and the dataset provided, context is of the essence when considering such implementations. It is, therefore, worthwhile to focus on ML

implementation in Ghana since it is considered underrepresented in the ML literature and differs in a variety of ways from developed countries (De-Arteaga, Herlands, Neill & Dubrawski 2018).

When exclusively considering development issues in ML in the developing world, there are notable challenges, such as the availability of data, computational capacity and Internet accessibility (De-Arteaga et al. 2018). However, the study on ML for the developing world has provided guidelines for ML development in developing countries (De-Arteaga et al. 2018). Therefore, this study discusses the importance of context in the development of an EWS in Ghana using the following properties.

a) The geographical scope of the applications and data is limited to developing countries.

The first property addressed the importance of using local datasets to solve specific problems in different areas in developing countries. Coleman (2021) established the necessity to rebuild and validate detectors based on different contexts because of the different data semantics and various risk factors. This study will be relevant to building predictive models for educational institutions in the Ghanaian context because studies have established that warning signs are context-specific (Coleman 2021).

b) The issues being addressed are related to a vital development area for the region in question.

ML techniques for predictive systems depend largely on data for training, and such adaptive systems may react differently from one environment to the other. (Goel, 2020). In this regard, This study will train ML-based models using the local dataset to ensure the efficient and effective operation of the model, which may be incorporated into the educational curriculum. For example, (Mgala & Mbogho 2015) created intervention prediction models with a total of 2,426 student data. Based on the research, it has been demonstrated that the development of successful intervention prediction models can be achieved with a relatively limited quantity of data. The use of LMS has evolved in most tertiary institutions in Ghana. These systems store data about

students' records, activities conducted, resources accessed, grades, number of forum posts, course participation, attendance etc. There also exist databases that store student demographics and other important records. This study will retrieve student data from LMS and integrate emerging LA tools and ML algorithms for the analysis.

c) The problems being addressed or contextual elements need solutions that differ from those that exist or are feasible in developed countries, and the proposed solution effectively handles these differences.

There have been instances technologies perform perfectly well in one environment but such systems turn to perform differently when implemented in different environment. In other circumstances, the problem may exist all across the world, but the context in underdeveloped countries necessitates that feasible remedies be fundamentally different (De-Arteaga et al. 2018). In developing predictive systems, some features, behaviours and characteristics of students in developing countries are mostly different from what pertains to developing countries. For example, if we consider time(hours) spent accessing resources online as a feature for determining the performance of students in the developing world context, there is a likelihood to have misleading results. The sense that students might not have access to internet service always and, therefore, might have downloaded most resources when the internet is available. Also, in measuring the percentage of activities conducted by the student to predict their failure in a particular course, there are well-known factors such as "adding terms to glossaries" and "contributing to Wikis" by authors such as Macfayden and Dawson (Macfadyen & Dawson 2010). However, these factors might be different in the perspective of Ghana and, therefore, may not influence the decisions of a model in the context of Ghana. Also, access to resources that are available to a student in developing countries tends to differ from what exists in developed countries; therefore, the criteria for predicting study success cannot be the same in every setting/ environment. For example, some major causes of dropout in Ghana have been identified as poverty, child labour, teenage pregnancy and distance to school(Adam, Adom & Bediako 2016), which differ from what may exist in developed countries. Given the inherent differences in

context and availability of resources in developing regions, this is a typical difficulty(De-Arteaga et al. 2018).

d) The proposed solutions incorporate ML as a key component of the projects.

The use of ML and LA methods in education offer enhanced teaching and learning practices in education (Flanagan & Flanagan 2018; Sciarrone 2018). This property addressed the importance of using ML as a component of a solution. While these methods/tools are well established in developed countries (Ferguson et al. 2019) in the development of models for predicting academic success, their implementation appears to be less common in Ghana. The proposed study, however, incorporates ML as a key component of the solution.

2.4 Overview of Learning Analytics and Commonly used tools

In recent times, there has been a plethora of research in the field of LA as a result of online gathering and analysis of data on student's academic progress, to evaluate academic performance, to forecast future outcomes and problems (Moubayed, Injadat, Nassif, Lutfiyya & Shami 2018; Baars, Stijnen & Splinter 2017; Martin & Ndoye 2016; Amara Atif, Ayse & Mauricio 2013; Romero, Espejo, Zafra, Romero & Ventura 2010; Zafra & Ventura 2009).

Various LA tools are used to analyse, visualize, predict, and provide personalized support for students. For example, the open learning Initiative is a teaching platform that Carnegie Mellon University in Pittsburgh developed. It creates and hosts web-based modules that give users access to teaching materials through various media elements such as text, animation and audio recordings. The systems allow users to practice what they have learnt and provide personalised feedback based on the users' performance.

Another well-known LA tool is the Social Network Adapting Pedagogical Practice (SNAPP). The University of Wollongong in Australia collaborated with other universities to develop the SNAPP LA tool. SNAPP aims to identify students at risk of academic underperformance because

of lower participation levels than their peers. The system generates visual representation through social network diagrams of students' interactions, activities and patterns of behaviour in learning management systems. Similarly, the Digital Dashboard for learning (DDL) is a tool that provides teachers with easy access to key student learning indicators through simple visual graphics within a web browser. The dashboard provides teachers real-time information about class performance and where students need special attention. The system provides a dashboard enabling teachers to design teaching that allows ARS to cover study materials and improve course design and assessment (Doneva, Gaftandzhieva & Bandeva 2021).

In the area of predictive analytics, the Degree Compass is a recommender system that helps students to predict specific course modules, review users' curricula, show modules history and previous grades are compared with the data of a previous student with a similar profile to recommend a module that best fit the user. Degree Compass is developed at Austin Peay State University in Tennessee (Whitten, Clarksville, Sanders & Stewart 2013). Other analytical tools developed and implemented in Higher educational institution includes the AWE early alert system designed by the University of England. The system uses data from student interaction with portals and a learning management system to identify students who may be experiencing difficulties in their studies (Atif, Richards & Bilgin 2015).

Most of the LA tools described above provide insight into how LA uses data footprints left by students when they engage with digital technology during the learning process to help teachers and administrators make better decisions, such as devising educational interventions. Higher Institutions have benefited immensely from LA tools in developing tailored solutions for improving students' performance. The tools have equally assisted teachers in monitoring the academic performance of students. LA depends largely on educational data, which has given rise to the use of Educational Data Mining (EDM) analysis (Kostopoulos, Karlos & Kotsiantis 2019). The EDM analysis identifies patterns in data using ML algorithms (Alboaneen et al. 2022). The tools used in LA for addressing educational challenges are usually similar to the tools used for EDM analysis. The study fields of LA and EDM are all intertwined (Viberg, Hatakka, Bälter & Mavroudi 2018) and are mainly used interchangeably.

Meanwhile, a distinct difference exists in their techniques and emphasis (Chen, Xie, Zou & Hwang 2020; Romero & Ventura 2020). EDM is interested in developing strategies for examining the distinct forms of data generated in educational settings (Romero & Ventura 2020). Methods like “Clustering”, “Classification”, “Bayesian” modelling, “relationship mining”, and “model discovery” are some of EDM's most often used techniques (Chen et al. 2020). These methods are typically used to create more effective learning environments by disclosing relevant information that may be used to adjust course structure or to aid in predicting student performance and behaviour (Kabathova & Drlik 2021).

LA, on the other hand, is concerned with the collecting, analysis, and reporting of student data and characteristics in order to understand better and improve learning and the settings in which it occurs (Kabathova & Drlik 2021; Romero & Ventura 2020). Statistics, “visualization”, “discourse analysis”, “social network analysis”, and “sense-making” models are among LA's most widely employed approaches (Chen et al. 2020; Tsai et al. 2020). LA focuses on using predictive models to make data-driven decisions and integrating technical, social, and pedagogical dimensions of learning (Romero & Ventura 2020). Data from LA are employed in three types of analyses: descriptive, predictive, and prescriptive (Nitu, Dascalu, Lazarou, Trifan & Bodea 2018). Predictive analysis forecasts expected trends and consequences for students before their experience (Nitu et al. 2018). In contrast, the descriptive analysis focuses on reporting student performance to other stakeholders through online interactions (Ifenthaler & Yau 2020). Interventions for many stakeholders in the educational community are dictated by prescriptive analysis. Prescriptive analytics makes decisions based on algorithmic models (Baker Seimens 2015).

The distinctions between the two communities (EDM and LA) are due to their respective “focus”, “research topics”, and eventual usage of models than to the methodologies employed (Romero & Ventura 2020). Deducing from the characteristics of both EDM and LA presented above, this study presents it in simple terms. EDM is the application of technology to find new patterns in data and develop new algorithms and models to address significant educational issues. In comparison, LA focuses on addressing educational issues by integrating data-driven decisions

with social, technical and pedagogical elements in learning by applying validated prediction models.

It is worth noting that both EDM and LA have an interest in data-intensive approaches to educational research, as well as a desire to improve educational practice (Romero & Ventura 2020; Liñán & Pérez 2015). Regardless of the differences between Learning Analytics (LA) and EDM, there exist similar characteristics, interests, and purposes (Chen et al. 2020). Other multidisciplinary topics are covered by EDM and LA. In reality, they can be depicted as a fusion of three major sections. Computer science, Education, and Statistics (see FIG 1).

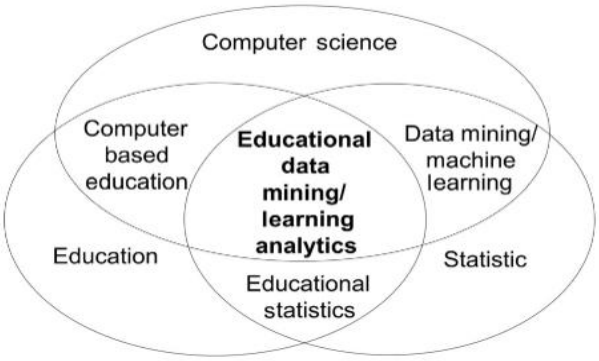

Figure 1:Relationship between LA, EDM and ML(adapted from (Romero & Ventura 2020))

Figure 1 shows the main areas related to educational data mining, machine learning and learning analytics, adapted from (Romero & Ventura 2020). Other subfields closely related to EDM and LA are “computer-based education” (CBE), “data mining” and “machine learning”, and “educational statistics”, which are formed by the intersection of these three sections (“Computer science”, “education” and “statistics”) (Romero & Ventura 2020). The goal of LA and EDM is to figure out how students learn. LA and EDM aim to support educational research and practice by analyzing large-scale educational data (Viberg et al. 2018). Continuous research is conducted in education using both LA and EDM approaches.

One of the most often investigated subjects in the EDM and LA areas is the EWS for predicting a student's learning success or failure (Kabathova & Drlik 2021). For example, EDM and similar LA methodologies have been utilized to investigate whether a learner is at risk of failing to learn

programming skills (Alyahyan & Düştegör 2020; Ihantola et al. 2015; Berland, Baker & Blikstein 2014). Studies (Adnan et al. 2021; Behr, Giese & K 2020; Lee & Chung 2019) have addressed predicting student attrition and dropout and increasing academic success using EWS. Other areas that EDM and LA have addressed include understanding learning behaviours, processes and strategies (Uzir et al. 2020; Matcha, Gašević, Uzir, Jovanović & Pardo 2019). For example, through the lens of LA, Martin & Ndoye (2016) analysed online learner-centred assessment and how it helps with online teaching and learning, as well as track students' progress and take remedial steps for evaluation based on the students' data. Another area of LA research is course recommendation systems for personalising learning (Mamcenko 2018; Mamcenko & Kurilovas 2017).

LA research and implementation started over two decades in developed countries; however, its implementation in developing countries is still low. Most literature on the field of LA is concentrated in developed countries. Different implications have been identified in previous studies of LA, including user behaviours and engagement modelling, predictive modelling and personalized and adaptive learning models. This study is particularly interested in predicting ARS in higher education using LA. The next section focused on the ML methods and variables used to implement predictive modelling or EWS in LA.

2.5 Methods and Variables for Implementing Predictive Systems in Education

This section discussed the commonly used input variables and methods for the design of predictive systems and EWS to address student dropouts and improve study success in education. The use of EWS and predictive systems are used to mean the same systems in this study. The use of features and the geographical context of the various studies are also presented. This study extends the literature on the characteristics and methods for implementing EWS to include the geographical context and the algorithms used for the implementation. The literature includes the period from 2015 to 2022.

Table 1:List of studies on Predictive models in education

Autho	Year	Method	Input data	Geographical context	Purpose	Result
Alboaneen, D., Almelihi, M., Alsubaie, R., Alghamdi, R., Alshehri, L. and Alharthi, R.,	2022	"Support vector machine(SVM)", "Random Forest (RF)", "K-Nearest Neighbor (KNN)", "Artificial neural network" and "Linear Regression (LR)"	"student academic data" and "demographic"	Saudi Arabia	To identify students with a high probability of failing a course	The findings indicated that predictions Mean Absolute Percentage Error(MAPE) reached 6.34%
Pellagatti, Ieva and Paganoni, (2021	2021	"Generalised Mixed-effect Random Forest (GMERF)"	"Personal information", "career track", and "previous studies"s	Italy	proposed a new statistical method called "Generalised Mixed-effect Random Forest(GMERF)" that works on hierarchical data	Random Forest predicted 90% of dropouts successfully
Dass, Gary and Cunningham,	2021	"Random Forest"	"MOOC platform Open edX class report", "assessment report", "progress report", "time" and "topic_report"	Arizona State University USA	Predict MOOC dropout and success	The Random Forest achieved a performance accuracy of 87.6%.
Bainbridge et al.	2021	"Logistic Regression"	"Demographic" and "Dynamic variables related to course participation"	USA	analyze characteristics and behaviours that best provide an early indication of a student being academically at risk	"Logistic Regression" model is the best technique for predicting at-risk students
Adnan et al	2021	"Random forest"	"clickstream data" and "student demographics"	UK	to predict at-risk students at different percentages of course length	"Random forest" outperformed the other models
Ghorbani, R. and Ghousi,	2020	"Random forest", "K-Nearest-	"students' performance"	Iran	Analyse students' performance and	The random forest

Autho	Year	Method	Input data	Geographical context	Purpose	Result
R., 2020.		Neighbour", "Artificial Neural Network", "XG-Boost", "Support, Vector Machine" "Radial Basis Function", "Decision Tree", "Logistic regression", and "Naïve Bayes"			address the class imbalance dataset problem	classifier outperformed the other models. SVM-SMOTE emerged as the best resampling method
Chui, K.T., Fung, D.C.L., Lytras, M.D. and Lam, T.M.,	2020	"Reduced training vector-based machine (RTV-SVM)"	"Student activities", "demographics"	UK	To predict at-risk and marginalized students	the proposed model achieved an overall accuracy of 92.2-93.5% in predicting at-risk students and 91.3-93.5 % in identifying marginalized students.
Bañeres, D., Rodríguez, M.E., Guerrero-Roldán, A.E. and Karadeniz, A.,	2020	"Decision Tree", "K-Nearest Neighbor", "Naïve Bayes", "Support Vector Machine"	"Corse work", "behaviour", "continuous assessment"	Canada	EWS to predict at-risk student	At-risk students were detected with higher accuracy outcome
Behr, Giese and K,	2020	"Random forest"	"grades of previous education", "determinants of student satisfaction", and "self-assessment"	Germany	Predict student dropout	The model achieves a measure of 0.86 using the Area under the curve (AUC) evaluated metric.
Hashim, Awadh and Hamoud	2020	"Decision Tree", "Naïve Bayes", "Logistic Regression", "Support Vector Machine", "K-Nearest Neighbour", "Sequential Minimal Optimisation and	"demographic", "academic background" and "behavioural features"	Iraq	compared different supervised ML algorithms for predicting student performance	"Logistic regression" method was the most accurate

Autho	Year	Method	Input data	Geographical context	Purpose	Result
		Neural Network"				
He et al	2020	"Joint RNN-GRU". "Joint Neural Network"	"Personal biographical information" and "sequential behaviour data"		model to predict at-risk students within a virtual learning environment	the model predicted 80% prediction accuracy
Vaughan and Durandt	2019	"logistic regression", "multiple regression tree and classification tree"	"Cumulative average quiz marks" and "final period marks" considered by the study.	South Africa	investigated the most suitable predictive model to identify at-risk students	The findings report that the logistic regression and multiple regression models showed accurate results.
Fairos, Yaacob and Nasir,	2019	"K-Nearest Neighbor", "Naïve Bayes", "Decision Tree", and "Logistic Regression Model"	"student name", "student ID", "gender", "course grade", and "final CGPA" for	Malaysia	to predict student performance	findings indicated that the "Naïve Bayes" outperformed all the other classification models
Burman and Som,	2019	"Multiclassifiers support vector machine"	"students' psychological parameters", "Learning strategies", and "Socioeconomic status"	India	predicted students' academic performance	the findings revealed that the RBK outperformed the linear kernel by producing better results
Jae Young Chunga et al	2019	"Random forest"	"attendances", "student activities", and "learning activities"	Korea	To predict students at risk of dropping out from high school students	binary classification performed excellently in predicting dropout students.
Xu, X., Wang, J., Peng, H. and Wu, R.,	2019	"Decision Trees", "Neural networks", and "Support Vector Machine"	"Online duration," "Internet traffic volume", and "connection frequency"	China	To predict student performance	behaviour discipline plays a vital role in student performance accuracies of all the techniques was improved as the number

Autho	Year	Method	Input data	Geographical context	Purpose	Result
						of features increased
Liao, S.N., Zingaro, D., Thai, K., Alvarado, C., Griswold, W.G. and Porter, L.,	2019	"Support vector machine"	"student clicker data" and "final exam scores"	USA	Identify a student who will fail a course	The proposed methodology works across different institutions
Gray, C.C. and Perkins, D.,	2019	"Descriptive statistics"	"Attendance"	UK	Predict at risk-risk student	A combination of algorithms yielded a significant outcome with an accuracy of 97%
Hussain, M., Zhu, W., Zhang, W., Abidi, S.M.R. and Ali, S., 2019	2019	"Artificial neural network (ANNs)". "Support vector machine(SVMs)", "Logistic Regression", Naïve Bayes classifiers", and "Decision Trees"	"average time, total number of activities", "average idle time", "average number of keystrokes", and "total related activity"	China	Predict student performance	ANNs and SVMs achieve higher accuracy
Lee, S. and Chung, J.Y.,	2019	"Random Forest", "Boosted Decision Tree", and "SMOTE"	"attendance", "behaviour", and "course performance"	South Korea	Predict dropout and improve the performance of EWS	boosted decision tree showed the best performance.
Jokhan, A., Sharma, B. and Singh, S.	2019	"Regression Model"	"student logins", "completion of online activities" and "online engagement"	University of the South Pacific	To predict student performance	The accuracy of the model was 60.8%.
Zhuping Wang	2018	"Bayesian Naïve"	"library records", "student behaviour in dormitories", "attendance, behaviour", and "course work"	China	to determine low grades, dropout and minimise delays in graduation	findings indicated that monitoring students' library borrowing and grades were the significant indicators
Emma Howard	2018	"Neural Networks", "K-Nearest	"students' background information";	University College Dublin	investigates how to approach developing an	"BART" outperformed the other

Autho	Year	Method	Input data	Geographical context	Purpose	Result
		Neighbours" and "Random Forest", "BART" and "XGBoost"	"students' engagement with LMS"; and "continuous assessment results"	(UCD).	accurate prediction model for an early warning systems	prediction models tested at the optimal time of weeks 5–6
Gerald Baars	2017	"Logistic regression"	"pre-admission variables" and "post-admission variables"	Netherlands	develop a model predicting students who fail or pass	The findings indicated that the earliest time with the highest specificity to predict student failure in the first-year curriculum is at six(6) months
Yulei Panga	2017	"Naive Bayesian classifier"	"Student demographic", "performance metrics" and "psychological" and "educational factors"	USA	development of a predictive model based on an ensemble support vector machine for predicting students' graduation	the proposed ensemble support vector machine was an effective model
Ornelas and Ordonez	2017	"Naive Bayesian classifier"	"engagement indicators" and "performance indicators"	USA	to forecast student success	"Naıve Bayesian" opted over the logistic regression model
Marbouti et al.	2016	"Support Vector Machine", "K-Nearest Neighbors", and "Naive Bayes Classifier"	"scores", "grades for quizzes", and "written exam"	USA	modelled early prediction of at-risk students	The "Naïve Bayes" model was among the algorithms that produced the best outcome.
Sangodiah *et al.*, 2015	2015	"Support Vector Machine"	"students' academic" and "non-academic data"	Malaysia	to minimize student attrition in higher learning Institutions	the "Support vector machine" obtained better accuracy
Siri	2015	"Artificial Neural Network"	"students' personal data" and "educational" and "academic	Italy	to create a model to forecast students at risk of dropping out of school	The "artificial neural network" model was

Autho	Year	Method	Input data	Geographical context	Purpose	Result
			careers".			considered a valid tool for predicting 84% of dropouts correctly
Mgala, M. and Mbogho, A.,	2015	"Logistic regression", "Multilayer perceptron", "Sequential minimal optimization algorithm (SMO)", "Bayesian network" classifiers, "Naive Bayes classifier", "Lazy learners", "Random forest classifier ", "J48 algorithm"	"Students' demographic data", "behaviour ", and "attitude data", parent and school factors	Kenya	Predict academic performance.	The result indicated that logistic regression was the most suitable prediction model for the type of dataset used in the study.
Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R. and Addison, K.L	2015	"Random Forests", "Adaboost", "Logistic Regression", "Support Vector Machines" and "Decision Trees"	"academic performance", "behaviour", and "attendance".	U.S. school districts	To predict students at risk of adverse academic outcome	Random For-est model outperforms all the other models
Kubayi, S.C., Jadhav, A. and Ajoodha, R.	ND	"K-nearest Neighbour", "Random forest", "Decision Trees", "Naïve Bayes", "Logistic regression" and "Multi-layer perceptron"	"Biographical characteristics", "Pre-College observations", "University Enrolment observations"	South Africa	Predict at- risk student	The random forest emerged as the highly performed classification model with an accuracy rate of 83%, a precision of 83%, a recall of 82% and an F1 score of 83%

This study combined studies on predictive analysis and EWS in education to address the most prevalent ML techniques or algorithms used in modelling LA predictive systems. The study further identified the nature of input data and the geographical context where they are tested. The following section presents discussions in Table 1.

2.5.1 Applications based on machine learning algorithms and Academic Dataset

This section focused on the ML techniques employed in the development of predictive systems, the variables used, and the geographical context of the studies retrieved. ML methods are used in the design of EWS in Education. Various ML techniques have been utilized in the design of LA predictive systems. The systems have contributed to the design of LA interventions for students, learners and teachers in higher education sectors.

RF is an “ensemble learning technique” (Behr et al. 2020). It works by constructing multiple decision trees during the training phase. “Decision trees” are based on nodes at each level. The results of each node are determined by a split into the subsequent node on a condition. This continues until the final output is achieved. The decision of the majority of the trees is voted by the random forest as the final decision. In predicting students at risk of dropping out of high school in Korea Lee and Chung, (2019), utilize the RF algorithm. The study obtained 165,715 high school students from the 2014 “National Education Information System” (NEIS). The “NEIS” contains national education information connected through the Internet with around 12,000 elementary and secondary schools, 17 city/provincial offices of education, and the Ministry of Education in Korea. The study used 12 features from the “NEIS” dataset encapsulated in “attendance”, “student activities”, and “learning activities” during classes to predict student dropout, given the various measures of performance for binary classification performed excellently in predicting dropout students (Chung & Lee 2019).

In a similar study, Ghorbani and Ghousi, (2020), designed a predictive model to analyse students' performance. The focus of this study was on addressing the class imbalance dataset problem by comparing different oversampling techniques to identify the best method to apply. The datasets utilized features that determined the academic performance of students at universities in Iran and Portugal. Evaluating the performance of the resampling methods to solve the class imbalance

problem, the “RF”, “KNN”, “ANN”, “XG-boost”, “SVM”, “Radial Basis Function”, “DT”, “LR”, and “Naïve Bayes” were employed. The study concluded that the “RF” classifier outperformed the other models, and the “SVM-SMOTE” emerged as the best resampling method among the others (Ghorbani & Ghousi 2020). This study considered classifiers that were uncommon in the literature for the model.

Another study (Behr et al. 2020) employed an uncommon classifier. The study predicted student dropout using the conditional inference tree similar to the random forest technique in solving the problem of a higher attrition rate among students in Germany. The study focused on predicting early student dropout based on binary classification. The developed model includes three phases, thus, “the pre-entry phase”, “the decision phase”, and “the early study phase”. The model achieved a measure of 0.86 using the “Area under the curve (AUC)” evaluation metric. The study pointed out that the early study phase contributed to the better performance of the model. The findings of the study indicated the best indicators for predicting graduate or dropout students in the grades of previous education, determinants of “student satisfaction and self-assessment”. The findings also reveal that study progress is influenced by the “final grade at secondary school (grade school)”, “the type of secondary school (school type)”, “the type of school leaving qualification (qualify max)”, and “the number of repeated classes (rep class)” (Behr et al. 2020).

Also, Pellagatti, Ieva and Paganoni, (2021) proposed a new statistical method called Generalised Mixed-effect Random Forest (GMERF) that works on hierarchical data to predict students' dropout. They designed a “RF” model based on decision tree aggregates that could analyse hierarchical data, thus data with varying responses and having an exponential sequence. The proposed” Generalised mixed-effect random forest (GMERF)” based on “RF” predicted 90% dropouts successfully. The study predicted students’ probability of graduating using “student-level characteristics” and “grouping structure of degree students in the Engineering programme” (Pellagatti et al. 2021). The study findings suggested early performance of students is an important indicator for predicting dropout, which is similar to the observation remarked in (Behr et al. 2020). Although the algorithm proposed could handle hierarchical data, the study concentrated on binary classification response in the simulation. Therefore, the effects of the algorithm on hierarchical data cannot be established in the study.

The model constructed by Dass, Gary and Cunningham, (2021) predicted whether students would dropout or continue in the “MOOC” course. The study obtained data from self-paced mathematics courses in the “MOOC platform open edX” of Arizona State University. The students' data were grouped into four thus, “class report”, “assessment report”, “progress report”, “time” and “topic_report”. The “RF” achieved a “performance accuracy” of 87.6%. The findings of this study imply that utilizing a ML technique with “RF” for predictions can provide trustworthy outcomes. The authors argue for further research into just-in-time LA intervention to support students learning, improve success and minimize dropouts (Dass et al. 2021). This study concentrated on the “MOOC platform”, suggesting that online student data pertaining to attendance and behaviour and coursework were utilized. These accounts differ in the dataset used in the previous studies (Pellagatti et al. 2021; Behr et al. 2020).

Similarly, Adnan *et al.* (2021) adopted different ML algorithms to predict ARS at different percentages of “course length”, and the metrics used to measure them were based on “accuracy”, “precision”, “support”, and “f-score” metrics. The study employed the freely available “Open University Learning Analytics Dataset (OULAD)” provided by the “Open University” UK students. The predictive system considered the use of “clickstream data” (students' interaction with the virtual learning environment in the form of numbers of clicks during the course timeline) and “student demographics” for training. The study concluded that the “RF” outperformed the other models by attaining good measuring scores (Adnan et al. 2021). Similar to the dataset used by Dass, Gary and Cunningham, (2021), this study focused on the student clicks during the course timeline and demographic for training the model. However, the study introduced evaluation metrics in their studies to measure the “accuracy”, “precision”, “support”, and “f-score” metrics of the models.

Alboaneen *et a.*, (2022) conducted a similar study on predicting student final scores by developing a web-based predictive system. The system used student “academic data” and “demographics” to predict students' “final scores” and identify students with a high probability of failing a course. The model used students'” academic data” and “demographics of female students” at the computer science department of the College of Science and Humanities at Imam Abdulrahman bin Faisal University (IAU). The ML algorithms employed include “Support

vector machine (SVM)", "Random Forest (RF)", "K-Nearest Neighbor (KNN)", "Artificial neural network", and "Linear Regression (LR)". The study findings indicated that "RF" was employed for the model because it had the "lowest Mean Absolute Percentage Error (MAPE)" of 6.34%. The study suggested that academic considerations had a greater influence on students' academic achievement than demographic factors, with the top midterm exam score (Alboaneen et al. 2022).

Lakkaraju *et al.*, (2015) focused on predicting students at risk of not finishing high school on time. The study considered a longitudinal study with two school districts with a combined enrollment of 200,000 students. The features considered include students' "GPAs", "absence rates", "tardiness", and "gender". In order to identify students who may be at risk of not graduating on time, the study conducted tests using various methods such as "RF", "Adaboost (AB)", "LR", "SVM", and "DT". The findings suggested that the "RF" outperformed all the other models (Lakkaraju et al. 2015).

Kubayi, Jadhav and Ajoodha, (no date) on the other hand, focused on predicting student academic performance using ML techniques and other statistical analysis and data mining tools in higher learning institutions in South Africa. The study used the following predictive models, "KNN", "RF", "DT", "Naïve Bayes", "LR" and "Multi-layer perceptron". The "RF" emerged as the highly performed classification model with an "accuracy rate" of 83%, "precision" of 83%, "recall" of 82% and "F1-score" of 83%. Although this study evaluated five (5) ML algorithms to select the best-performing method, the model cannot be used in a real-life scenario because the training dataset was generated synthetically (Kubayi et al. n.d.).

2.5.1.1 Artificial Neural Network

Many studies have used ML and deep learning approaches, such as "Artificial neural networks (ANN)", to predict student performance (Chen et al. 2020). The human brain serves as a model for "ANN". Neurons are the fundamental building blocks of artificial neural networks (Tahiru & Agbesi 2021). Neurons are organized into three layers: the input layer receives data, the hidden layer processes it, and the output layer predicts it. Perceptrons take in data and process it by sending it from the input layer to the hidden layer and then to the output layer. A weight

functions similarly to a score that is passed from one neuron to the next. Every weight is multiplied by a bias, which is a constant. A prediction is made using “neurons”, “weights”, and “biases”. For example, in predicting students’ dropout at University, the “ANN” was used to create a model to forecast students at risk of dropping out of school. The study population was made up of 810 students enrolled for the first time in a healthcare professions degree course at the University of Genoa. The variables considered were “students' personal data” and data on their “educational” and “academic careers”. The “ANN” model was considered a valid tool for predicting 84% of dropouts correctly (Siri 2015).

Artificial Neural Network has also been indicated to perform well on both static and sequential data; for example, He *et al.* (2020) used a joint RNN-GRU (joint Neural Network) model to predict ARS within a virtual learning environment. The proposed joint model predicted 80% prediction accuracy for students at risk of failing at the end of the semester. The study made use of statistics, personal biographical information and sequential behaviour data from the Open University learning analytics dataset (OULAD) (He et al. 2020).

A similar study by (Alkhasawneh & Hobson 2011) developed two neural network models using a feed-forward backpropagation network to predict retention for students in the science and engineering fields. The study used a total of 338 samples used, and 70.1% of students were classified correctly. This project used neural network models to build a framework that predicts incoming freshmen in Science and Engineering disciplines retention at Virginia Commonwealth University (VCU) using students' current available overall GPA as the dataset (Alkhasawneh & Hobson 2011).

2.5.1.2 Support Vector machine

The “SVM” technique utilises a hyperplane that optimises margins to distinguish the instances of two classes. Support vectors are, therefore, the points close to the border. “SVM” techniques can be utilised in both classification and regression problems (Marapelli 2020).

Burman and Som, (2019) predicted students’ academic performance using a “Support Vector Machine” in order to improve the performance of students using data mining tools. The

researchers used a “Multi classifiers support vector machine” to classify the learners into low, average and high categories based on their scores. The study focuses on non-intellectual parameters of students which affect their study and academic growth. The variables considered were students' psychological parameters which cover “Personality”, “Motivation”, “Psychosocial contextual influences”, “Learning strategies”, “Approach to learning”, and “Socioeconomic status”. Comparing the “Radial Basis Kernel” (RBK) with the “Linear Kernel” (LK), the findings revealed that the RBK outperformed the linear kernel by producing better results (Burman & Som 2019).

Another study was carried out to minimize student attrition in higher learning Institutions in Malaysia using “SVM” to predict students with the likelihood of failing before it happens. The features considered include students' academic and non-academic data. The probation factor was considered one of the features that can be used to predict ARS and the “SVM” was used to predict the probation rate, which leads to dismissal. According to the study, the “SVM” showed improved performance accuracy (Sangodiah et al. 2015).

Al-shehri *et al*, (2017) utilized both “SVM’ and “KNN” to estimate student performance in final exams at the University of Minho in Portugal. The “SVM” outperformed “KNN” with a correlation coefficient of 0.96. The authors used solely online Learning Management System (LMS) data to predict student outcomes in a course (Al-shehri et al. 2017).

Alternatively, Chui *et al.*, (2020) proposed the use of a “training vector-based machine” (“RTV-SVM”) to predict at-risk and marginalized students. The research presented a reduced “RTV-SVM” that eliminates redundant training vectors while keeping the support vectors. A dataset consisting of 32593 university students from seven courses was chosen from the OULA dataset of the largest university in the UK for performance evaluation. The authors analyzed factors like "demographic information of students" and "engagement with a digital learning platform". The results of the study indicated that the proposed model achieved an overall accuracy of 92.2-93.5% in predicting ARS and 91.3-93.5 % in identifying marginalized students. Even though this study used the same method, the dataset was varied to include marginalised students (Chui et al. 2020).

Liao *et al.*, (2019) developed a model that can identify students who are likely to fail a course early on in the semester. This helps in predicting their overall academic performance. The study employed the “SVM” to train and identify ARS. The features used include student clicker data and exam scores in one term of students at two universities in North America. The study demonstrated that the “SVM” classifier effectively predicted ARS in different courses across different universities (Liao et al. 2019).

Pang *et al.*, (2017) proposed a predictive model that predicts students' likelihood of graduating using ensemble methods. The model uses an ensemble “SVM” that has proven to be effective in anticipating graduation outcomes. The model designed by the authors incorporates diverse features, including demographic information of students, university performance metrics like grade point averages and earned credits, as well as psychological and educational factors. The model used data from the Southern Connecticut State University (SCSU) in the USA. The findings indicated that the proposed ensemble support vector machine was an effective model (Pang et al. 2017).

Hussain *et al.*,(2019) utilized the digital electronics education and design suite (DEEDS), a technology-enhanced learning (TEL) system, to analyze students' data and anticipate any potential difficulties they may encounter while studying digital design courses. The authors employed “ANNs”. “SVMs”, “LG”, “naïve bayes classifiers” and “decision trees”. The study used the following indicators, “average time”, “total number of activities”, “average idle time”, “average number of keystrokes”, and “total related activity”. The study results showed that “ANNs” and “SVM” achieve higher accuracy than other algorithms. The study considered the “k-fold cross-validation” for calculating the characteristics and root mean square error metrics as the metrics for evaluating the model's performance (Hussain et al. 2019).

2.5.1.3 Naïve Bayes

Study Fairos, Yaacob and Nasir, (2019) developed a classification model using a supervised ML approach to predict student performance at a university in Malaysia. ML algorithms such as the “KNN”,” Naïve Bayes”, “Decision Tree”, and “Logistic Regression” were used to predict students’ performance. The study considered the following variables; “student name”, “student

identity number", "gender", "course grade", and "final CGPA" for training the model. The researcher evaluated the performance of the models using metrics such as "accuracy", "precision", "recall" and "ROC curve". The findings indicated that the "Naïve Bayes" outperformed all the other classification models (Fairos et al. 2019).

In a study conducted by Marbouti, Diefes-dux and Madhavan, (2016), authors developed a model for predicting ARS in a course using standards-based grading. Furthermore, feature selection methods were used in selecting the number of variables in the model in order to improve the generalizability of the model. The study used secondary data collected during the Spring 2013 and Spring 2014 semester offerings of a first-year engineering course at a large Midwestern U.S. university. The variables used included the "homework scores" and grades for "quizzes" and written "exams". The study utilised the" Naive Bayes Classifier" model along with "SVM", and "KNN". The "Naïve Bayes" model was among the algorithms that produced the best outcome (Marbouti, Diefes-Dux & Madhavan 2016).

Alternatively, Wang *et al.*, (2019) proposed a new dimension of adding "library records" and "student behaviour" in dormitories to the "attendance", "behaviour", and "course work" (ABCs) to determine low grades and dropouts and minimise delays in graduation. The system works by tagging students at risk level as having low grades, the potential of dropout or delayed graduation. The study employed the "Naive Bayesian" algorithm to determine the risk level of 1712 students at Hangzhou Normal University. The findings indicated that monitoring students' library borrowing and grades were the most significant indicator, as the algorithm achieved a higher accuracy of 86% with the selected features (Wang et al. 2019).

In the work of Ornelas and Ordonez, (2017), the authors suggested a "Naive Bayesian" classifier that was implemented in a dozen Rio Salado Community College (Arizona, USA) courses. The authors used data from the institution's LMS as input, which was separated into two categories: engagement indicators, which include LMS logins and participation in online activities, and performance indicators, such as performance in online activities and points earned in course tasks. For eleven distinct courses, the classifier was able to forecast success, that is, a student receiving a C or better with an accuracy of over 90%, though not early enough to function as an

EWS. With a training sample of 5936 students and a validation sample of 2722, this experiment was conducted on a rather large population.

2.5.1.4 Logistic regression

Logistic regression predicts the correlation between numerical variables. Simple linear regression will have one dependent variable and one independent variable; however, in multiple regression, more than one independent variable will describe the value of the dependent variable (Marapelli 2020).

Hashim, Awadh and Hamoud, (2020) established a student performance prediction model based on Supervised ML Algorithms. The study compared different supervised ML algorithms for predicting student performance at the University of Basra. The study employed characteristics of students such as demographic, academic background and behavioural features. The following supervised classification methods were compared "DT", Naïve Bayes, "LR", "SVM", "NN", "Sequential Minimal Optimisation" and "ANN". The study concluded that the "LG" method was the most accurate in predicting the final grade of students, with 68.7% passing and 88.8% failing. Also, Mgala and Mbogho, (2015) predicted the academic performance of students in Kenya. The study compared different classifiers such as "LR", Multilayer perceptron, "Sequential minimal optimization algorithm" ("SMO"), "Bayesian network", "Naive Bayes", "Lazy learners", and "RF". Features that were trained included "Students' demographic", "behaviour" and "attitude data", "parent", and "school factors". The result indicated that "LR" was the most suitable prediction model for the type of dataset used in the study (Hashim, Awadh & Hamoud 2020b).

In a similar study, Bainbridge et al. (2015) reported on the "LR" model as the best technique for predicting ARS in an online graduate course on Public Affairs and Administration Education. Eventually, the author selected the "LR" model after testing different data mining analytics, which produced a similar result. The study used a standard set of" demographic" characteristics, such as student "age" and "gender", as well as static variables at the time of the course, such as their "full-time" or "part-time" status, the "class size", whether or not they are on academic probation and their cumulative "GPA". Other dynamic data related to course participation include a "partial grade" in the class, the number of "forum posts", the number of times the

“content is read”, the number of “forums read”, the number of “sessions opened”, number of “assignments submitted” by the student, number of “exams taken and submitted” by the student, and number of “assignments read” by the student. The author discovered that relatively simple LA models are effective. The study identified and analysed characteristics and behaviours that best provide an early indication of a student being academically at risk. The authors paid particular attention to the use of online tools, as well as using data generated by students of an online Master of Public Administration program drawn from the Marist College Open Academic Analytics Initiative (Bainbridge et al. 2015).

Baars, Stijnen and Splinter, (2017) develop a model predicting students who fail or pass the first year of an undergraduate medical curriculum at Erasmus medical school. Predictive variables included pre-admission variables such as age, gender, pre-university education GPA, the way students were selected, and post-admission variables such as the number of credits obtained, degree of participation in exams, and exam success rate of data on pre-and post-admission variables was collected from 1819 students of five consecutive cohorts. The findings indicated that the earliest time with the highest specificity to predict student failure in the first-year curriculum is at six (6) months (Baars et al. 2017).

Also, Van Appel and Durandt, (2019) investigated the most suitable predictive model to identify ARS in the Business statistics course at the University of Johannesburg. The predictive analysis methods employed include a based method,” LR”, “multiple regression”, “regression tree”, and “classification tree”. Cumulative average quiz marks and final period marks were the variables considered by the study. The findings report that the “LR” and “multiple regression” models showed accurate results (Van Appel & Durandt 2019).

In the same way, Jokhan, Sharma and Singh, (2019) developed An EWS to predict student performance in first-year IT literacy courses at the University of South Pacific. The study used captured data from students' interaction with Moodle, such as “logins” and “activities completion” rates. The “LR” was able to determine the correlation between students' online behaviour and their performance with an accuracy of 60.8% (Jokhan et al. 2019).

2.5.1.5 Bayesian network

Lacave, Molina and Cruz-Lemus, (2018) employed the Bayesian Network algorithm to predict the dropout of computer science students at the University of Castilla-LaMancha(Spain). The study made use of student academics and demographic data. Academic and social data of the students enrolled in the CS degree were used to train the model. Bayesian Network was considered an appropriate model to be used in the context of LA, according to the findings (Lacave et al. 2018).

Similarly, Howard, Meehan and Parnell, (2018) investigated how to approach developing an accurate prediction model for EWS at an undergraduate level at University College Dublin (UCD). The author considered a dataset from the statistics university course's weekly continuous assessment and resources from the learning management system Blackboard. The variables used in this study were divided into three categories: students' background information, students' engagement with LMS, and continuous assessment results. The study identified weeks 5-6 as the optimal period for implementing an EWS. They used predictive methods like BART and XGBoost, which are uncommon in the data analytics literature as well as common predictive methods such as Neural Networks, K-Nearest Neighbours and Random Forest. The Bayesian Addictive Regression Trees (BART) outperformed the other prediction models tested at the optimal time of weeks 5–6 (Howard et al. 2018).

2.6 Other Ensemble Learning Methods.

This section presents other ensemble learning methods that have been utilized in developing predictive systems in education. The Gradient boosting and Catboost algorithms are discussed.

2.6.1 Gradient Boosting

Nagy & Molontay 2018, used ML algorithms to identify students at risk of dropout and predict student dropout of university programs based on data from 15,825 undergraduate students from Budapest University of Technology and Economics. The models were tested using 10-fold cross-

validation, and the best models were “Gradient Boosted Trees” and “Deep Learning”. The study identified key features that were predictive of student dropouts, such as program ID, Freshman or re-enrolled, financial and APS calculating methods (Nagy & Molontay 2018).

Tenpipat & Akkarajitsakul 2020, researched factors affecting undergraduates' educational status and created binary classification models for predicting their educational status. DM and ML techniques were used to collect data from “KMUTT's” internal data sources. The results showed that the prediction accuracy of the “GB”, “DT”, and “RF” models was 93%, 92%, and 92%, respectively. The top 5 important features were the student's academic year’ high school GPA’, ‘channels of university admission’, ‘student's faculty’, and ‘gender’ (Tenpipat & Akkarajitsakul 2020).

This study used “GB” and “DT” to predict dropout in MOOC using students' learning activities data. The authors found that their model achieved 89% accuracy in predicting dropout students (Liang, Li & Zheng 2016).

Hussain *et al.*, (2019) used ML algorithms to identify low-engagement students using the Open University dataset. The input variables were the “highest education level”, “final results”, “score on the assessment”, and the number of “clicks on VLE” activities. Results showed that “J48”, DT, JRIP, and “Gradient-boosted” classifiers exhibited better performance than other tested models. A dashboard designed for instructors was established to aid in the assessment of student involvement in VLE courses. (Hassan et al. 2019).

2.6.2 CatBoost

Catboost is a well-known gradient-boosting library created by the Russian search engine Yandex(Abdulkareem, Foh, Lee, Carrez & Moessner 2022). It is designed to work well with categorical features and has been proven effective in various ML tasks, including classification, regression, and ranking (Abdulkareem et al. 2022). Oreshin et al. (2020) used ML approach to predict academic outcomes using socio-demographic, psychometric, and educational data from the LMS and social network. The catBoost accurately predicted students' outcomes with an accuracy of 91% (Oreshin et al. 2020). Another study demonstrated how a generic predictive

model can be developed to identify ARS across a wide variety of courses. Experiments were conducted using a range of algorithms, with the CatBoost algorithm performing the best with an accuracy of 75% on the dataset. The study also noted that assignment grades, both current and prior grades are very important in predictions (Ramaswami, Susnjak & Mathrani 2022). Also, Mingyu, Sutong, Yanzhang & Dujuan (2022) proposed a predictive method for student academic crisis warning. The proposed method using Catboost has significant advantages over common ML algorithms in terms of achievement prediction, providing intuitive decision support and guidance assistance for education administrators (Mingyu et al. 2022).

2.7 Feature Importance

This section discussed studies that have employed feature selection methods and their relevance.

Feature importance is a technique used in ML to determine which features or variables have the greatest impact on the output or prediction of a model(Pudjihartono, Fadason, Kempa-Liehr & O'Sullivan 2022). It helps to identify the most relevant and useful feature in the dataset and can aid in feature selection or engineering. The importance of feature selection can significantly affect the performance of predictive models. Selecting the right set or redundant feature can lead to overfitting and poor model performance (Odhiambo Omuya, Onyango Okeyo & Waema Kimwele 2021). In higher education, it has been employed to identify influential factors for identifying student performance and each level of the course (Lu et al. 2018; Marbouti, Diefes-dux, et al. 2016; Márquez-Vera et al. 2016).

Several methodologies have been proposed for feature importance analysis, addressing different types of data and algorithms. In traditional statistical approaches, methods such as "chi-square test"," t-test", and "analysis of variance" ("ANOVA") (Faulconer, Griffith & Frank 2021; Nácher, Badenes-Ribera, Torrijos, Ballesteros & Cebadera 2021; Valle et al. 2021; Yerel et al. 2021) have been used to evaluate the significance of individual features. Information-theoretic approaches, such as mutual information and entropy, have been employed to measure the information gain provided by each feature (Mainali et al. 2021; Chen & Yang 2016). Tree-based

algorithms, including “Random Forest” and “Gradient Boosting”, have gained popularity for their intrinsic feature importance measures based on node impurity or information gain (Hasan et al. 2020; Akçap, Altun & Petek 2019; Baranauskas, Netto, Nozawa & Macedo 2018). Additionally, permutation-based methods, such as permutation importance and mean, decrease accuracy, have been introduced to estimate feature importance by permuting feature values (Oh 2022; Gómez-Ramírez, Ávila-Villanueva & Fernández-Blázquez 2020; Fisher, Rudin & Dominici 2019). Research has shown that incorporating feature importance can improve “model interpretability”, “reduce dimensionality”, and enhance” prediction accuracy”. Models can achieve better generalisation and robustness by removing or down-weighting less important features. Feature importance analysis has also been used for feature selection and feature engineering, enabling the identification of relevant subsets of features or the creation of new features derived from the most important ones. Different feature importance methods may produce varying results, necessitating careful consideration and validation. It is crucial to assess the stability and reliability of feature importance estimates through cross-validation and sensitivity analysis (Heinze, Wallisch & Dunkler 2018; Taskin, Kaya & Bruzzone 2017). Moreover, feature importance should be interpreted in the context of the specific problem domain, as certain features may exhibit complex interactions and dependencies.

2.8 Relationship between students' online habits and performance

There is a growing body of research that suggests a relationship between student “behaviour” and “interactions” (the number of clicks) on a virtual learning environment (VLE) and” students' final results” (Martin & Bolliger 2018; Cho & Tobias 2016; Kent, Laslo & Rafaeli 2016). Student interactions on VLEs encompass various activities, such as accessing course materials, participating in online discussions, submitting assignments, and engaging with learning resources. Numerous studies have highlighted the positive relationship between student interactions on VLEs and academic performance. Higher levels of engagement and frequent interactions with course content have been associated with improved learning outcomes and higher final results. For example, (Alqurashi 2019), explored how “online learning self-efficacy” (“OLSE”), “Learner-content interaction” (“LCI”), “learner-instructor”(“LII”), and “learner-learner interaction” (“LLI”) can predict student satisfaction. The author concluded that “Learner-

content interaction" ("LCI") emerged as the most significant and important predictor of student satisfaction. Another study analysed the relationship between students' online work habits and academic performance. Using a data-driven approach, the authors measured student habits on timelines, regularity and intensity. The study found that students who work on assignments early and regularly achieved high GPAs and high grades in the course. The study also concluded that high-achieving students show vastly different work habits from low-achieving students (Dvorak & Jia 2016). Another study (Coldwell, Craig, Paterson & Mustard 2008) investigated the relationship between participation, "demographics", and "academic performance" of students in an online course. The authors found a relationship between students participation in in online learning environment and their final result. The study also identified the relationship between gender, nationality, participation and final result. On the other hand, there was no relationship between age, participation and final results. The study suggested that student demographics should be considered an important determinant for the design of online learning. Also, (Kuzilek, Hlosta & Zdrahal 2017), used Markov chain modelling to identify behavioural patterns leading to drop-out or passive withdrawal in VLE. The results show that interesting patterns can be uncovered when combined with the information about the submission of the first assessment.

Alternatively, research has also identified negative associations between specific online habits and academic performance. For example, (Cho & Tobias 2016) examined the role of online discussion in student learning experiences. The study used the Unified Theory of Acceptance and Use of Technology to investigate the students' perceptions. The metrics for the study included community inquiry, learning time satisfaction and achievement. The findings showed no significant difference among time spent on the blackboard, course satisfaction and student achievement.

Deducing from the (above) studies that have investigated the relationships among student interactions on VLEs, late submission of assignments, the length of revision, and final results, their findings suggest that student interactions on VLEs positively influence assignment submission patterns and the length of revision. Engaged students who actively participate in online discussions and utilize learning resources are more likely to submit assignments on time and allocate sufficient time for revision. Furthermore, there is evidence to suggest that timely

assignment submission and adequate revision positively impact final results, leading to improved academic performance.

2.9 Research Gap

Many studies in LA are concentrated in developed countries, where the USA (Jokhan et al. 2019; Liao et al. 2019; Ornelas & Ordonez 2017) and the UK (Adnan et al. 2021; Chui et al. 2020; Gray & Perkins 2019) contribute more of the research in the field. Other countries like China (Hussain et al. 2019; Wang et al. 2019; Xu, Wang, Peng & Wu 2019) and Australia (Pellagatti et al. 2021; Behr et al. 2020; Siri 2015) are also doing better in LA research and implementation. Meanwhile, limited studies were found in developing countries. Only two studies in South Africa (Van Appel & Durandt 2019; Kubayi et al. n.d.), one article in Kenya (Mgala & Mbogho 2015) and barely any study on the research focus was recorded in Ghana from the databases the search was conducted, and the period specified. This indicates a gap in research and implementation of LA and ML application in education in Ghana. In this regard, this study focused on developing a predictive model to identify students with a high risk of dropping out of a higher educational institution in Ghana. This will serve as a basis for developing tailored LA interventions in higher institutions in Ghana.

Many studies have designed models predicting student failure and success, student attrition, student performance and final grades. Most of these studies have focused on dataset from only VLE. Nonetheless, this study focused on the design of a model to predict students at risk of not graduating from the university using student activity logs, student courses, assessments, and student demographics and social features.

Previous studies have designed predictive models using variables such as students' academic data, student demographic factors, final grades, attendance, online activities and behaviours. However, few studies have considered students previous school attended and previous grades in the design. Also, it was noted that researchers did not focus on the socio-economic status of students, distance travel by students, and economic status. This model is designed to include the

highest education obtained by students and students' socio-economic backgrounds. As emphasized by (Liz-Domínguez, Caeiro-Rodríguez, Llamas-Nistal & Mikic-Fonte 2019), geographical context and student demographics are considered significant aspects in the design of a good predictive algorithm. A study (Dawson et al. 2017) indicated that including student-specific characteristics has a greater influence on developing predictive systems.

The most common classifiers employed in the literature include “SVM”, “RF”, “KNN”, “ANN”, and “LR”. Studies (Marbouti, Diefes-dux and Madhavan, 2016; Alboaneen *et al.*, 2022; Fairos, Yaacob and Nasir, 2019; Hashim, Awadh and Hamoud, 2020) considered this sequence of classifiers in a single study and evaluated their performance on the “accuracy”, “precision”, “f1-score”, “recall” and “AUC” metrics. The current study developed and compared ensemble classifiers such as “RF”, “GB”,” Catboost”, “KNN”, “LR” and “NB” and evaluated the strength and weaknesses of the algorithms based on “accuracy”, “precision, ”recall”, and “f1-score”.

The SLR indicated that most studies employed the RF algorithm (Lee and Chung, 2019; Behr, Giese and K, 2020; Ghorbani and Ghousi, 2020; Pellagatti, Ieva and Paganoni, 2021; Dass, Gary and Cunningham, 2021; Adnan *et al.*, 2021; Alboaneen *et al.*, 2022; Kubayi, Jadhav and Ajoodha, (no date)), SVM (Burman and Som, (no date); Sangodiah *et al.*, 2015; Al-shehri *et al.*, 2017; Chui *et al.*, 2020 ; Liao *et al.*, 2019; Pang *et al.*, 2017; Hussain) and the LR algorithm to develop predictive models. In contrast, the least recorded studies were presented on the Bayesian network (Lacave, Molina and Cruz-Lemus, 2018; Howard, Meehan and Parnell, 2018). No one analytical method or ML method was considered supreme or best over the other. All the algorithms used indicated a reasonable accuracy rate based on their training dataset. Therefore, this study cannot decide based on only literature to establish which methods are considered the best design for predictive models. The current study, therefore, developed and compared six (6) different traditional and ensemble ML algorithms and evaluated the models' performance-based “accuracy”, “precision”, “recall,” and “f1-score”. The findings outlined in section 4.6 align with the literature discussed. The “RF”, which extends decision tree algorithms, served as the base model for the comparison. The RF ensemble algorithms are known to address issues of class imbalance (Ghorbani & Ghousi 2020; Lee & Chung 2019) and have shown evidence of better performance with predictions (Behr et al. 2020; Ghorbani & Ghousi 2020; Lee & Chung 2019).

There is also a dearth of research that distinguish the behaviour and characteristics of students that graduate from students who do not graduate from higher educational institutions. Studies (Martin & Bolliger 2018; Cho & Tobias 2016; Kent et al. 2016) have investigated the relationship between students' performance and online activities, but no study has addressed the differences in graduating and non-graduating by analysing behaviour in assessing learning materials before module start time, revision pattern before exams and assignment submission in VLE. The current study utilized an exploratory analytical approach to distinguish between graduating and non-graduating students.

Previous studies have utilized supervised ML algorithms on various datasets (Lee and Chung, 2019; Behr, Giese and K, 2020; Ghorbani and Ghousi, 2020; Pellagatti, Ieva and Paganoni, 2021; Dass, Gary and Cunningham, 2021; Adnan *et al.*, 2021; Alboaneen *et al.*, 2022; Kubayi, Jadhav and Ajoodha, (no date)). However, this approach may not be optimal due to the issue of dimensionality, which results in lengthy training times and an overfitted predictive model. Therefore, it is recommended to reduce the total number of features to a more manageable level by using the tree-based feature importance model to select the most crucial features before training the model. This study extends the literature on the best features for developing predictive systems to identify students at risk of dropout in higher education using the tree-based feature selection to predict student dropout in minimal training time to achieve good performance.

2.10 Summary

This chapter presented an extensive review of literature relevant to the study. The chapter addressed the tertiary landscape in Ghana, describing the current trend of mode delivery and the presence of technology to enhance educational data analysis. It further discussed the need for context-based predictive systems in Ghana to serve as a contribution to the research and implementation in the field of ML. Subsequently, an overview of LA and the common tools used were presented. Also, other fields that constitute LA systems, such as educational data mining and ML, were discussed. The chapter also presented literature on the ML techniques used in the

development of LA predictive systems focusing on the most used algorithms, variables used and the geographical context. The chapter also discussed other ensemble learning algorithms used in predictive systems, the importance of feature selection in developing predictive systems and the effects on students' online engagement and performance. Finally, the research gap and conclusion of the chapter followed.

The next section discussed the methodology used in conducting this study.

3 Chapter Three – Model development using an integrated learning analytics and machine learning framework

3.1 3.1 Introduction

The methodology for this study is divided into two parts. The former described the systematic procedure of the search strategy used for conducting the literature review. The latter presented the description of data used for the study, the integrated LA framework and the ML workflow utilized for the overall model development in this thesis. This study focuses on predicting students at risk of dropping out of a module using the OULAD. The result of this study can be used to inform academic stakeholders early on whether students will dropout of a course or program. Hence, the result can be used to identify significant characteristics that impact students dropping out of Higher Learning Institutions and equip academic stakeholders with the requisite information in designing an intervention that will retain and assist students to progress in their education.

3.2 Search strategy for the literature review

This section presents the search strategy for the systematic literature review on the most used ML algorithms in predictive systems.

Data retrieval was conducted using keywords from existing studies, including “machine learning”, “learning analytics”, “higher education”, and “predictive systems”. The Web of Science databases, Scopus, Science Direct, and springer link databases were considered for collecting articles related to the research area. These databases are of high quality and have higher indexes. The published articles, journal papers and conference proceedings were searched using titles, abstracts, and keywords. In order to obtain more other references, the forward and backward snowballing search approach was adopted. The search was conducted with the following search strings (“Learning Analytics” OR “learning analytics” OR “LA” OR “Machine Learning” OR “machine learning” OR “ML” OR “Predictive Systems” OR “prediction” OR “algorithm” OR “Algorithm” OR “at-risk student” OR “performance” OR “Performance” AND

"higher education" OR "Education."). The literature search resulted in 215 articles as of 1st February 2022. The screening was conducted to exclude studies irrelevant to LA and ML methods in education. The inclusive and exclusive criteria adopted include articles published from 2015 to 2022 in peer-reviewed journals, conference proceedings, articles written in English and articles related to the subject of the research. The abstract of the relevant 30 articles was manually read with specific data extracted from each article. The extracted data includes the name of the publication, the year of publication, the student characteristics/ variables used, the geographical context where the research was conducted, the aim of the research and the result or findings obtained. The discussion was grouped into LA predictive models with specified ML methods. For example, studies that utilized the Random Forest method to predict ARS were grouped under one heading, and this followed through with all the other methods obtained. After the groupings and discussions, the study presented a summary that focused on the research gap and justification of the proposed method for the study.

3.3 Method for Model Development

The LA process designed by Campbell and Oblinger (2007) and Pardo (2014) described five stages process of capture, report, predict, act and refine. The five-stage process was adopted for this study to design and develop a predictive model to predict students at risk of dropping out. The study integrated ML workflow into the five processes of LA to develop the model to answer the following research questions.

1. What are the most prominent features/characteristics of students who graduated and those who did not graduate from higher education?
2. Which emerging machine learning tools/techniques have been applied successfully in the design of predictive systems in education?
3. Which emerging machine learning model can be adapted to predict student dropout in higher institutions in Ghana?

3.4 Overall data and participant description

For this study, a student is considered at risk of not graduating or not completing a module if the student's graduation status indicates "withdrawn". This definition was chosen because the OULAD dataset clearly and succinctly categorises student outcomes as "Distinction", "Pass", "Fail", or "Withdrawn" for all module presentations. An overview of the dataset is provided below.

3.4.1 Description of the Open University Learning Analytics Dataset (Participants)

Data collected for this study were raw data of student information stored on the OULAD database. It includes data on student registration, courses, assessment, virtual environment and demographics. The dataset contains data from courses presented at the Open University (OU). OU is a British University that principally offers undergraduate and postgraduate studies off campus. It is considered one of the largest distance-learning institutions, with over 170,000 students. It stores recorded information about student demographics and interaction with the learning materials in a data warehouse spanning 2013 and 2014. Modules represent courses in the OU dataset. This dataset is available publicly to support the research in LA. The dataset was selected because it contains data on student demographics, courses, as well as student interaction in the form of a click-stream from an online Learning Management System. The dataset is suitable for this study because it is structured and devoid of all ethical issues. Comparatively, the OULAD dataset reflects the structure and standards of educational datasets required for institutions in Ghana. An overview of the student Information system and the LMS of GCTU in Ghana confirms that about 80% of the features in the OULAD are present and could be used to predict ARS if such data is made available.

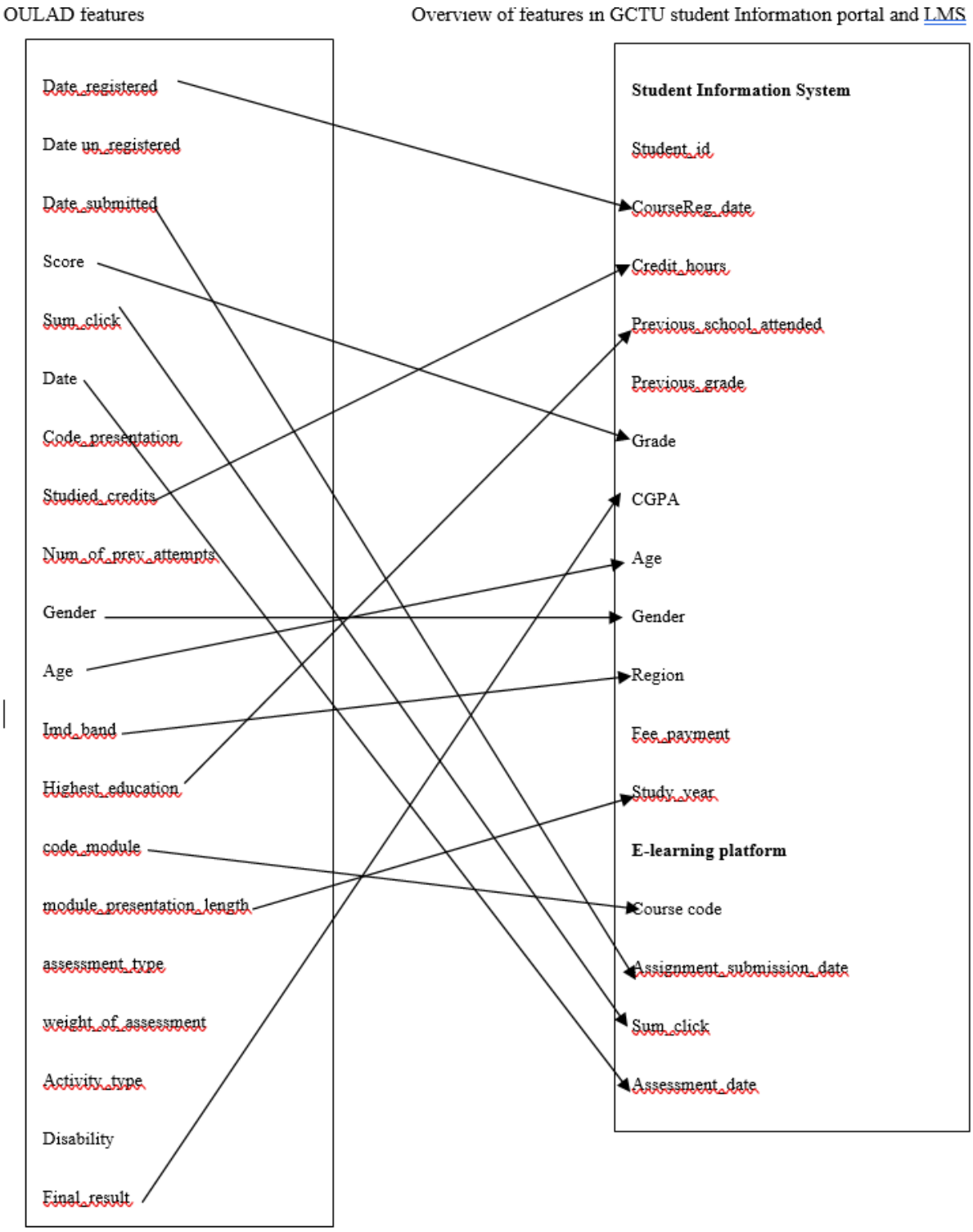

Figure 2:Mapping target features in OULAD to GCTU

Therefore, utilizing the OULAD dataset will produce equivalent results to employing a dataset sourced from Ghana. It is widely recognized in the LA field that utilising online datasets can offer valuable insights for addressing educational issues but can be challenging. One of the reasons why gathering data from educational settings can be challenging is because it takes a lot of time and involves strict adherence to principles of data ethics, privacy, and consent in order to safeguard the rights of individuals involved (Romero & Ventura 2020). For example, (Aljohani, Fayoumi & Hassan 2019; Hlosta, Zdrahal & Zendulka 2017) and other researchers have researched using the student click stream in the OULAD and achieved varied results in their works. Furthermore, the OULAD is selected based on the portable feature of the datasets. Thus, it addresses all principles of data ethics, privacy and protection consent, unlike generating data from an educational environment that can be time-consuming. The OULAD dataset can be extremely helpful in analyzing student characteristics and identifying those at risk of not completing their higher education. This dataset can also be utilized to create accurate models predicting which students are likely to dropout, improving the ability to support and help them succeed.

3.4.2 Data variables in the OULAD dataset

Modules are presented multiple times in the year and the month (denoted by alphabet), and the year of the module is used to differentiate each module. For example, a module starting in January ends with A, in February with B and so on; so that '2013J' means that the presentation started in October 2013. Each module is an independent module with no pre-requisites qualification from previous modules. Module presentation represents student study groups consisting of 20 students in a group. A tutor is assigned to each Module-presentation for guidance and support throughout the module presentation. The length of a presentation is nine (9) months. An assessment usually represents the final exam at the end of each module. The dataset contains information about 32593 students and 10,655,280 summaries of student click

streams of the student during interactions with the virtual learning environment. The data on student interaction consist of different activities on various modules, with each activity signifying a different aspect of the online learning environment. Examples of activities include students' interaction with course content, submission of an assessment, visits to discussion forums, participation in video discussions, etc. 20 different types of activities are represented in the dataset, which includes ‘data plus’, ‘dualpane’, ‘external’ ‘quiz’, ‘folder’, ‘forums’, ‘glossary’, ‘homepage’, ‘html activity’, ‘oucollaborate’, ‘oucontent’, ‘ouelluminate’, ‘ouwiki’, ‘page’, ‘questionnaire’, ‘quiz’, ‘repeat activity’, ‘resource’, ‘shared subpage’, ‘subpage’, and ‘url’. Each activity denotes a specific behaviour in the learning environment. The data is represented in tables using identifier columns and presenting the student as the central point. The dataset consists of several files, including student demographics, VLE data, interactions with the VLE represented by clickstream data, assessment, and module information, as shown in Figure 1. OULAD is certified by the Open Data Institute (http://theodi.org/) and can be downloaded from https://analyse.kmi.open.ac.uk/open_dataset.

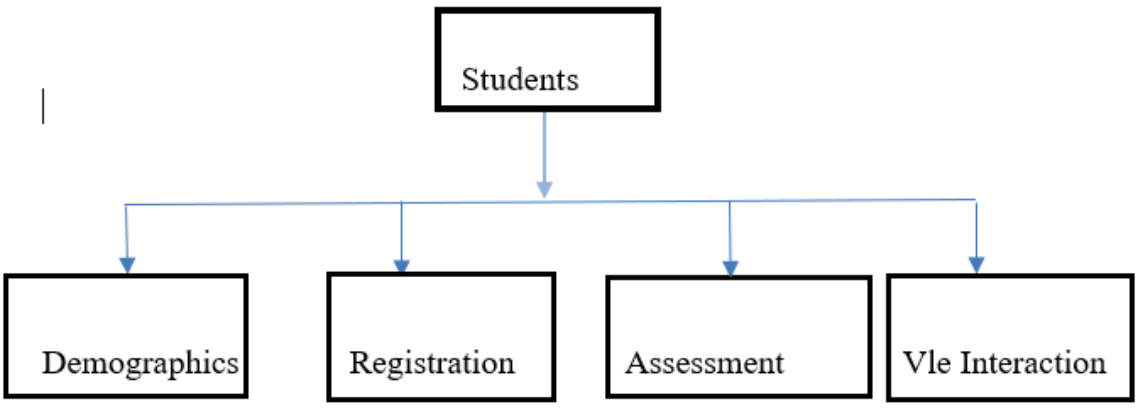

Figure 3:Structure of OULAD Dataset

3.4.3 List of Data Variables

The table below describes variables in the various tables

Table 2:List of Data Variables

Variable Name	Variable Type
Course Table	
Code_module	categorical
Code_presentation	categorical
Module_presentation_Length	numerical
Assessment Table	
Assessment_type	numerical
Date	numerical
Weight of assessment (%)	numerical
Virtual Learning Environment (Vle) Table	
Activity_type	categorical
Week-from	numerical
week-to	numerical
Student Information Table	
Gender	categorical
Age_band	numerical
Highest_education	categorical
Region	categorical

Imd_band	categorical
Disability	categorical
Number of prev attempts	numerical
Studied credit	numerical
Final_result	categorical
Student Assessment Table	
Date_submitted	numerical
Is_banked	numerical
Score	numerical
studentVle Table	
Date of interaction with learning material presented in days	numerical
Sum_clicks	numerical
Student Registration Table	
Date_registration	numerical
Date_unregistration	numerical

3.5 Data Preparation and Experiment

The LA Framework was used to guide data preprocessing, and the ML workflow was used to develop the predictive model. The study proposed an integrated ML workflow and the LA

framework to develop the predictive model. Figure 2 shows the LA framework processes used for this study. The processes of the framework are discussed below.

3.5.1 Integrated Learning Analytics and Mahine Learning Framework

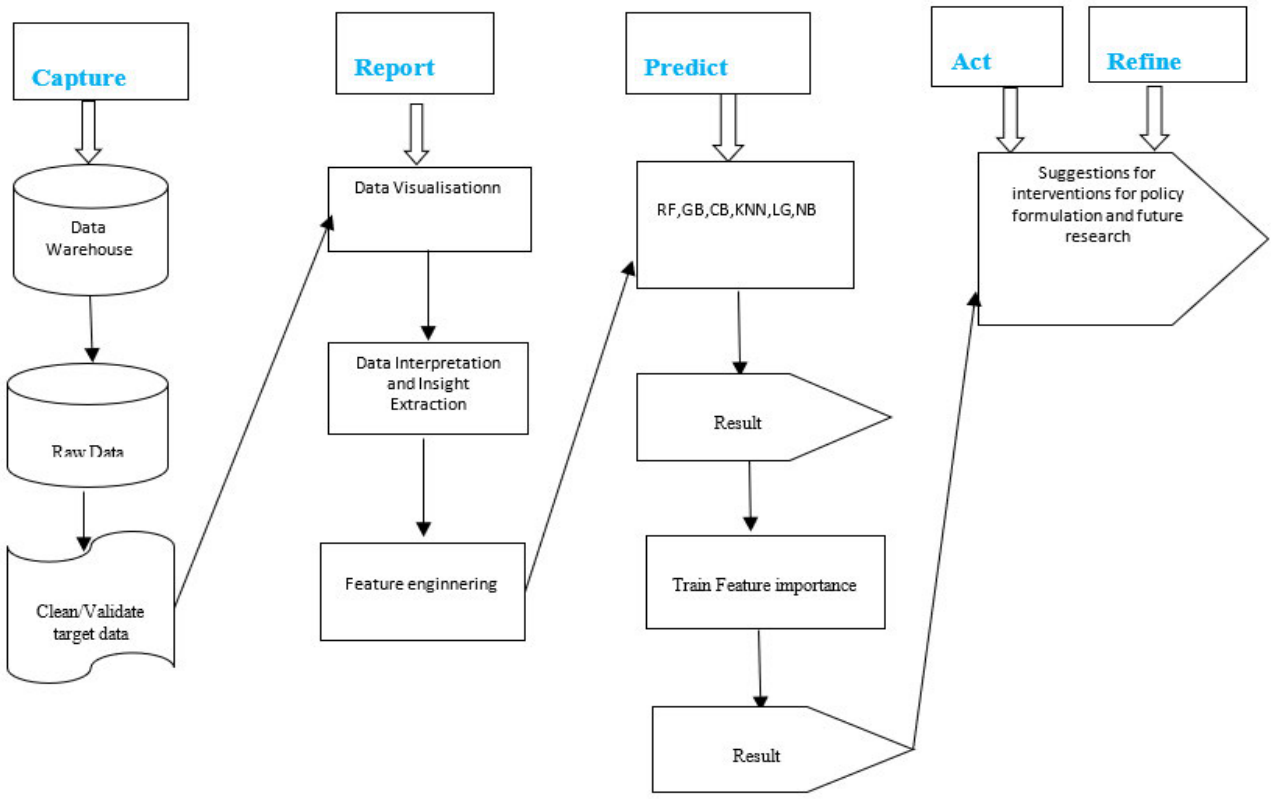

Figure 4:Intergrated Learning Analytics Framework and ML workflow (Adapted from (Campbell & Oblinger 2007))

3.5.2 Capture

The capture process comprises data collection, data organisation and data cleaning.

3.5.2.1 Data Collection

This process involved collecting data from heterogeneous sources such as the learning management system or data stored in institutional databases and student information systems. The captured data must include actor identifier, timestamp and event type (Nguyen, Wandabwa, Rasco & Le 2021; Macfadyen & Dawson 2010). Figure 3 described the data acquisition and anonymized process of the OULAD dataset.

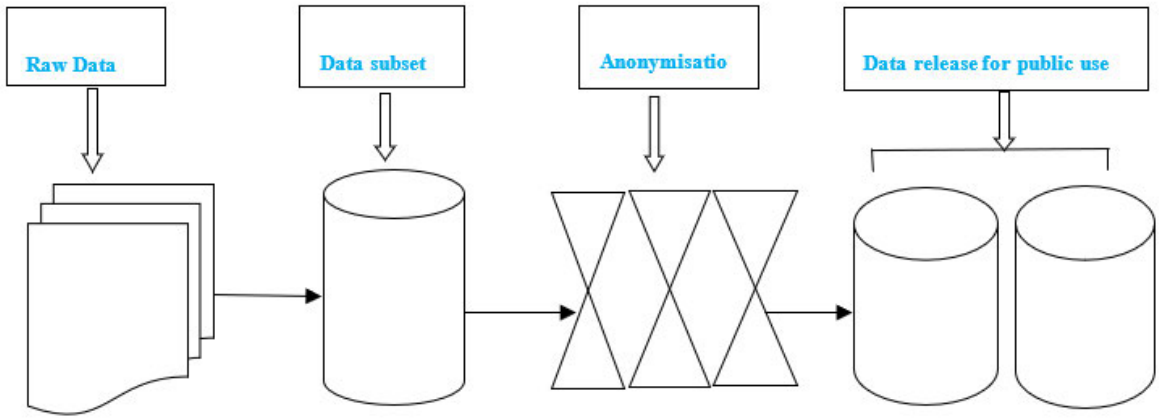

Figure 5:Data acquisition and anonymized process of the OULAD dataset (Adapted from (Kuzilek et al. 2017))

3.5.2.2 Data Organisation

The OULAD was obtained in a structured format with seven (7) data files: ‘student registration’, ‘student info’, ‘student assessment’, ‘student Vle’, ‘assessments’, ‘courses’ and ‘Vle’ as detailed in Table 1. The ‘course table’ identifies different courses and the course length. However, the dataset does not contain students offering multiple courses. Thus, the dataset represents distinct courses offered by students. The data consisted of 7 courses, of which four (4) represent Science, Technology, Engineering and Mathematics (STEM) and Three (3) represent social science modules. A summary of the module and presentations are shown in Appendix A. Each student was represented with a unique ID in the dataset. Each file was loaded into in Python 3 environment for further processing. This study intends to predict students at risk of dropping out of a module. For modelling the OULAD, the following terms were designated for classifying the

performance of students into three (3) categories. This study aims to accurately predict which students are at risk of dropping out of a module. To effectively design model using the OULAD dataset, specific terms have been assigned to classify student performance into three distinct categories.

a. Withdrawn – students who dropped out or did not complete the module.

b. Pass - students who had ‘distinction’ and ‘pass’, thus students who completed the module with good performance.

c. Fail – students who completed the module but did not pass the assessment.

3.5.2.3 Data cleaning and validity

The selected courses were already anonymized, certified and released publicly by the Open Data Institute. When data is anonymized, it is free of any identifiers that can link a specific person to the information stored. The criteria for selected courses for the Exploratory analysis were inferred from the Open University’s data criteria selection. These include the VLE data available for the module presentation (since not all the modules are studied via VLE), and the module has a significant number of withdrawn students.

The modules with the high number of withdrawn students are ‘BBB’, ‘CCC’, ‘DDD’ and ‘FFF’. All modules were filtered and explored to identify significant characteristics among withdrawn and passed students. Details of all the modules showing students' final results are presented in Figure 10.

- Merging Dataset for Exploratory Data Analysis

The dataset was first examined to establish that the course presentations were distinct and did not necessitate any prerequisites for students to partake. Table 24 shows the summary of students in each of the course modules in Appendix A. Upon observation, it was noted that 2013J and 2014J are distinct modules. Other presentations, such as 2013B and 2014B, are also present in the

dataset. Some student IDs were duplicated due to previous attempts at the module and failing. The number of attempts feature reflects either 1 or 2 attempts made by the student during these presentations.

After establishing that each students' registered for a distinct presentation under the modules, different tables were merged for the exploratory analysis. The 'studentVle' table and the 'vle' table were first merged using the 'site_id' and the 'activity_type'. The merged dataset with activities and sum click was sorted to reflect the activities with the highest 'sum_click', which indicate the activities with the highest interactions. After that, the author merged the 'studentVle_all_' (contains 'studentVle' and 'vle' tables) with the 'studentInfor' table (which contains information on withdrawn students) to create an 'overall_studInfo_df' (see fig 6) data frame. In order to keep the data frame precise for the EDA some labels were altered. The 'final result' column had labels for 'Distinction', 'Pass', 'Fail' and 'Withdrawn'. For instance, the Distinction labels were replaced with Pass labels for the purposes of the EDA using the replace function in Python. After merging the studentInfo, students, and Vle tables, the file is huge and therefore had to be divided into dates and months categories for a better visual effect. The visuals were presented for each module. To achieve this, each module was filtered out from the data frame. For example, module AAA was filtered out, and the graph was plotted on students' revision patterns to ascertain the characteristics and learning patterns of students who dropout, fail and pass. Figures 5-7 illustrate the merged tables.

```
studentVle_all_df = studentVle_df.merge(vle_df[['id_site','activity_type']], on='id_site', how ='left')
studentVle_all_df.head()
```

	code_module	code_presentation	id_student	id_site	date	sum_click	activity_type
0	AAA	2013J	28400	546652	-10	4	forumng
1	AAA	2013J	28400	546652	-10	1	forumng
2	AAA	2013J	28400	546652	-10	1	forumng
3	AAA	2013J	28400	546614	-10	11	homepage
4	AAA	2013J	28400	546714	-10	1	oucontent

Figure 6:Merged Vle dataframe

```
In [63]: #Showing the highest activities in percentages sorted

highest_all_activity = over_all_activity.sort_values('percentage', ascending= False)
highest_all_activity
```

Out[63]:

	activity_type	sum_click	percentage
9	oucontent	11206803	28.30
4	forumng	7973390	20.13
14	quiz	6981240	17.63
6	homepage	6949064	17.55
18	subpage	3411582	8.61
16	resource	1110132	2.80
11	ouwiki	894512	2.26
19	url	566702	1.43
8	oucollaborate	108974	0.28
5	glossary	87962	0.22
13	questionnaire	64764	0.16
2	externalquiz	64292	0.16
12	page	63631	0.16
0	dataplus	47468	0.12
10	ouelluminate	39028	0.10
1	dualpane	20716	0.05
7	htmlactivity	9239	0.02
3	folder	5420	0.01
15	repeatactivity	9	0.00
17	sharedsubpage	171	0.00

Figure 7:List of activity types

```
In [74]: overall_studInfo_df.head()
```

Out[74]:

	code_module	code_presentation	id_student	id_site	date	sum_click	activity_type	gender	region	studied_credits	highest_education	imd_band	age_ba
0	AAA	2013J	28400	546652	-10	4	forumng	F	Scotland	60	HE Qualification	20-30%	35-
1	AAA	2013J	28400	546652	-10	1	forumng	F	Scotland	60	HE Qualification	20-30%	35-
2	AAA	2013J	28400	546652	-10	1	forumng	F	Scotland	60	HE Qualification	20-30%	35-
3	AAA	2013J	28400	546614	-10	11	homepage	F	Scotland	60	HE Qualification	20-30%	35-
4	AAA	2013J	28400	546714	-10	1	oucontent	F	Scotland	60	HE Qualification	20-30%	35-

Figure 8: Vle merged demographic data frame

- Getting the Data ready for the Predictive model

To effectively use ML algorithms, numerical datasets generally work well. However, the OULAD dataset is a mix of categorical and numerical data, requiring a transformation to make it

suitable for the algorithm. The dataset consisted of labelled data and was best suited for implementing a supervised learning algorithm. In order to ensure accuracy and effectiveness in the predictive model, the study executed a series of data transformation procedures. These steps were crucial in preparing the data for analysis and producing reliable predictions. To effectively handle the data, the study utilized Python to load all 7 CSV files into specific data frames.

- Outlier detection

Each file was viewed using the .describe function in scikit learn to find the minimum and maximum values as well as the mean frequencies. The threshold for the score feature in the student assessment table was 0 to 100, with a score below 40 representing a fail. However, no student scored more than 100 suggesting all scores were within the threshold. On the other hand, an outlier was detected in the student assessment table with a threshold of 0-480. For example, one student recorded 630 credits and passed the module; therefore, this row was not removed since this module focused on withdrawn students.

- Transforming dates to ‘mean’ and clickstream to ‘sum’

The ‘studentVle' table records the daily click stream of students' interactions with the virtual learning environment (Vle). The total number of clicks for each student's activity type and the cumulative scores for each assessment were calculated. The dataset was organized by computing the mean date for each activity and the sum of all clicks. The merged data frame of mean dates and sum clicks is displayed in Fig 4.

- Merging data frame

Firstly, the ‘studentVle’ and ‘assessment’ tables were merged using student_id and grouped by sum_clicks and the mean dates to form ‘assessment_df’. Secondly, the ‘assessment_df’ was merged with student registration data on student ID to form ‘studReg_df’. Lastly, ‘assessment_df’, ‘studReg_df’ and ‘studentInfo_df’ data frames were merged to create ‘student_df’ data frame for developing the predictive system. Columns such as ‘disability’ and ‘region’ were dropped in the ‘studentInfor’ table before the merging. These columns were not considered in the final predictive model. The merged data frame consisted of 29741 rows and 5

columns. Figure 7 shows the merged dataframe of mean dates, sum of clicks and the student infor table.

Duplicates

The merged data frame was checked for duplicates. There were 3667 rows which were duplicated. All duplicates were removed for the next stage of the process.

Encoding

In dealing with the categorical data, the author used different functions in Python. For example, the .get dummies() function was used to encode the highest_education, imd_band,age_band and gender columns. However, this function adds extra columns to the dataset, increasing the number of columns and making it difficult to visualize the dataset. Therefore, the replacement or the ordinal method was used to convert the code_presentation and final result columns into 0's and 1's. All the changes made were concatenated with the original dataset into the variable name "studentFin_df" for the predictive model. Figure 34 in Appendix A shows the encoded features.

Missing Values

All missing values were imputed to the mean using the transformation function in scikit learn. For example, the isna() function was used to identify missing values in data_registration, date_unregistration, date_submitted and score columns and the transformation function was used to impute the mean of the value in the columns.

NaN values

Columns containing NaN values were omitted. For example, the student registration data frame contains columns such as 'week-to' and 'week-from', which consist mainly of NaN values. These two columns were not included in the merge data frame. Also, NaN values identified in

date_registration, date_unregistration, date_submitted and score were all dropped from the data frame.

Feature importance

Feature encoding is important in the design and implementation of a predictive outcome, and this is because the model gets to train on the most important features in order to achieve a good performance. In this study, the tree-based feature selection in sklearn was used to determine the important features in the dataset. This method works by calculating each feature's contribution in a decision tree-based model such as a Random Forest. The output for the feature selection was trained and tested on the random forest algorithm. The Random Forest algorithm was applied for feature importance analysis and performance evaluation. The dataset was randomly split into training and testing sets using an 80:20 ratio. The Random Forest model was trained on the training set, and feature importance was calculated using the algorithm's built-in mechanisms. Subsequently, the model was evaluated on the testing set, and accuracy, precision, recall, and F1-score were computed.

Sampling

The preprocessed dataset, such as the imputed numerical data and encoded categorical were concatenated with the original dataset into the variable “studentFin_df” for the design of the predictive model. The dataset was further split into features and labels that could fit in the ML model. The dataset was split into 80% and 20% ratios. This represents the training data of 80 % and testing data of 20%, meaning that the model gets trained on 80% of the data and tested on 20 % of the data, of which the model is unfamiliar. After splitting the dataset, the suitable model and the predictive system were considered.

Report

At this stage, the LA methods were used to analyse data according to the study's goal (to identify ARS) and to understand the general overview of the dataset. Before the predictive system was

designed, this stage explored the data to determine the factors distinguishing dropout students from graduate students (Research question 1). The following assumptions were formulated to analyse the dataset.

- Does consistency in activities such as participating in quizzes, accessing learning materials early in the semester, and submitting assignments early distinguish students who withdraw from students who pass? (results or contribute to good results and student graduating?)
- What habits/behaviours peculiar to drop-out students cause them to fail or dropout from higher institutions?
- Do successfully graduated students do anything different from those who fail or dropout?

An initial analysis of the cleaned filtered data was completed using the Python libraries such as Numpy, Pandas, Matplotlib, and Seaborn libraries in Scikit Learn. As indicated in the assumptions, the analysis was to identify characteristics of dropout students in modules 'AAA','BBB', 'CCC', 'DDD', 'EEE' and 'FFF'. This was achieved by filtering students that offered the modules from the 'overall_studInfo_df' data frame. The data was summarized through visualization and descriptive statistics to identify patterns and distribution variables. Data was presented in graphs to identify the various features and labels that could be used for the analysis. Variables were categorised using bar scatter plots, bar graphs and line graphs to represent the distribution of the student interaction and describe and report on the average date of submission of assignments and student revision patterns. Further correlation analysis was performed to identify which demographic, academic information, and activity information was related to students at risk of dropout.

3.5.2.4 Exploratory Data Analysis

Exploratory data analysis provided insight into the student dataset using graphical charts. The data exploration actions included visual techniques that examined the dataset to identify the relationships between datasets concerning students' final results. Selected variables such as student interaction with VLE and submission of quizzes and assignments in the dataset were explored to find patterns in students withdrawn from the module.

3.5.2.4.1 Final results by module

The study visualized the final result per module to understand the dropout level. To achieve this, the author filtered the specific modules from the ‘studentInfo’ table to obtain the total number of students in modules AAA, BBB, CCC, DDD, EEE, FFF and GGG. And then grouped by the final results, which are Distinction, Pass, Fail and Withdrawn. This study focuses on predicting students at risk of dropping out of higher education. Henceforth, the Withdrawn category in the final result was considered dropouts.

3.5.2.4.2 Frequency of revision and the final result

The study focused on individual modules to explore the trends and behaviour of students who dropout and those who pass. Here, the author compared student frequency of revising the learning materials and their final result to understand the characteristics of students who dropout from each module. To achieve this, the author filtered students' records based on ‘code_presentaion’ from the merged data frame (‘overall_studInfo_df’). Dates in the modules were revised from days to months to make it easy for visualisation. For example, module BBB has two (2) presentations which are 2014 B and 2013J, so the revised dates were from February to October and September to June, respectively. Each month was given a number. After that, the revised date and the number associated were sorted in increasing order and used for plotting the graphs in Figures 9-15.

3.5.3 Predict

This process involves modelling predictive outcomes based on past student data. Studies (Nguyen et al. 2021; Arnold et al. 2012) have suggested that statistical inference and ML are the most common forms of predictive analytics used to design predictive models. Literature reveals that some of the most popular ML models for predictive models include Random Forest (RF), Gradient Boosting (GB), Cat Boost (CB), logistic Regression (LG), Naïve Bayes (NB) and K-Nearest Neighbor (KNN). RF ensemble models have performed better than most individual ML models used in predictive models (Dass et al. 2021). In this regard, this session applies the

ensembled learning method, RF, as the base model in the predictive model and compares performance with other ensemble learning and classical ML algorithms such as KNN, LR, and NB. The selection of the models was based on the ML methods that are most used as well as least used for predicting dropout students in the literature. The models, therefore, were evaluated on accuracy, precision, recall, and f-score.

Withdrawn/Dropout was the primary outcome variable used in the empirical analysis. This indicator variable has a value of one if there is at least one known withdrawn occurrence about the student and zero if there are no known withdrawn occurrences. It is important to note that the model did not include "Fail" students in predicting dropouts.

This study also evaluated the accuracy of the predictive model to provide more information on the best or high-performing predictive model suitable for higher educational institutions. The model was evaluated based on accuracy, precision, sensitivity(recall), and f1-score.

The research tests were conducted on a 64-bit Windows 11 computer using Python 3. The system consists of an Intel Core i5 processor and 8GB of RAM. The ML classifiers were developed using the Scikit-Learn toolkit with the Jupyter Notebook IDE within the Anaconda Navigator GUI environment.

3.5.3.1 Machine Learning (ML) Models

This section provides different ML algorithms and their application. The RF, GB, Catboost, KNN, LR and NB are explained below.

Random Forest

This study chose the RF ensemble learning method due to its robustness to outlier data and the ability to evaluate the model for insights into the most effective discriminating variables. These factors contribute to the longer-term research goal of using these insights to target interventions at these characteristics. The RF performed efficiently in classification and regression tree problems (Dass et al. 2021). In the random forest method, a group of input coordinates known as features or variables are selected at random, and the node splits at each node of the collection of

trees. Secondly, the best split was calculated on the features in the training set. The RF tree methodology can increase in size without pruning. Pruning is a method used to address model generalization and model complexity problems. The RF method is blended with bagging to resample and perform replacements on the training dataset each time a new individual tree is grown (Biau & Scornet 2016). Results from each tree were aggregated to give a prediction for each observation. Based on the decision, the data were split between nodes, and the algorithm continued until the dropout class was predicted. An "information gain" metric was used to decide how to split the data (Equation 2). Information gain is a way to ensure that the most useful features for each step are selected. It is the anticipated decrease in entropy following the completion of sorting at a particular node. “Entropy” is a metric in information theory that quantifies the impureness or uncertainty of a set of observations (Equation 1). It determines how a decision tree splits data. Decreased entropy and increased information gain are the means through which a superior model for ML are acquired. Equations (1) and (2) yielded entropy and information gain, where pi is the probability of class i. Every decision tree within the random forest produces an output class. The dropout category is determined using the plurality voting of all outputs (Rajendran, Sinha & Chamundeswari 2021).

$$Entropy = \sum_i -\text{pi} \log 2 \text{ pi} \quad (1)$$

$$\text{Information Gain} = \text{Entropy (parent)} - \text{Weighted Average [Entropy (children)]} \quad (2)$$

Gradient Boosting

The algorithm consists of a collection of sequentially trained decision trees, with each new tree attempting to correct the errors made by its predecessors. A simple model is typically trained on the training set, such as a decision tree or a linear regression model. The residual errors between the base model's predicted and actual target values are calculated for each training example. Afterwards, a new model was fitted to predict these residual errors. Adding the new model's predictions to the previously predicted target values reduces the residual error (i.e., the difference

between the predicted and actual values). The final prediction was calculated by summing the predictions of all models.

CatBoost

Catboost is a well-known ML ensemble algorithm based on gradient boosting. It is designed to work well with high-dimensional data and has been successfully implemented in numerous applications, including image recognition, recommendation systems, and natural language processing. CatBoost constructs decision trees using a gradient-boosting technique. Each decision tree is trained using a subset of the available data and features. After training each decision tree, CatBoost employs gradient boosting to improve the model's accuracy. The algorithm begins by predicting the target variable for every training set sample. The residuals (the difference between the predicted and actual values) were then computed, and the subsequent decision tree was trained to predict the residuals. This procedure was repeated until the desired quantity of trees had been trained. CatBoost combines the predictions of all the ensemble's decision trees to generate a prediction for a new sample.

KNN

KNN is a simple, supervised ML technique that is commonly used for missing value imputation and classification or regression problems. It is predicated on the notion that the observations nearest to a given data point are the most "similar" observations in a data collection. Therefore, we can classify unanticipated points based on the values of the closest existing points. By selecting K, the user can specify the number of neighbouring observations to be incorporated into the algorithm. K is the number of neighbours to utilize. For classification, a majority vote determines a new observation's category. Larger values of K are frequently more resistant to

outliers and generate more stable decision boundaries than very tiny values (K=3 is better than K=1, which could lead to unacceptable results).

Logistic Regression

Logistic regression is a statistical technique used for binary classification tasks that involve predicting the probability of an event based on a set of input variables. It is widely utilized in many fields, including medicine, economics, social sciences, ML, and data science. The dependent variable in logistic regression is binary and has only two possible values (e.g. 0 or 1, yes or no. The aim is to establish a connection between the independent variables (also called characteristics) and the dependent variable. The model output is the predicted probability that the dependent variable is 1, which can be transformed into a binary prediction by setting a threshold.

Naïve Bayes

The Nave Bayes algorithm is a probabilistic ML algorithm that classifies data. It is based on the Bayes theorem and assumes that, given the class label, the features are conditionally independent. The algorithm is widely utilized in natural language processing and text classification. The probability of each class label is estimated using the training set. The probabilities were estimated using the number of samples in each class for binary classification. For each trait, the conditional probability of the trait given the class label was estimated. The probability was estimated based on the frequency of each class's features in the training set. The conditional probability of each class label is computed for a new sample using the Bayes theorem. A new sample was labelled with the class label with the highest probability.

3.5.3.2 Measure of Predictive Performance

The Confusion matrix gives numerous relevant measures for evaluating our classification model. The parameters considered for evaluating the MLA's outputs are Accuracy, Precision, Sensitivity (Recall), Specificity, and the F-score. Figure 15 shows the confusion matrix.

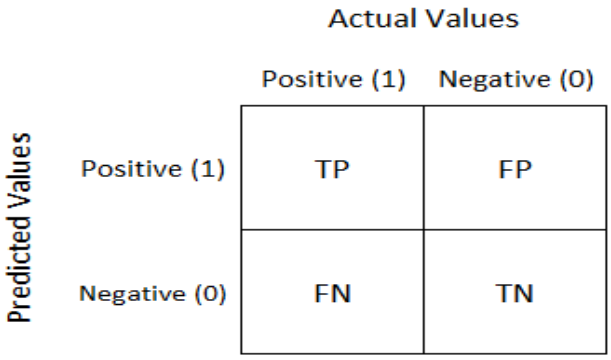

Figure 9:Confusion Matrix

Accuracy

Accuracy measures the correct classification that the model trains. However, it is insufficient as a criterion by itself to evaluate the effectiveness of a model in circumstances where class features are not distributed. The ratio of subjects correctly classified by the model to the total number of subjects were computed using the formula presented below.

Accuracy = (True Positive + True Negative)/Total Predictions (3)

Precision measures the percentage of positives predicted truly, but does not evaluate the correctly predicted negative cases. The formula provided below can be used to find the percentage predicted truly.

Precision = True Positive/(True Positive + False Positive) (4)

Sensitivity (also known as Recall) assesses how well a model predicts positives. This indicates that it evaluates true positives and false negatives (which are positives that have been incorrectly predicted as negative). Sensitivity is effective at determining how well a model predicts that something is positive. The formula below measures what percentage are predicted positive of all the positive cases:

Sensitivity = True Positive/(True Positive + False Negative) (5)

Specificity quantifies how well a model predicts negative outcomes. Similar to sensitivity but viewed from the perspective of negative outcomes. Specificity were calculated using the formular below:

Specificity = True Negative/(True Negative + False Positive) (6)

F-score

F-score represents the "harmonic mean" of precision and sensitivity. It considers both false positive and false negative cases and is useful for unbalanced datasets. This score does not take the True Negative values into account:

F-score = 2 * ((Precision * Sensitivity)/(Precision + Sensitivity)) (7)

3.5.4 Act

This stage involves the action the learner and the teacher take based on the prediction system's responses. The responses to ARS identified were presented graphically on the school administrators' dashboard to further act upon and to suggest the needed intervention for the students. This stage provides an analytical model for teachers and administrators to identify ARS earlier in their studies to suggest interventions that will help improve their performance and avoid dropouts. For future recommendations, this serves as a repository for student variable monitoring to identify the trends in student learning activities that correlate with success as well as dropouts.

3.5.5 Refine

The final stage of the methodology focused on the developmental and implementational factors in LA. After identifying the ML algorithm that results in accurate and high precision on the dataset, the algorithm was saved in the Jupyter Notebook for future implementation of similar datasets. This was used to ensure the generality of the model and build an educational data repository for ML research in higher education institutions. As this study has established the various features and the best algorithm for predicting ARS, the need for intervention to prevent students from dropping out should be the focus of future studies. This will be considered as a recommendation for future studies.

3.6 Summary

This chapter discussed the detailed methodology for developing an ensemble-based predictive system to identify students at risk of dropping out of higher education. The chapter elaborated on the choice of the dataset obtained for the study. It provided steps for cleaning and modelling the predictive system using the Scikit learn library in Python. The five (5) steps of the Learning Analytic process, described as capture, report, predict, act, and refine, were adopted. The capture step was utilized for obtaining and cleaning the dataset.

In contrast, the reported step was used to visualize the data with Scikit Learn, Matplotlib and Seaborn libraries in Python 3 environment. This was followed by the predictive step, which explained how ML workflow was integrated into the LA framework to design a predictive model with the RF, GB, CB, KNN, LR and NB. The Act and Refine stages discussed the implementation and feedback process of the model. The next chapter presents the result of the Exploratory Data analysis to determine the characteristics of graduating and non-graduating students and, subsequently, the predictive model development and evaluation performance results.

4 Chapter Four – Data Analysis and Result

4.1 Introduction

This chapter presents Data Analyses for the research questions using the OULAD dataset. The presentation was organised in the following order. Firstly, the Dataset schema and variables were presented; secondly, an exploratory analysis of each module and final results. Thirdly, the analysis of the dataset based on module presentations to illustrate the characteristics of graduating students against non-graduating students. Lastly, the results and evaluation of a predictive model of ARS using the "RF" ensemble classifier, "GB", "CB", "KNN", "LR", and "NB" algorithms and evaluating their performance.

4.2 Variables in the Dataset

This section described the dataset after loading the individual tables in the Python environment. Each table was viewed to identify the dataset's variables and features. Some variables were extracted and plotted in graphs to make the dataset meaningful. The tables and structure of data frames are listed in Appendix A.

4.2.1.1 StudentInfo Table

This table consists of student demographics and their final results. The dataset has 32593 rows and 12 columns. They are shown in Table 16 in Appendix A.

4.2.1.2 studentRegistration Table

This table consists of students' registration details. The dataset contains 32593 rows and 5 columns. Shown in Table 17 in Appendix A

4.2.1.3 studentVle Table

This table consists of student interaction with the Virtual learning environment represented as clickstreams. The dataset consists of 10655280 rows and six (6) columns. Table 18 in Appendix A.

4.2.1.4 studentAssessment Table

This table consists of student assessments in the various modules. The dataset has 173912 rows and five (5) columns. Table 19 in Appendix A.

4.2.1.5 Assessments Table

This table consists of assessment type and dates of submission of students. The dataset is made up of 206 rows and six (6) columns. Table 20 in Appendix A.

4.2.1.6 Courses Table

This table consists of all seven (7) courses, including four (4) Stem courses and three (3) social science courses represented as AAA to GGG. The dataset has 22 rows and 33 columns. Table 21 in Appendix A.

4.2.1.7 Vle

This table consists of the various activities that students engage in when accessing the Virtual learning environment (VLE). This dataset contains 6364 rows and six (6) columns. Table 22 in Appendix A.

4.3 Exploratory Data Analysis Result

The following charts represent the result of the EDA. This section utilised bar graphs to present the students' final results in each module.

4.3.1 Final result rates by module

The following bar graphs illustrate the student final results in the seven (7) modules with dropout students highlighted.

Final result by module AAA

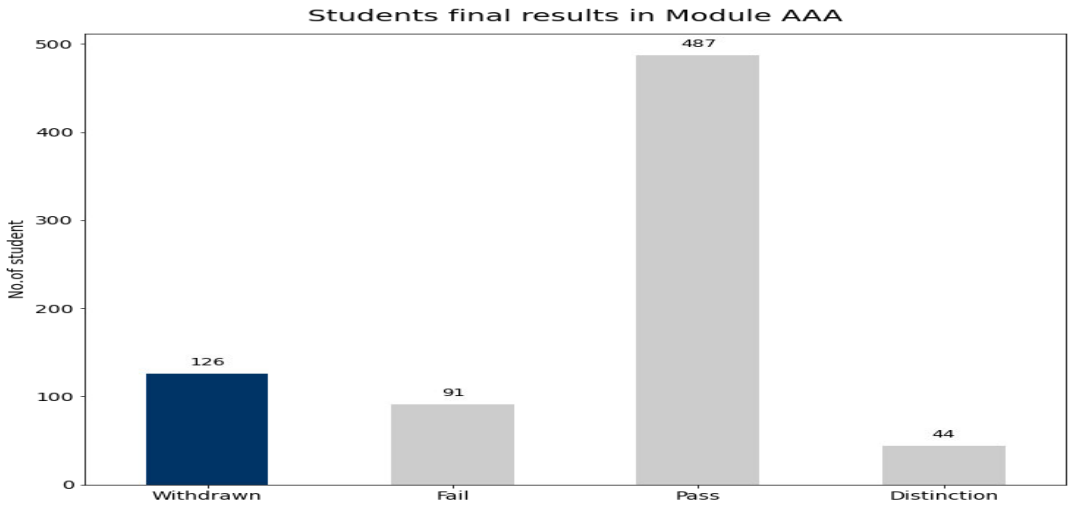

Figure 10:Final result by module AAA

Fig 9 focused on student offering module AAA. The total students for the module was 748. 126 students out of 748 students dropped from this module. However, this module had an impressive pass performance of 487.

Final result by module BBB

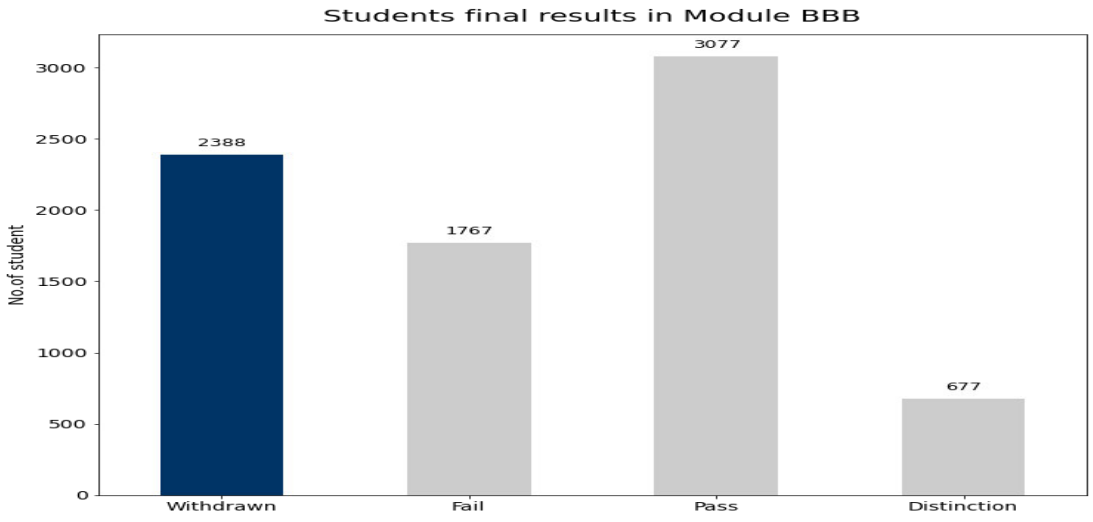

Figure 11:Final result by module BBB

Fig 10 shows the total number of students in module BBB was 7909. The final student result in Module BBB shows a more significant population of the student passing than failing. It can be observed that the total number of students that withdrew (2388) was more than those who failed (1767) but lesser than the passed (3077) students.

Final result by module CCC

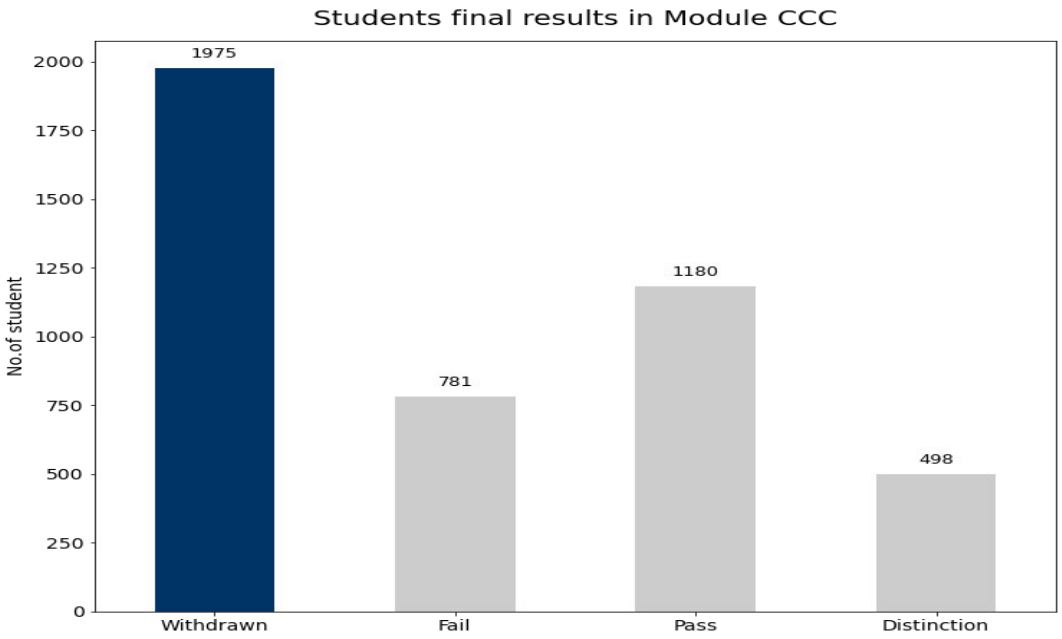

Figure 12:Final result by module CCC

Fig 11 shows the total number of students in module CCC was 4434. Focusing on course CCC, it can be observed that 1975 students dropped out from the module whilst 498 and 1180 had distinction and pass, respectively.

Final result by module DDD

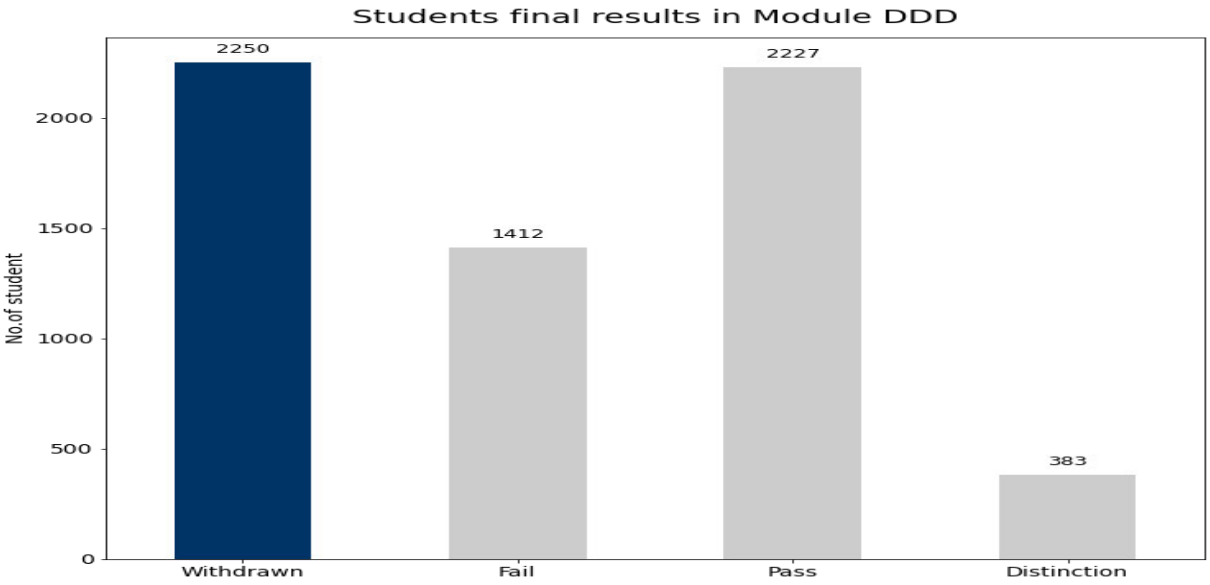

Figure 13:Final result by module DDD

Fig 12 shows the total number of students in module DDD was 6272, of which 2250 withdrew while 383 had distinction. This module recorded higher pass results of 2227 and a failure result 1412.

Final result by module EEE

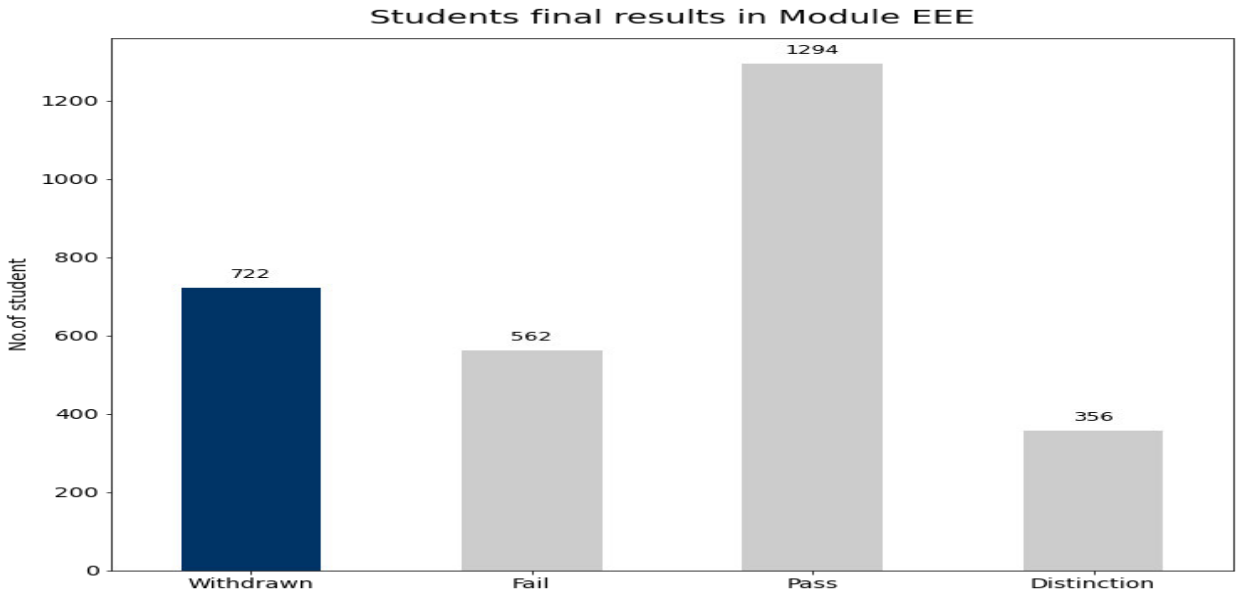

Figure 14:Final result by module EEE

Fig 13 shows the total number of students in module EEE was 2934. There were 722 students labelled as withdrawn, while 1294 and 356 students completed the module with Pass and Distinction, respectively.

Final result by module FFF

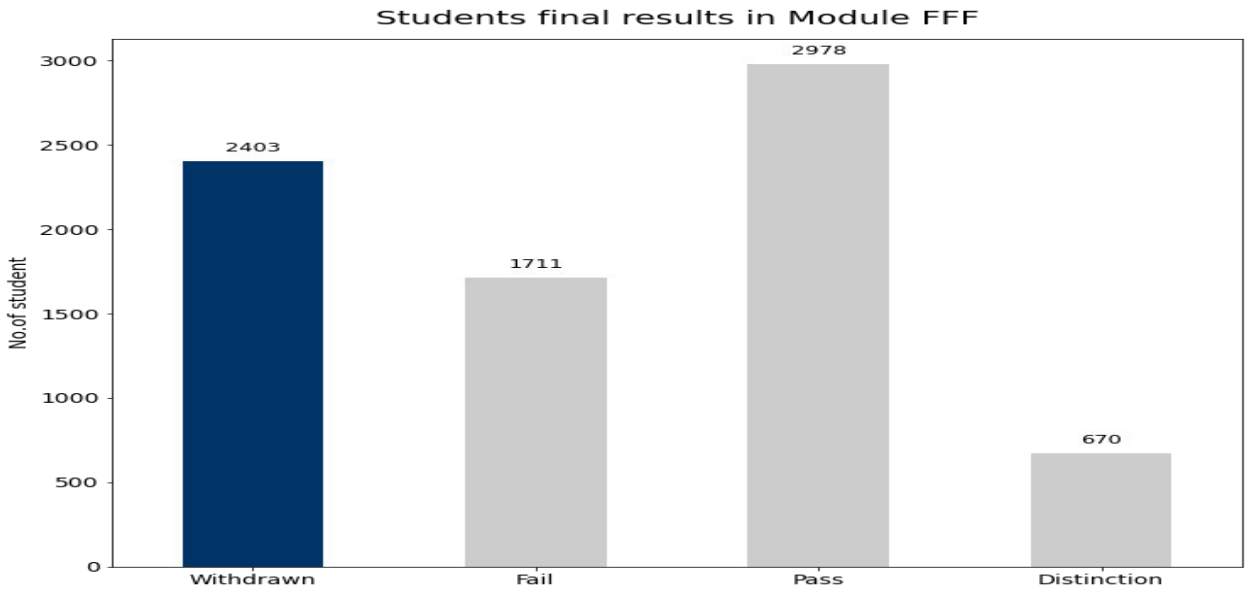

Figure 15:Final result by module FFF

The total number of students in module FFF was 7762. It was observed that 2403 students dropped out from the module while 670 had distinction. This module also had a higher number of students passing.

Final Result by module GGG

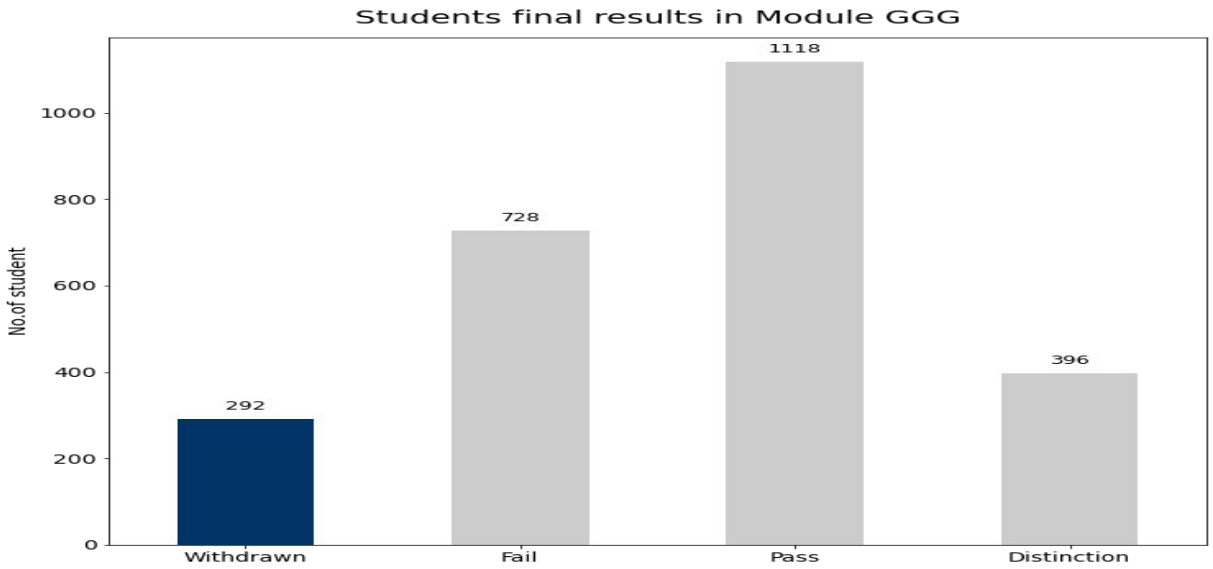

Figure 16:Final result by module GGG

Fig 15 shows the total number of students in module GGG was 2534. This module recorded more pass results of 1118. Students with distinction results were 396, whiles 728 students failed. The module recorded 292 withdrawn students.

4.3.1.1 Summary of all modules and Final result

Figure 16 illustrates a homogenous presentation of the code modules and the final result.

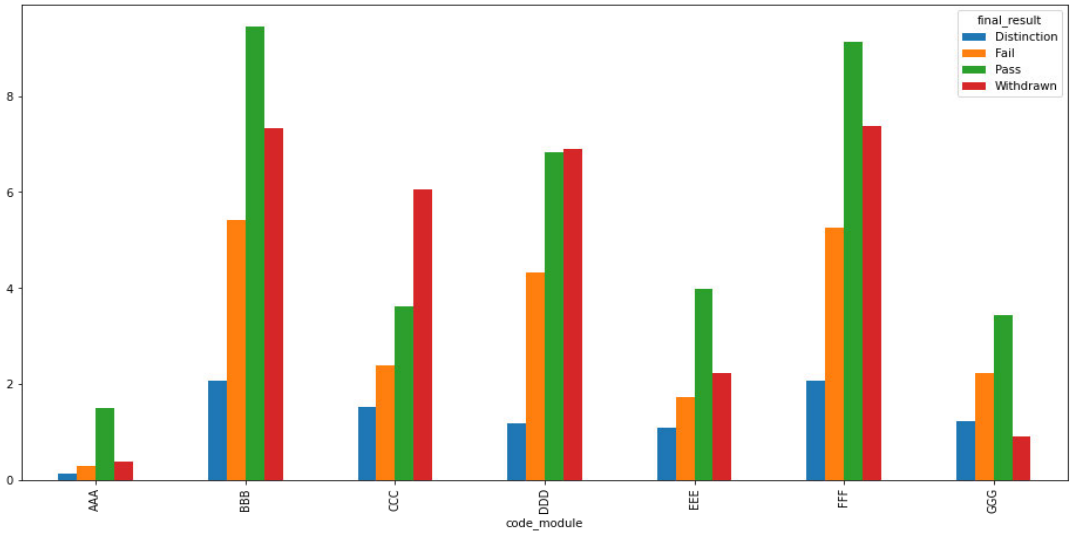

Figure 17:Final result by all modules

As indicated in Figure 16, modules BBB and FFF recorded the highest pass students among the modules, followed by DDD and CCC. Module AAA enrolled fewer students (748) than the other modules, whereas module BBB had the highest student enrollment of 7909. Also, modules BBB, DDD and FFF have the highest number of pass students, whereas modules AAA, CCC, EEE and GGG projected fewer pass students.

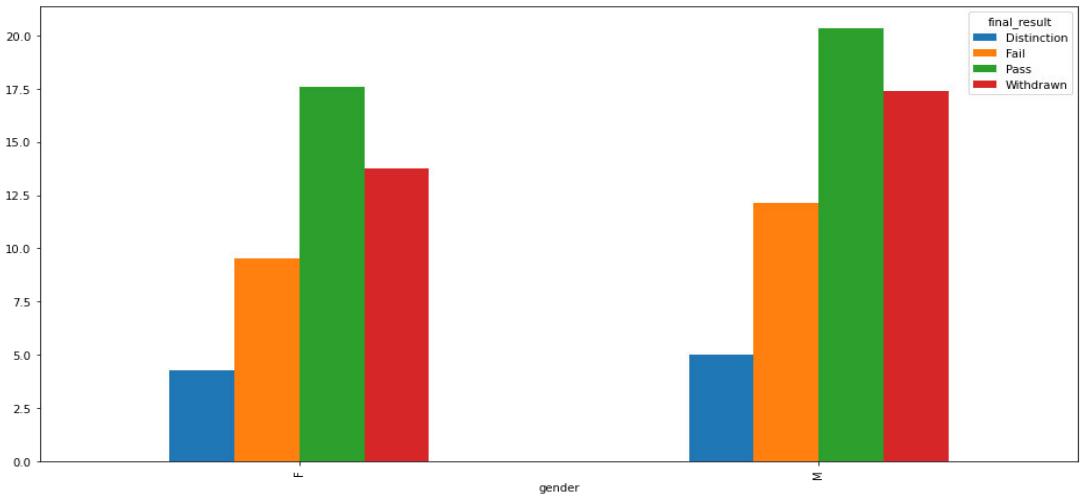

Figure 18:Final Result by Gender

Fig. 17 presents the final results of students in the modules based on gender. It can be observed that males recorded higher pass records as well as high records for withdrawal.

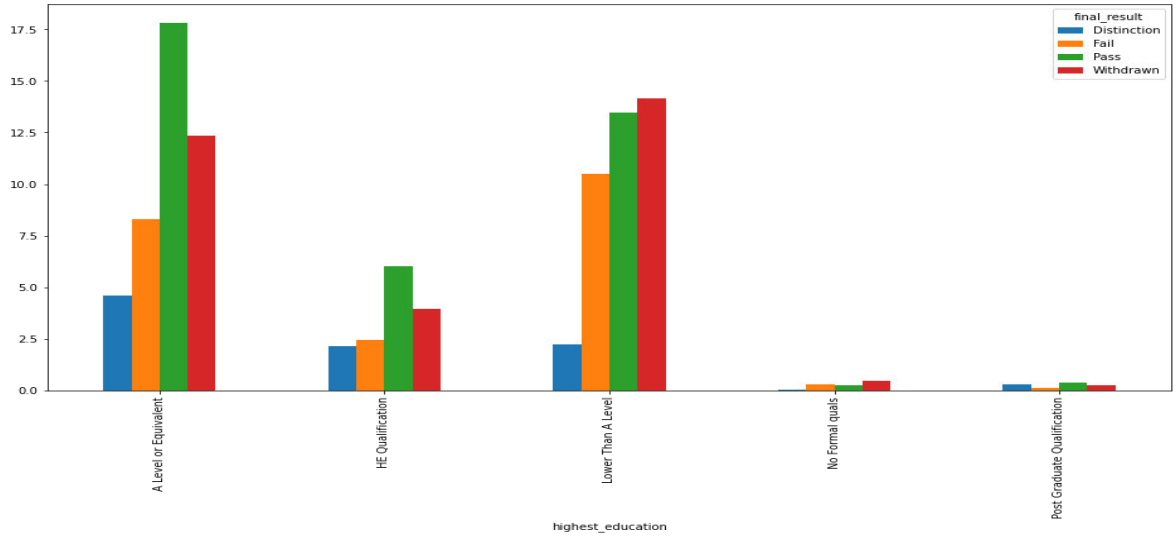

Figure 19:Result by Highest Education

Fig. 18 illustrates the correlation between the highest education and final results. It can be observed that students who had an A level or equivalent in previous education excel more than qualifications lower than an A level qualification.

Summary of Exploratory Analysis

Table 3:Summary of Dataset

ITEMS	Total number	Representation
Module/Course	7	
Presentations	22	3.14
Total num of registration	32593	
Total num of student	28785	1.13
vle interactions	10655280	
vle	6364	1674.31

4.4 Results for Research Question One

The first research question analysed the most prominent characteristics or behaviours that distinguish dropout students from graduated students related to interactions with the virtual learning environment (vle). The student activities and online logs were used to demonstrate the distinctions between students who graduated and those who dropped out of module. RQ 1 explores how frequently students interact with the VLE and revision trends before the examination and the effect on the final result. The following assumptions were formulated to explore the data to determine the factors distinguishing dropout students from graduate students.

- Does consistency in activities such as participating in quizzes, accessing learning materials early in the semester, and submitting assignments early, contribute to students graduating?
- What habits/behaviours are peculiar to drop-out students and cause them to dropout of higher institutions?

- Do students who graduate do anything different from students who dropout?

To answer RQ1, the most interactive activities student access virtually and the associated "sumclicks" percentages were retrieved and plotted into a bar graph visual presentation and explanations.

4.4.1 Most interactive activity type

Table 10 shows the various student activity types and the "sumclicks" records of their interaction with the virtual environment in ascending order. The result shows that students' most indulgent activity is 'oucontent', with a "sumclicks" of 11206803 representing a significant 28.3% of the total "sumclicks". Table 10 illustrates the sorted virtual learning environment activities percentages in ascending order.

Table 4:Most interacted activities

activity_type	sum_click	percentage	activity_type	sum_click	percentage
oucontent	11206803	28.3	questionnaire	64764	0.16
forumng	7973390	20.13	externalquiz	64292	0.16
quiz	6981240	17.63	page	63631	0.16
homepage	6949064	17.55	dataplus	47468	0.12
subpage	3411582	8.61	ouelluminate	39028	0.1
resource	1110132	2.8	dualpane	20716	0.05
ouwiki	894512	2.26	htmlactivity	9239	0.02
url	566702	1.43	folder	5420	0.01
oucollaborate	108974	0.28	repeatactivity	9	0
glossary	87962	0.22	sharedsubpage	171	0

This provide insight into the most used activities in the dataset.

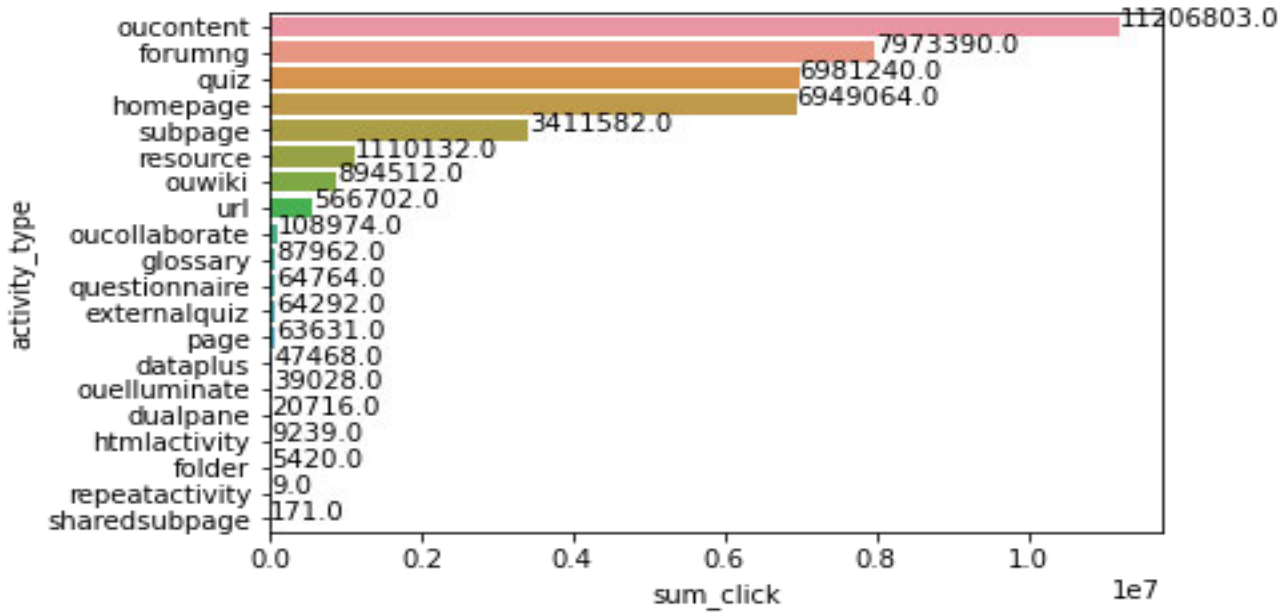

Figure 20:Most interactive activities by sum clicks (Own work adapted from Tahiru and Parbanath 2023)

Fig.19, clearly shows that generally, most students utilized the 'oucontent' as it garnered the highest sum clicks at 28.3%. The assumption is that the 'oucontent' is the learning materials of the modules that are uploaded online during the presentation of the module. The second most engaged activity was the 'forumng', with 20.033% suggesting that lecturers post questions of interest to the forum and students engage in posting contributions to the forum. Hence, more clicks were generated for the 'forumng' activity. The next activity was 'Quiz' with 17.63%, suggesting students took part in quizzes and other forms of assignments associated with the module. 'Homepage' activity was followed by 17.55 %, suggesting that some students only visited the home page of the learning environment without accessing the content. This is possible because 'homepage' is the primary page for students to access when engaging with the virtual learning environment. Figure 13 presents the overall activities and student sum clicks in the virtual learning environment.

4.4.2 Frequency of revision and the final result

The results on the frequency of revision to distinguish graduated students from dropout students are presented below. In order to achieve this, the code presentations 2013B, 2013J, 2014B and

2014 J were retrieved from the dataset and plotted against the average ‘sumclicks’ with the duration of the code module, which was represented in months. The plots were generated with the Seaborn and Matplolib.

Fig.20 indicates that prior to the start date of presentation 2013B, both graduated and non-graduated students accessed the presentation's course materials. After that, the activity of Graduated students on the VLE declined, then began to increase in June, then declined again in August and September, and finally ended in October. The increase in activity in June indicates the review of course material by students prior to the examination period at the conclusion of the presentation. For non-graduated students, the graph indicated a decline in activity between May, after accessing the course materials and October, when the presentation concluded. After accessing the course material at the beginning of the presentation, there were limited activities for non-graduated students in presentation 2013B.

4.4.2.1 Frequency of revision and the final result for 2013B, 2014B

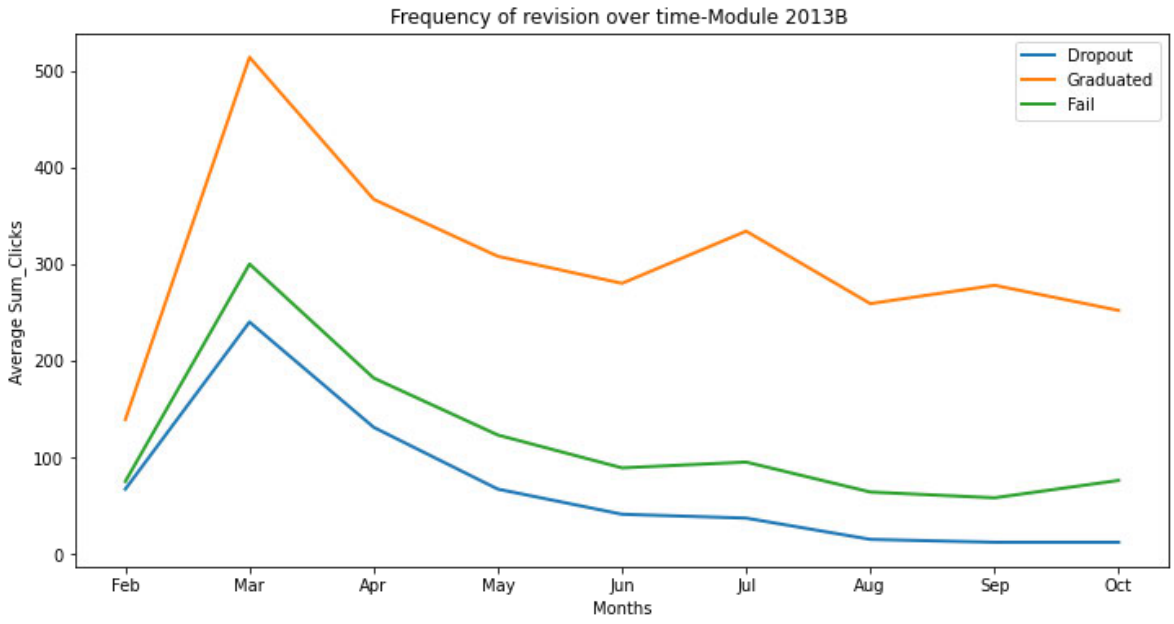

Figure 21:Revision trend for 2013B (Own work adapted from Tahiru and Parbana 2023)

Figure 21 illustrates the trend of students accessing and revising course materials in presentation 2014B. Access to course materials follows the same pattern as depicted in Fig. 20 for presentation 2013B, with the exception of a slight variation in the revision pattern observed among non-graduated students. This suggests that both non-graduated and graduated students accessed course materials and reviewed them prior to the exam; however, non-graduated students' performance did not improve. Access to course materials and revision, of course, materials had less impact on the final grades of non-graduated students for presentation 2014B.

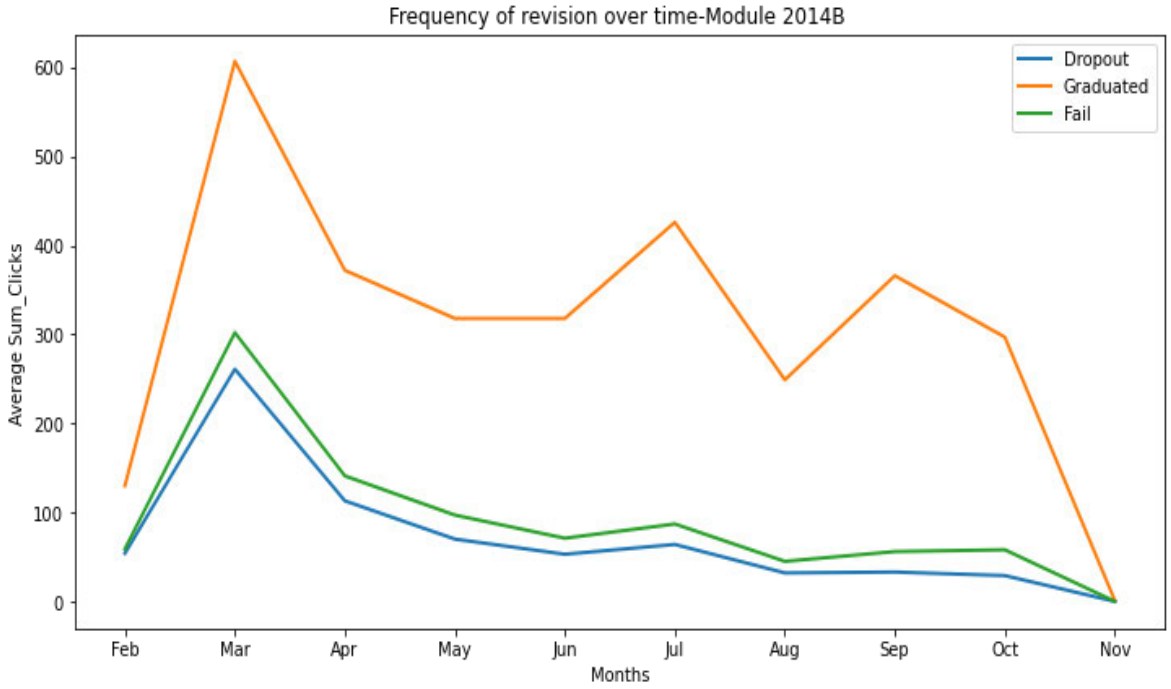

Figure 22:Revision trend for 2014B (Own work adapted from Tahiru and Parbanath 2023)

Using the total number of clicks of students enrolled in the 2013B and 2014B presentations, Fig.22 and Fig.23 illustrate students' engagement with VLE. According to the data, more graduated students participated in VLE activities than non-graduated students. This indicates that ‘graduated’ students engaged in online activities more frequently than those who did not. Nonetheless, a few ‘graduated’ students who did not participate in many activities were able to graduate. These could also depend on the types of activities engaged in by graduated students.

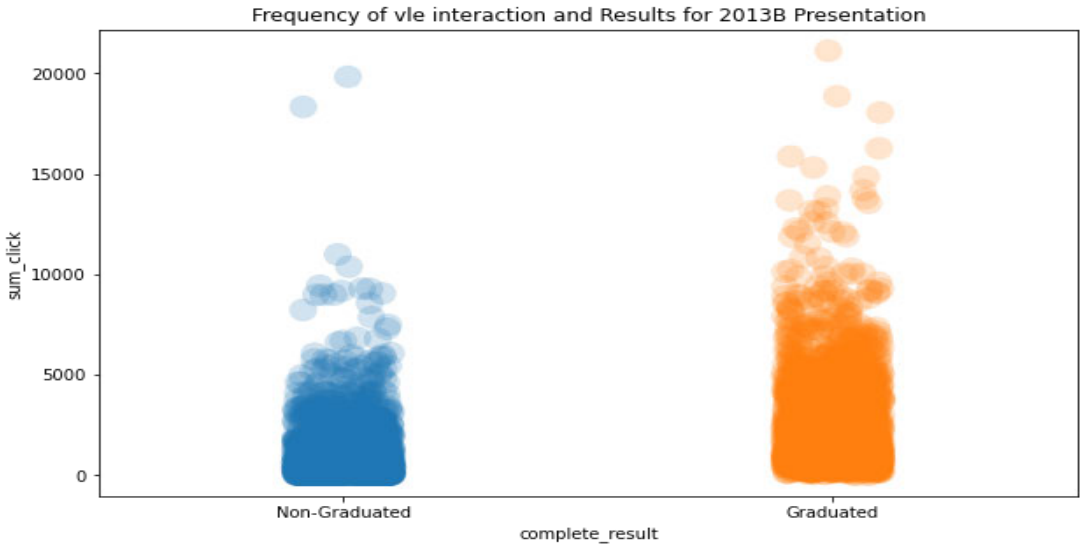

Figure 23:Frequency of revision in 2013B (Own work adapted from Tahiru and Parbanath 2023)

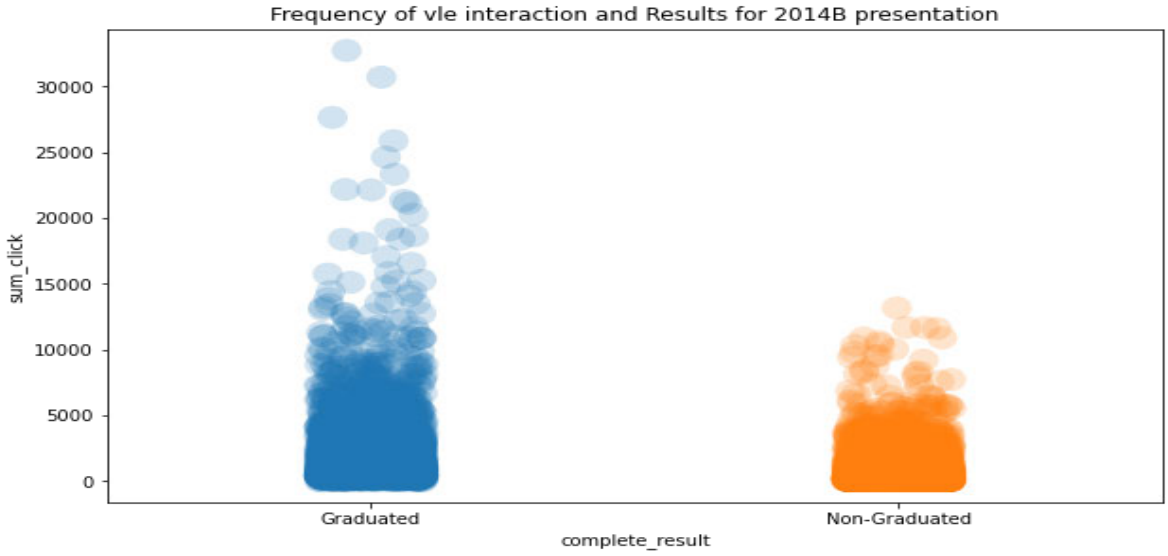

Figure 24:Final result by sumclicks 2014B (Own work adapted from Tahiru and Parbanath 2023)

Fig. 24 illustrates that students in both categories (graduated and non-graduated) accessed course materials similarly at the start of the 2013J presentation. In this instance, learning materials were accessed prior to the start of the presentation. From December to April, when the final

examination of the presentation was due, the revision pattern for graduates remained consistent. The non-graduated plot is flat, indicating less preparation for the final examination.

4.4.2.2 Frequency of revision and the final result for 2013J, 2014J

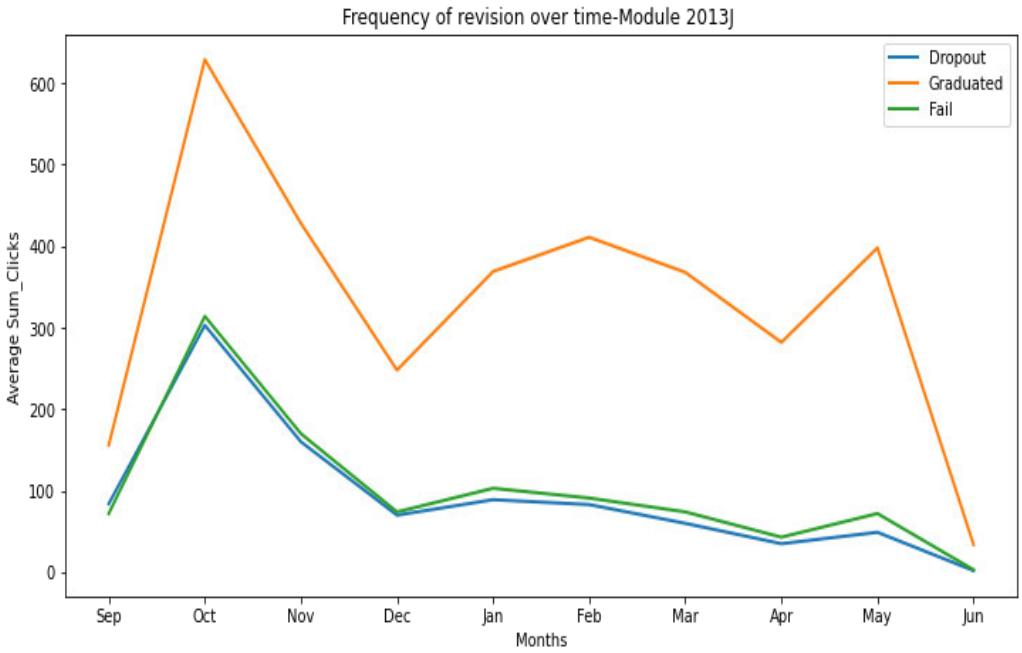

Figure 25:Revision trend for 2013J (Own work adapted from Tahiru and Parbanath 2023)

Fig 25 depicts access to course materials and revision trends by ‘graduated’ and ‘non-graduated’ students in presentation 2014J. Accessing course materials and revising for the final exam follows a similar pattern for both graduated and non-graduated students, as depicted in Figure 25. This indicates that students in both groups can access course materials and review them before final exams. However, the final outcomes did not reflect the efforts of the students who did not graduate.

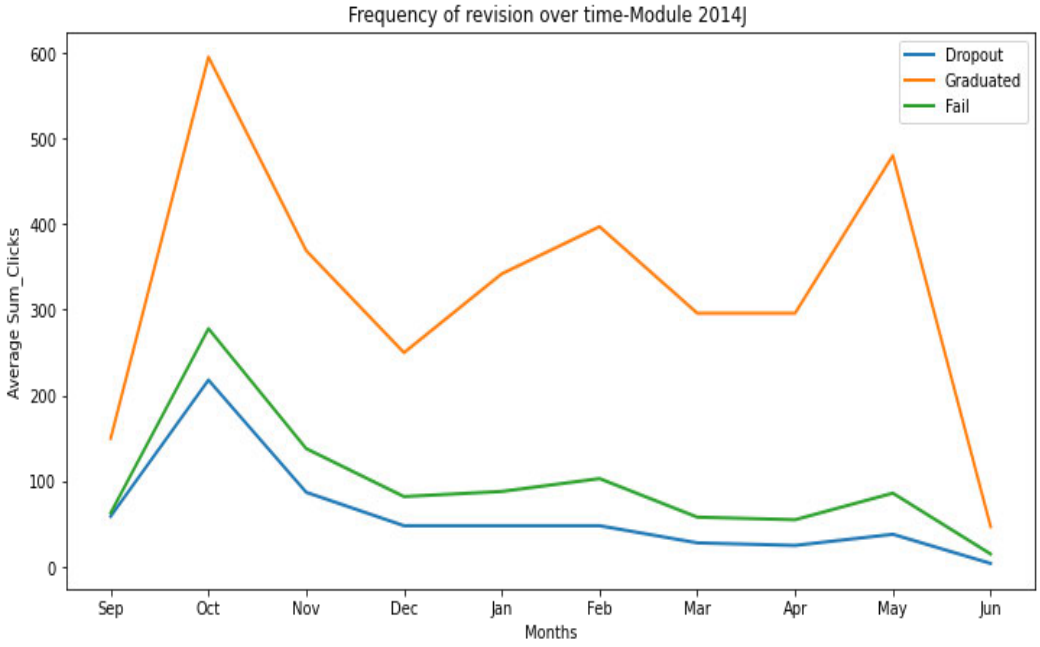

Figure 26:Revision trend for 2014J (Own work adapted from Tahiru and Parbanath 2023)

Figures 26 and 27 depict the ‘sum-of-clicks’ indicating student engagement with the VLE for students enrolled in 2013J and 2014J presentations, respectively. According to the data, more graduates participated in VLE activities than non-graduates. This indicates that students who graduate engage in online activities more frequently than those who do not. Nonetheless, a few graduate students who did not participate in many activities were able to graduate. These could also depend on the types of activities engaged in by graduates. The percentage of non-graduates in presentations 2014J is significantly lower than in 2013J.

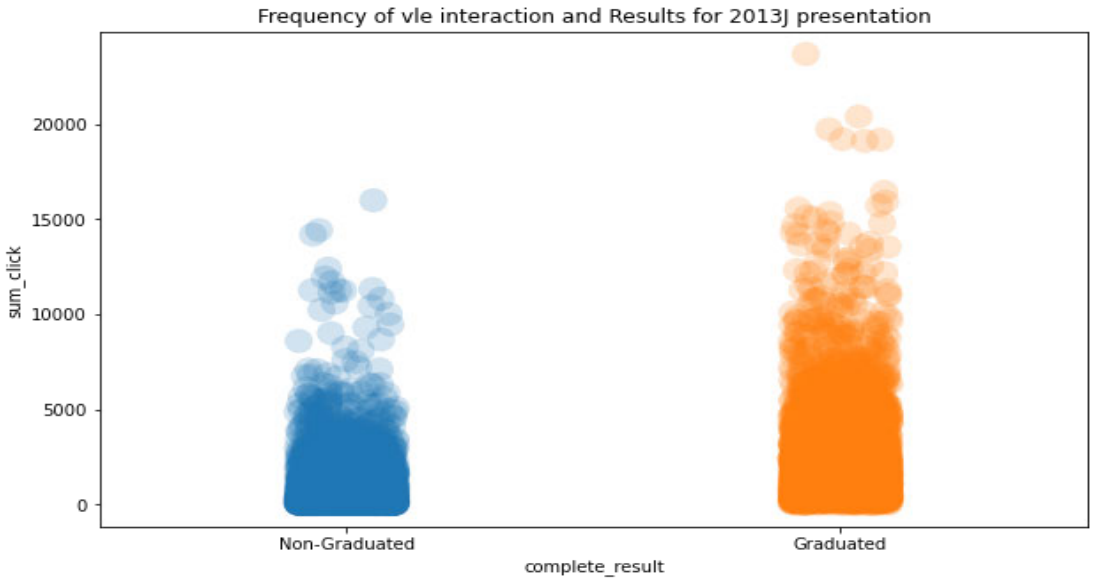

Figure 27:Frequency of revision in 2013J (Own work adapted from Tahiru and Parbanath 2023)

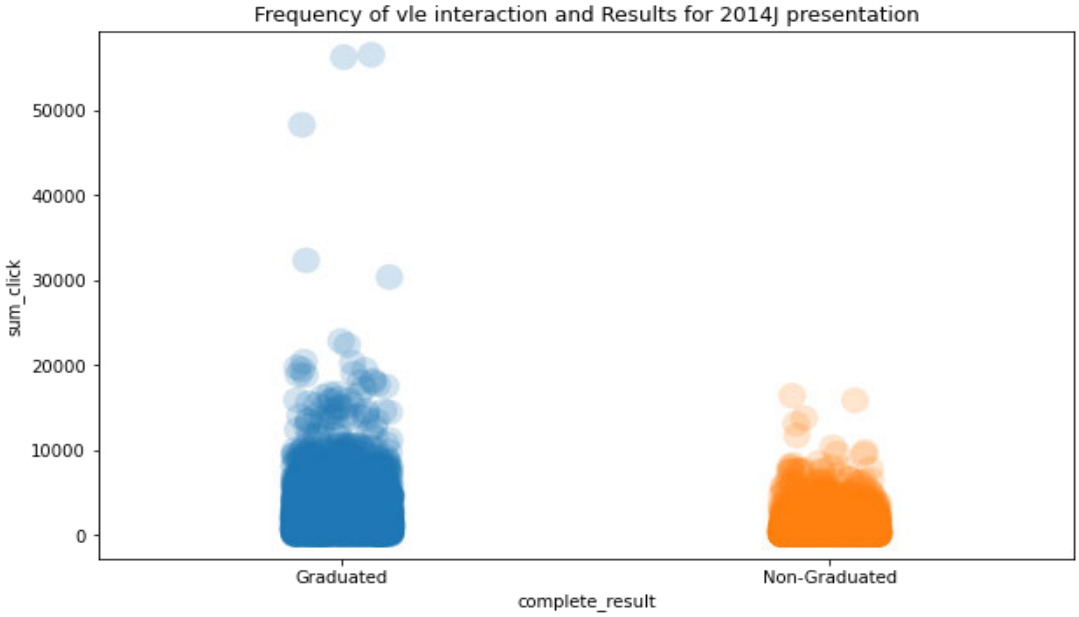

Figure 28:Final result by sumclicks 2014J (Own work adapted from Tahiru and Parbanat 2023)

Figure 28 shows that non-graduates made up only 20% of the presentation's ‘sum-clicks’. This means failed students recorded 13.03% of interactions with the VLE, whiles the Withdrawn students interacted with the VLE at 6.56%. Students who received a distinction had 46.3% of the

‘sum-clicks’, while those who received a pass had 34.28 clicks. Students who completed the module appear to be more engaged with the VLE's resources. Educators' primary concern is identifying the causes that prevent students from interacting with VLE and, by extension, not passing or graduating from a presentation, even though it is evident that with less interaction with course materials, students would not be able to pass or graduate in a cause.

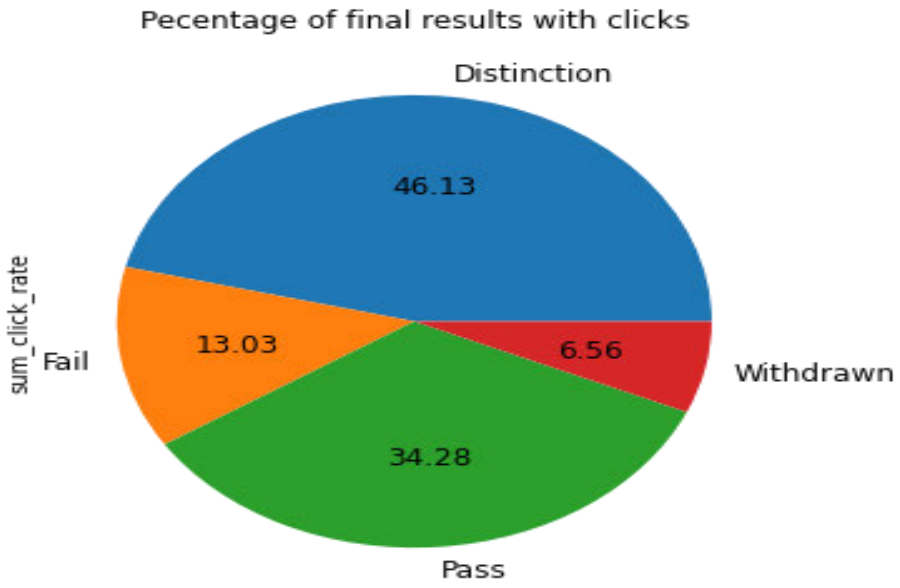

Figure 29:Percentage of final results with clicks (Own work adapted from Tahiru and Parbanath 2023)

As indicated in Figures 20, 21, 24, and 25, there was an increase in activities at the beginning of the module, ie. From September to November for all the classes (Pass, Fail and Withdrawn). April and May saw another high interaction between Pass and Fail students. It can be observed that there was a decline in student activities in June and November, which implies that the modules for 2013 and 2014 have been successfully completed. Figure 28 shows the percentages of ‘sum-clicks’ for each of the final results. It was observed that withdrawn students recorded the lowest ‘sum-clicks’, and distinction students recorded the highest ‘sum-clicks’ of 46.113%,

confirming that students who engage more with the Vle perform better than those who engage less with the Vle.

4.4.3 Late submission of Assignments and quizzes and final result

It was observed that 878 submissions out of a total submission of 171047 represented assessments in the form of quizzes and assignments late. The study observed that students who did not graduate from the module recorded 0.33% of the total late submissions. This could also be related to the characteristics that non-graduate students exhibit in educational institutions. This was achieved by merging the student assessment and assessment tables and retrieving date_submitted greater than the assessment due date. The late-submitted date counts were divided by the total submission count of withdrawn students and multiple by 100. Fig 40 shows the total number of late submissions and the final result.

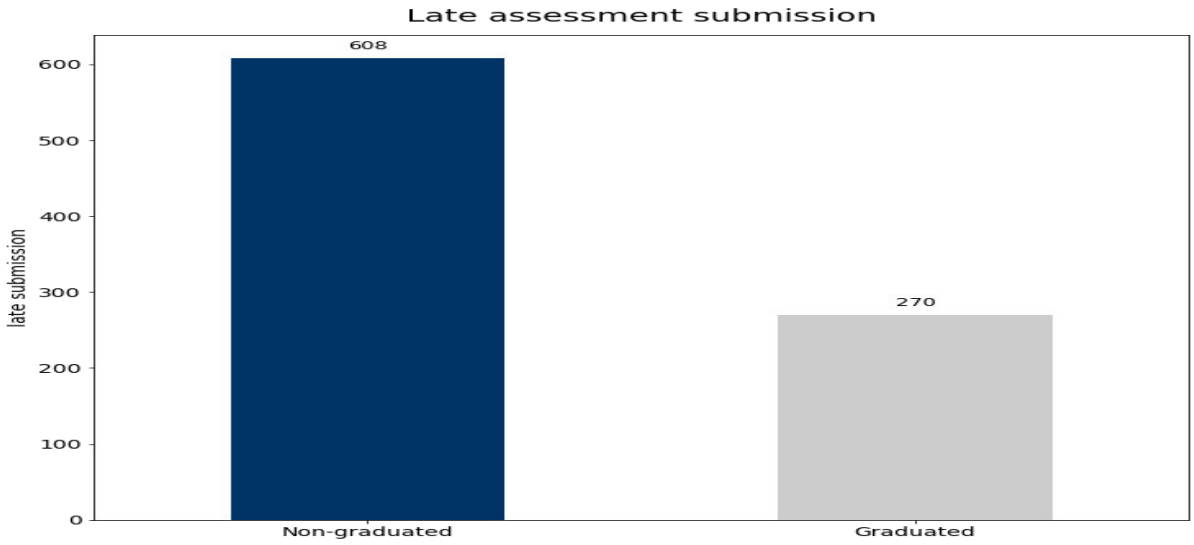

Figure 30:Late submission for graduate and non-graduated students (own work)

Summary of Assumptions

Characteristics that distinguish graduating students from students who dropout.

1. Graduating students recorded higher sum clicks meaning they engaged more with the learning management system.
2. Graduated students revised their learning materials more and earlier than dropout students.
3. Graduating students submit assignments early before the submission deadline.

4.5 Results for Research Question Two

Which emerging machine learning model can be adapted to predict student dropout in higher educational institutions?

To answer the second research question, different ML models were investigated through a SLR to obtain insight into the performance of algorithms that have been utilized successfully in the design of predictive models. The result of the systematic literature review indicated that the most common classifiers employed in the literature for developing predictive systems in education include “SVM”, “RF”, “KNN”, “ANN”, and “LR”. Besides the SLR results, other ensemble learning algorithms discussed in the literature, such as “GB” and “CatBoost” algorithms, were also employed. Having identified different ML algorithms; the current study implemented the algorithms through a design model to determine the good performance model. In this study, the three ensemble learning algorithms (“RF”, “GB”, and “CatBoost”) and three traditional learning algorithms (“KNN”, “LR”, and “Naïve Bayes”) were developed and compared. Each algorithm trained the model, and the predictive performance was evaluated on “accuracy”, “precision”, “recall”, and “F1-score”.

The Scikit learn library in Python was employed to build and compare the following models:

a. “Random Forest” (RF)
b. “Gradient Boosting” (GB)
c. “CatBoost”(CB)
d. “K-Nearest Neighbor” (KNN)
e. “Logistic Regression” (LG)

f. “Naïve Bayes” (NB)

The above ML models were measured and compared using accuracy, sensitivity(recall), specificity, precision, and f1-score metrics. The dataset was loaded to the Python 3 environments using the ‘.read_csv’ command. The Scikit learn in Python has the capability to run all the ML models specified.

4.6 Predictive Model Results

This section presents the model's results developed for the six (algorithms). The target dataset was presented, followed by the metrics result for each model and the confusion matrix of the model performance.

4.6.1 Comparing Ensemble learning algorithms to traditional machine learning algorithms

To address RQ 2, ML models were designed and compared to determine how well they predict the drop-out of students. The total dataset of 17488 was used to perform this analysis; the input features are shown in Figure 35. The distribution among the classes is shown in Table 11. The ‘1.0’ represents the ‘Withdrawn’ class, with 5811 values, which is what the model is to predict, while the ‘0.0’ class represents the Pass class with a value of 11677. The results show that all the ensemble learners outperformed the traditional ML models. As shown in Table 12, the Random Forest classifier and the Gradient Boosting algorithms had an accuracy of 0.93%, respectively. Whiles the CatBoost had 0.92 accuracies. (Though the ensemble learners outperformed the traditional ML models, the LG and NB performed better among the traditional ML models with 0.91% accuracy, respectively) The KNN had the lowest accuracy of 0.73. Table 12 shows the performance of the classifiers.

Table 5:Values of the target dataset

Target dataset value	Representation	Value	Percentage
Withdrawn	1.0	5811	0.33
Pass	0.0	11677	0.67

Table 6:Result of analysis

	Accuracy	Precision	Sensitivity (Recall)	F1- score
RF	0.93	0.94	0.82	0.88
GB	0.93	0.92	0.83	0.88
CatBoost	0.92	0.93	0.82	0.87
LG	0.91	0.92	0.80	0.85
NB	0.91	0.90	0.83	0.86
KNN	0.73	0.68	0.34	0.45

The KNN, LG and NB algorithms recorded low accuracy performance as compared to the ensemble learning algorithms. However, the KNN recorded the lowest accuracy of 73 % when compared with the other traditional ML algorithms as well as the ensemble learning algorithms.

Confusion Matrix of Analysis

The confusion matrix illustrates the performance of each algorithm on how correctly the model predicted the positive and negative classes.

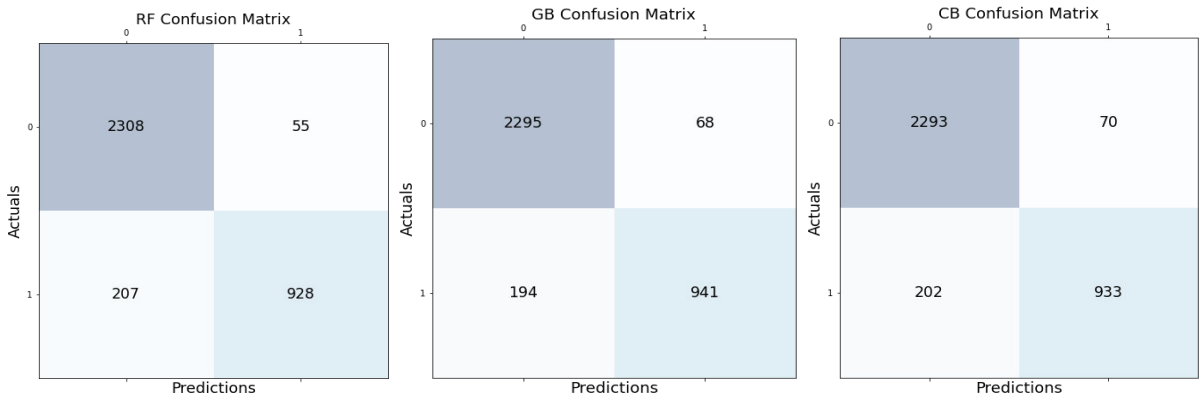

Figure 31:Confusion Matrix for ensemble learning models

The RF Confusion matrix shows a TP of 2308, meaning the model predicted positive values of 2308, which were positive, and TN of 928, meaning 928 predicted values were negative, and predicted negative. Then, a FN of 207, meaning the 207 predicted value was negative but actually positive (Type II error), and an FP of 55, meaning the 55 predicted value was positive but actually false (Type I error).

The GB Confusion matrix shows a TP of 2295, meaning the model predicted positive values of 2295, which were positive, and TN of 941, meaning the 941 predicted values were negative, and it predicted negative. Then, a FN of 194, meaning the 194 predicted value was negative but actually positive (Type II error), and an FP of 68, meaning the 68 predicted value was positive but actually false (Type I error)

The CB Confusion matrix shows a TP of 2293, meaning the model predicted positive values 0f 2293, which were positive, and TN of 933, meaning the 9333 predicted values were negative, and it predicted negative. Then, a FN of 202, meaning the 202 predicted value was negative but actually positive (Type II error), and an FP of 70, meaning the 70 predicted value was positive but actually false (Type I error).

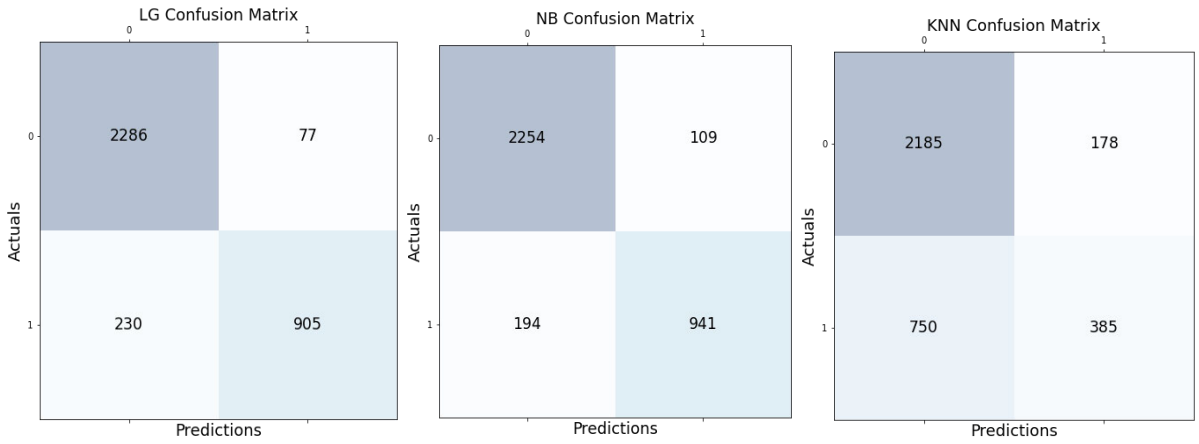

Figure 32:Confusion Matrix for Traditional ML models

The LR Confusion matrix shows a TP of 2286, implying the model predicted 2286 positive values, which were positive, and TN of 905, meaning 905 predicted values were negative, and predicted as negative. Then, a FN of 230, meaning 230 values, were predicted as negative but actually positive (Type II error), and an FP of 77, meaning 77 values, were predicted as positive but actually false (Type I error).

The NB Confusion matrix shows a TP of 2254, implying the model predicted 2254 positive values, which were positive, and TN of 941, meaning 941 predicted values were negative, and predicted as negative. Then, a FN of 194, meaning 194 values were predicted as negative but were actually positive (Type II error), and an FP of 109, meaning 109 values were predicted as positive but actually false (Type I error).

The KNN Confusion matrix shows a TP of 2185, implying the model predicted 2185 positive values, which were positive, and TN of 385, meaning 385 predicted values were negative, and predicted as negative. Then, a FN of 750, meaning 750 values were predicted as negative but were actually positive (Type II error), and an FP of 395, meaning 395 values were predicted as positive but actually false (Type I error).

4.7 Result for Research Question Three

Which emerging machine learning model can be adapted to predict student dropout in higher educational institutions?

Based on the performance of the models, the “RF” and the “GB” achieved high-performance accuracy. The author performed feature selection on the training data to determine which model to adopt in higher education. The Random Forest was used to select and predict the important feature, after which the outcome was evaluated on the training time and accuracy of the training set and confusion matrix.

Table 7:Feature Selection of Data

Target dataset value	Representation	Value
Withdrawn	1.0	5811
Pass	0.0	11677

```
In [102]: x.columns

Out[102]: Index(['date_registered', 'date_unregistered', 'date_submitted', 'score',
                 'sum_click', 'date', 'code_presentation', 'studied_credits',
                 'num_of_prev_attempts', 'gender_F', 'gender_M', 'imd_band', 'imd_band1',
                 'imd_band2', 'imd_band3', 'imd_band4', 'imd_band5', 'imd_band6',
                 'imd_band7', 'imd_band8', 'imd_band9', 'age_band1', 'age_band2',
                 'age_band3', 'highest_education_A Level or Equivalent',
                 'highest_education_HE Qualification',
                 'highest_education_Lower Than A Level',
                 'highest_education_No Formal quals',
                 'highest_education_Post Graduate Qualification'],
                dtype='object')
```

Figure 33:Shows all the 29 features in the dataset

Training a Random Forest classifier using all the features

As shown in Table 14, training a "RF" classifier using all the features resulted in 92% accuracy in about 11.56s of training time.

Table 8: Model performance using all features

	Training Time	Accuracy	Precision	recall	f1-score
RF	11.56	0.92	0.94	0.82	0.87

The feature importance analysis results obtained from the "RF" model are presented in Table 15. Among the input variables, the feature 'date' exhibited the highest importance score of 0.56, indicating its strong influence on the model's predictions Feature 'sum_click' followed with an importance score of 0.16, suggesting its significant impact. Conversely, the feature 'age_band' displayed a relatively lower importance score of 0.0003, implying its weaker influence on the model's decision-making process.

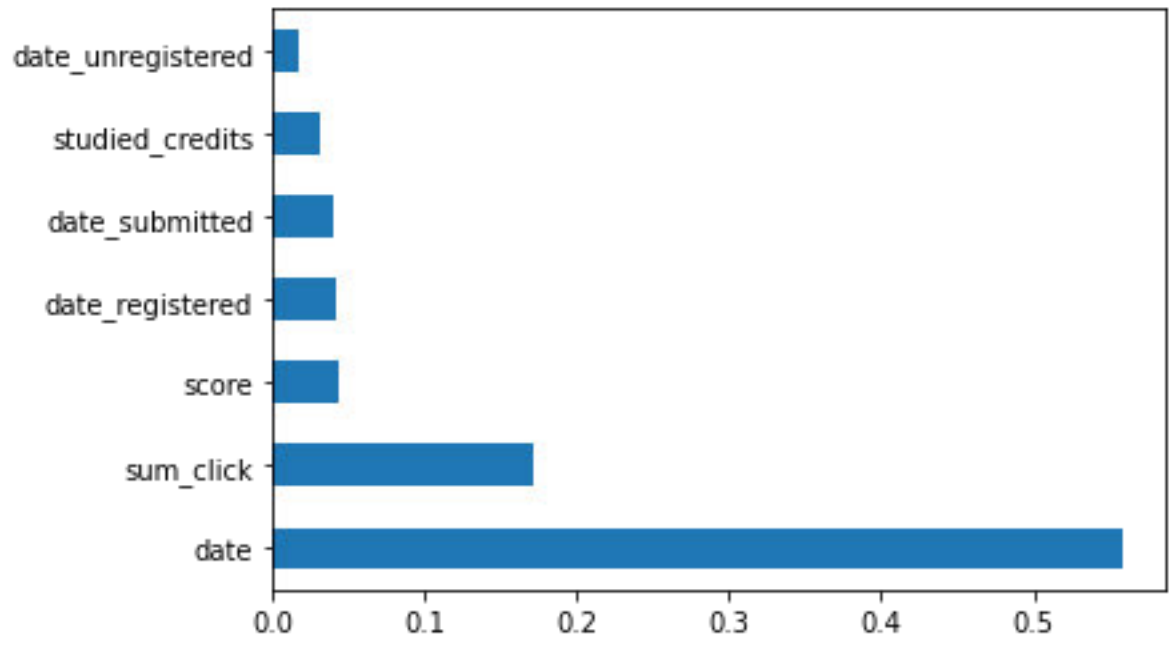

Figure 34:Feature importance ranking.

Train the model using the top 8 features.

Haven identified the features considered to be important by the RF model; the model was trained using the top 8 features. Figure 47 shows the results for the top 8 features.

Table 9:Model performance with feature selection

	Training Time	Accuracy	Precision	Recall	F1-score
RF	9.68s	0.92	0.94	0.81	0.87

The performance evaluation results on the testing set revealed that the RF model achieved an accuracy of 0.92, precision of 0.94, recall of 0.81, and F1-score of 0.87. The training time was reduced by 1.88s to achieve the same module performance.

4.8 Summary of Analysis

This study's data analysis provided answers to the three research questions. The study described each table used in the analysis. In order to distinguish graduating students from non-graduating students, features in the VLE log were categorized and labelled into months to plot patterns in students learning behaviours during the presentation. In an attempt to find the most used ML algorithm that can be adapted in HEI institutions, the result of the SLR identified some traditional and ensemble methods that authors commonly used. The current study compared six (6) ML models and evaluated their accuracy, precision, recall and f1-score performance. The analysis performed across the ensemble learning algorithms and the traditional ML algorithms showed that the ensemble learners performed better than traditional methods. The RF model performed better with high precision among the ensemble learners and the traditional ML algorithms. The confusion matrix showed the pictorial form of the predicted and actual classes for straightforward interpretations.

The next chapter will address each research question and include interpretations of the analysis provided in this chapter. The three (3) research questions provided context for the discussion and interpretations.

5 Chapter Five - Discussion

5.1 Introduction

This study utilised the best ML algorithm to develop a predictive model for a higher educational institution. This chapter presented a summary of the EDA performed on the dataset. It further discussed and provided implications of the results presented in the previous chapter making connections between the literature. The research questions formed the themes for the discussion in this chapter. First, research question one was addressed, followed by the second and the last research question.

The research questions for the study are:

1. What are the most prominent features/characteristics of students who graduated and those who did not graduate from higher educational institutions?
2. Which emerging machine learning methods/techniques have been applied successfully in the design of predictive systems in higher educational institutions?
3. Which emerging machine learning method can be successfully adapted to predict student dropout in higher educational institutions?

5.2 Summary of EDA on Dataset

The Dataset contains seven (7) modules offered in 22 presentations. That is approximately 3.14 presentations per module. 32593 registrations recorded for 28785 students, signifying 1.13 registrations per module. Which means each student registers for approximately one course. However, the previous attempts column indicated that some students had attempted a particular module more than once, Suggesting that students' IDs may duplicate. An aggregate of students' interaction with the Virtual learning environment (VLE) per course in a day is 10655280 recorded on 20 VLE activities. 31.1% of students dropped out of the course, and 21.6% failed. As a result, only 50% of the total number of students who registered for the course passed the

exam, indicating an issue with the student's success in completing the module. From the data, it can be observed that modules BBB, CCC, DDD and FFF recorded the highest numbers of dropouts. As indicated in the statistics provided in Table 24 (Appendix A), it can be observed that there is a high percentage of dropouts in the OULAD dataset. There were also differences in the final result and gender, based on the data that more males had passed marks and also more males withdrew from the module. This could be related to a lower enrolment of females in the modules. Also, it was observed that students who had an A level or equivalent in previous education passed the module more than qualifications students who had lower than A-level qualifications.

5.3 RQ1

What are the most prominent features/characteristics of students who graduate and those that did not graduate from Tertiary education?

Presentation modules AAA to GGG may have varied results for activities that students engaged in the most; however, this study focussed on the collective activities that students engage in the most in order to have a high-level insight into the relationship between students' online interaction and their final results. The observation in Table 10 and Figure 19 implies students interact with the ‘oucontent’ frequently, followed by ‘forumings’. This could be generalised to mean that both graduated and non-graduated students access the learning materials, engage in forums, and take part in quizzes. An interesting observation made on students' interaction percentages for quizzes and the homepage was 17.63% and 17.5%5, respectively, which could imply the percentages of students who visit the homepage are the same student the submitted quizzes. Students hardly repeated activities nor shared as lower sum-clicks were recorded for such activities. The implication is that students who engage in distance learning do not share their learning materials with their colleagues. Further insight into the analysis suggested that non-graduated students performed repeated activities 0.06% more than ‘graduated’ students. This could be related to the many times non-graduated students had’ to retake assessments in the module.

Frequency of revision and the final result of graduated and non-graduated students

The revision pattern for graduated and non-graduated students is similar during the early months of the module. As indicated in Fig. 20 and Fig 21, there was an increase in activities at the beginning of the module, ie. February to October presentations and October to May presentations for both graduated and non-graduated students. There are similar patterns for all the presentations in Figures 20, 21, 24, and 25; hence the discussion will incorporate all the presentations 2013B, 2014B, 2013J and 2014J. From the explanation in the dataset, students are given access to the materials before the module's start date, which explains why modules 2013J and 2014J are starting in September, which implies that students started accessing the learning materials before the start of the module. As indicated in Figures 24 and 25 of students, the sudden decline in activities in December may be due to the holidays. However, from Jan to March, revisions were increased for students who graduated and those who did not. This could have been a period approaching the final examination, which carries high weight. As observed, students who graduated and those who did not start revising their materials three (3) months before the examination period. However, those who graduated revised their materials more times than the non-graduated students; for example, in module 2013J, as shown in Fig. 26, the average frequency for a graduated student in the module was 1%, meaning the graduated students revised materials a few times more than the non-graduating students. Also, April and May saw another high interaction between graduated and non-graduated students, probably due to examination activities at the end of the module presentation. Then finally, in June and October, student activities dropped, signifying the end of the modules 2013B and 2014J, respectively.

However, the graph shows that students dropout of the modules as early as three (3) months from the module start date. For example, in the case of modules 2013J and 2014J, students dropout as early as November. In Figures 24 and 25 even though the graph showed significant inactivity of withdrawn students in the early months, some clicks were observed from withdrawn students until the end of the module. This suggested that students withdraw from the module at different levels. If such students are detected early an intervention can be recommended for them to assist them through their studies. It can also be observed in Figures 20 and 21 that the trend of students

who graduated and non-graduated students who withdraw from the module are quite similar, signifying that students who fail at the end of the module are potential withdrawal students who might need intervention as well.

In conclusion, it was observed that graduated students have the habit of interacting more with the VLE than non-graduated students. This corroborates with research that has looked at students' engagement with ‘vle’ and academic performance (Alqurashi 2019; Dvorak & Jia 2016; Coldwell et al. 2008). Also, students who graduated differ in character from non-graduated students when submitting assignments. Students who submitted assignments late ended up not graduating. Graduated students were observed to submit assignments earlier than the deadline, and non-graduated students submitted assignments mostly after the deadline. In my opinion, such students must be given attention and investigated the appropriate intervention to address their performance. Although some students labelled as graduated submit their assignments late, the number of non-graduated students exceeds that of the graduated students in the dataset. These findings agree with (Kuzilek et al. 2017), who predicted student performance based on the early assessment submission. The acquisition of data pertaining to students' unique characteristics and conduct is of significant benefit to educational stakeholders seeking to identify and address potential hurdles to teaching and learning while simultaneously facilitating the development of individualized solutions. Additionally, these findings can be utilized to enhance the efficacy of teaching methodologies by tailoring them to align with the learning tendencies of students.

5.4 RQ2

Which emerging machine learning methods/techniques have been applied successfully in the design of predictive systems in education?

Comparing Ensemble learning algorithms to traditional machine learning algorithms using

The performance evaluation of ML algorithms for predicting student dropout is crucial for selecting the most appropriate model for the task. This section aims to compare and evaluate the

performance of different algorithms using metrics such as accuracy, precision, recall, and F1-score. The algorithms considered for evaluation are “RF”,” GB”, “Catboost”, “KNN”, “LR” and “NB” algorithms. The results demonstrate the variations in performance across the algorithms, providing valuable insights into their strength and limitations.

The results in Table 12 and Fig. 30 and Fig.31 show the performance of ensemble and traditional ML algorithms. The comparison of the algorithms based on accuracy reveals that “RF” and “GB” achieved the highest accuracy of 93%, followed closely by “CB” with an accuracy of 92%. The lowest accuracy was 73% recorded by “KNN”. This result implies that the “RF” performed well in correctly classifying most of the data samples and making predictions for approximately 92% of the cases it encounters. Based on the analysis, the RF classifier has demonstrated superior accuracy compared to other ML algorithms (Valecha 2018). This insight could prove beneficial in enhancing our approach to identifying students who may be at risk and implementing successful interventions. However, accuracy alone may not provide a complete picture of a model's performance, as it is influenced by the class distribution and the occurrence of false positives and false negatives; other metrics, such as precision, recall and f1-score are also discussed.

Precision measures the ability of an algorithm to correctly identify positive samples. “RF” demonstrated the highest precision of 94%, indicating a low rate of false positives signifying that it is less likely to label students at-risk of dropping out when they are actually not. The potential benefits of achieving such a high level of precision are significant for predicting ARS in higher education. Schools and educational institutions could use this model to accurately identify students who are at a greater risk of dropping out. With this valuable information, they could provide targeted support and interventions tailored to address these ARS specific needs and challenges. In contrast, “KNN” exhibited a slightly lower precision of 63%. This suggests that “KNN” may generate more false positives, potentially affecting its performance in the application. It is concerning to consider the implications of a model with lower precision. This suggests that a considerable number of students identified as potential dropouts may not actually be at risk. This misidentification can result in misplaced allocation of resources and interventions, which may be directed towards students who do not require immediate support.

Additionally, inaccurately flagged students as at-risk may experience unnecessary burden and stress.

Recall, also known as sensitivity or true positive rate, assesses the ability of an algorithm to correctly identify all positive samples. “GB” and “NB” exhibited the highest recall of 83%, respectively, indicating their proficiency in capturing a larger proportion of positive samples. A higher recall score suggests that the model has a lower likelihood of missing out on identifying students at risk of dropping out, reducing the occurrence of false negatives. The “RF” and “GB” models appear to be effectively identifying many students who are truly at risk of leaving the module. This allows educational institutions to take proactive measures and provide timely interventions to support these students, enhancing their chances of remaining in school.

Conversely, “KNN” showed a slightly lower recall of 34%, suggesting that it may miss some positive instances. The “KNN” exhibits lower recall in predicting student dropout, indicating its suboptimal ability to correctly identify actual cases of students who will dropout, compared to all the positive cases present in the dataset. Ensuring the model's recall is optimized is crucial to prevent erroneous predictions that could negatively impact students' academic future. It is imperative for Educational institutions to identify a broader spectrum of students who could benefit from targeted support programs, mentoring, academic counselling, or other interventions aimed at enhancing student retention and success. This approach to early intervention is particularly valuable and should be prioritized.

F1-score is a harmonic mean of precision and recall, providing a balanced measure of an algorithm's performance. “RF” and “GB” achieved the highest F1-score of 88%, indicating a better balance between precision and recall. When predicting student dropout, a model with a higher F1 score value indicates a good balance between precision and recall. These are two essential evaluation metrics in binary classification tasks. “Catboost” and “NB” followed closely with an F1-score of 87% and 86%, respectively. Whiles the LG achieved an F1 score of 85%, “KNN” recorded the lowest F1-score of 45%, indicating a lower balance between precision and recall. This finding indicates that all the ensemble learning methods and traditional methods recorded good performance on the f1-score, meaning it can accurately identify a significant

portion of students at risk (recall) while keeping the rate of false positives low (precision). This is beneficial for educational institutions because it allows them to direct resources and interventions towards students who genuinely require support, enhancing their likelihood of success and decreasing dropout rates.

On the other hand, the “KNN” perform poorly on F1-score, suggesting a higher incidence of false positives or false negatives. False positives mean that students may be incorrectly identified as at-risk when they are not, leading to unnecessary interventions. False negatives imply that ARS are missed and do not receive timely support and intervention. This can have serious implications for resource allocation and intervention strategies in higher educational institutions.

The performance of the model illustrated in all the metrics indicates the higher performance of the ensemble learning algorithms (“RF”, “GB”, “CatBoost”) over the traditional learning algorithms (“LR”, “KNN”, and “NB”). This observation is consistent with studies (Dass et al. 2021; Behr et al. 2020; Lee & Chung 2019; Kubayi et al. n.d.) that have conducted similar experiments using the ensemble learning methods and traditional ML methods. Although the result of the studies conforms to similar studies in the domain, the results obtained in this current study show better performance of “RF” with an accuracy of 0.933% and precision of 0.94%. The results indicate that different algorithms exhibit variations in their performance. For instance, It can be noticed that the random forest and the Gradient boosting algorithms also performed better among the ensemble learning algorithms. The “RF” outperformed both the traditional and ensemble learning methods with 93% accuracy and 94% precision. This implies that “RF” algorithms make accurate predictions on the dataset by limiting the number of samples incorrectly predicted as negative.

5.5 RQ3

Which emerging machine learning model can be adapted to predict student dropout in higher institutions in Ghana?

The results obtained in the comparison of the models informed the ML technique that could be adapted to predict student dropout in higher institutions. The Random Forest algorithm was applied to conduct feature importance analysis and performance evaluation.

The feature importance analysis using the “RF” algorithm provides valuable insights into the relevance and impact of input variables. ‘Date of assessment’ was identified as the most important predictor, indicating its substantial influence on the model's decision-making process. Sumclick of student logs in the virtual learning environment closely followed, highlighting its significant contribution to the model's predictions, followed by score, date of registration, date of assessment submission, studied credits and date the student unregistered from the course (Figure 33 shows the top 8 important features). These findings are consistent with other studies. For example, (Marbouti, Diefes-Dux, et al. 2016) used a feature selection method to reduce the number of variables used in order to increase the generalizability of the model's accuracy in predicting ARS. On the other hand, the age of students demonstrated a relatively lower importance score, suggesting its diminished role in the decision-making of predicting ARS. This finding contradicts the findings of (Lakkaraju et al. 2015), which indicated that age is an important feature for prediction.

The model training time for 26 features was 11.56, indicating that each model takes 2.25s to train. However, the training time was reduced to 9s with a feature selection of 8 features. As indicated in Tables 14 and 15, both models achieved the same accuracy performance of 0.92% but with different time rates. This indicates that given the right feature, the model can perform better within the shortest time. This can go a long way to solve the problem of limited resources and high computing power (De-Arteaga et al. 2018), which inhibit the deployment of ML implementations, especially in Ghana. This implies that feature selection techniques could be utilized to achieve better performance using less computing resources.

The effect of feature selection on classification was tested. In this case, the eight (8) most important features were compared with those that use all features. The results indicated in Table 14 that the classification performance with 8 features selected according to the tree-based feature importance method achieved 92% in 9.68s compared to the classification performances obtained

using all features. The feature selection method is important because it enables the formation of prediction models through the use of a lower number of features, and using a lower number of features means the model can be interpreted more easily. Another important aspect of using feature selection is to gain an understanding of which features have the greatest impact on the class that needs to be predicted.

The performance evaluation results indicate that the tree-based feature selection based on the Random Forest model achieved satisfactory performance across accuracy, precision, recall, and F1-score and the model training time. The model leveraged the insights from feature importance analysis to classify and make accurate predictions effectively. The performance of the model of 0.92% accuracy reflects the overall correctness of the model's predictions, while precision quantifies the model's ability to correctly identify positive samples at 0.94%. The recall value demonstrates the model's capability to identify all positive samples at 0.81%, and the F1-score provides a balanced measure between precision and recall at 0.87%.

5.6 Summary of Discussion

This section discussed the results in relation to the study's research question. Student behaviour and habits associated with ARS include accessing learning materials, revision patterns, and early submission of assessments. ‘Non-graduated’ students had habits of not revising the learning materials early before the final exams. Although it was noted that both graduated and non-graduated students access the learning materials simultaneously, variations were recorded in the habits of assignment submission and revision patterns. ‘Graduated’ students recorded higher clicks for activities than non-graduated students, which signifies that the graduated students interacted more with course activities than non-graduated students. However, few graduated students recorded lower clicks. This study has compared different ML algorithms and determined the method that achieved the best predictive accuracy that could be adapted in higher educational institutions. The discussion of the algorithm evaluation concluded that the ensemble ML methods outperformed the traditional methods. The “RF” ensemble learning algorithms outperformed the “Catboost”, “KNN”, “LR”, and “NB” on the accuracy, precision, recall and f-1 score whereas the GB and the RF had the same accuracy, recall and f-score, the “RF” performed

the “GB” on precision. The “RF” emerged as the best model for predicting ARS based on the findings of this study. This study based performance on higher precision as predicting ARS requires correctly identifying the positive class to avoid incorrectly raising false information and allocating resources to students who might not need it. The study further designed a predictive system using a tree-based feature selection method. The study identified important features such as “date of assignment submission”, “sum_clicks of activities”, “score on the assessment”, ”date of registration”, “date of assignment submission”, “studied credits”, and “date the student unregistered” for predicting at-risk students. The model used the important features to predict ARS and achieved better accuracy of 92%.

5.7 Implications for practice and/or policy

The PhD study presented significant insights and contributions to the field of LA and the application of ML in HEI. Firstly, the systematic literature review revealed that studies in LA are limited in Ghana. The current study showed the importance of context in the design and development of AI solutions in higher educational institutions since there is no one size fits all solution. The study contributes to the importance of context in the institution of LA in higher education in Ghana. Practically, educational stakeholders must make an effort to store and make student data readily available to researchers to analyse and improve teaching and learning (Tahiru & Parbanath 2023). This contribution serves as a motivation to educational stakeholders in facilitating and institutionalising LA in higher education.

Secondly, the study proposed a comprehensive LA framework that integrates ML workflow for predicting at-risk. The framework provides a simplified and easy-to-reproduce methodological approach to identify the best ML algorithm to predict ARS. The framework is specifically designed to predict students who are at risk and provide educational stakeholders with effective intervention strategies. By utilising feature engineering techniques, feature selection and carefully selected ML algorithms, it offers highly accurate and reliable predictions that enable educators to identify struggling students who are at risk of dropping out and offer them the

necessary support to succeed. This approach will transform the learning environment by creating a safer and more successful space for all students.

Thirdly, this study systematically reviewed the commonly used algorithms used in the design of predictive systems in higher education. The Random Forest, Support Vector Mchine. K-nearest Neighbour, Logistic regression, and Naive Bayes algorithms emerged as the commonly used algorithms. The performance of the algorithms was compared and contrasted through the design of a model. This contribution provides valuable insight into the different algorithms' strengths and weakness and their suitability for predicting ARS. This contribution also served as a basis and guide in the trend and direction for researchers in LA to improve.

Fourthly, the study contributed by identifying the best features for determining ARS in higher educational institutions. For example, the findings of (Baker 2010) suggest that the feature selection method aims to gain insight into which features have the most significant impact on academic performance that needs to be predicted. The best features serve as a guide to educational stakeholders in designing curriculums and course instructions for students. The features identified as important in the online dataset are equally important to student success in higher educational institutions in Ghana. For example, In the light of the pandemic, almost all institutions in Ghana resorted to the use of online education as an additional mode of instruction (Edumadze et al. 2022; Adarkwah 2021; Ali 2020; Upoalkpajor & Upoalkpajor 2020), which translates that data about students learning and interactions can be accessed and interpreted. Therefore, feature selection would enhance the development of predictive systems using fewer features about students in a higher educational institution in designing interventions and improving studies. These findings can motivate educational stakeholders to facilitate and formulate policy guidelines for LA implementation in higher education.

Lastly, building upon comparative analysis, this study provides practical evidence of the performance of the ensemble learning models over the traditional learning ML methods in designing Predictive systems in higher education Institutions. Predictive systems in higher education research continue to be at the forefront of ML and LA research. It emerged as the top

theme for research in the domain of ML research (Tahiru, Parbanath & Agbesi,2023). The use of ensemble learning in developing a predictive system demonstrated better performance with high precision. The study results serve as a blueprint for researchers using similar data to identify ARS in higher educational institutions. The methodological steps, algorithms employed, and the identified features could serve as a guide to educators and stakeholders in the early detection of students at risk of dropout and the design of the needed intervention to prevent dropout. This contribution demonstrates the effectiveness of ensemble learning in enhancing the accuracy and robustness of student dropout prediction models.

6 Chapter Six – Conclusion, Limitations and Recommendations

This section presents the overall conclusion of the studies providing summaries of all significant findings, followed by the limitations of the study and the recommendation for future studies.

This study aimed to solve the critical issue of student dropout in higher educational institutions by designing a ML model for predicting ARS in higher educational institutions in Ghana. Even though online data was utilised for this research analysis, the facts exist that both online and face-to-face modes of study experience students' dropout that need immediate interventions. The ML models were designed using RF, GB, Catboost, KNN, LR and NB algorithms. An integrated approach of the 5-step LA process and the ML workflow with scikit learn is employed for the model development. This chapter, therefore, concludes the entire thesis as follows:

- The results of determining behaviours among students who graduate and those who did not from higher educational institutions using the VLE logs.
- Results on comparative analysis of ensemble learning models and traditional ML models
- The results of key features for determining student dropout using the RF model.

Differentiating between the habits and trends of students who graduate from students who do not by analysing students' logs in the virtual learning environment is crucial in identifying students who need assistance in higher education. The findings of the EDA suggested that students who engage more with the learning materials provided online by the Open University, such as submitting assignments early, taking part in forums, and revising their learning materials early, usually pass the course and, therefore, graduate. On the other hand, non-graduated students have the habit of interacting with the learning materials mostly at the inception of the module and showed no consistency in revising learning materials three months earlier before the examination period. This insight gathered about student characteristics and behaviour provides educational stakeholders with first-hand information to investigate and provide tailored solutions to students with challenges in learning. Also, insight into student learning habits is identified to inform the pedagogical structure in higher educational institutions.

Different ML methods were employed in this study to determine the best approach that increases predictive performance to model students' dropout using student VLE logs, student courses, assessments, and student demographics and social features. The results indicated that the ensemble learning algorithms outperformed all the traditional ML algorithms after training and testing algorithms across the traditional and ensemble algorithms. The author evaluated the models' performance on the accuracy, precision, recall and f1-score metrics. The findings indicated that the RF emerged as the best-performing predictive model, having attained an accuracy of 93% and a precision of 94%. The Gradient boosting followed closely with the same accuracy of 93% but was outperformed by the RF with a precision of 92%.

In analyzing the characteristics and behaviours that best indicate students needing intervention to improve academic performance, a predictive system using a tree-based feature selection method was designed. The results of the study indicated that the most important features to predict ARS include the "date of assessment", "sum clicks", "score", "date of registration", "date of assessment submission", "studied credits", and date the student unregistered from the course. These features emerged as the most important features for predicting students' dropout in higher educational institutions within a minimal time of 9s, compared to using all features with a running time of 11.56s. Similar features can be utilized in higher educational institutions in Ghana to identify students at risk of dropout with less computing resources.

The study concluded that student dropout could be predicted by utilising LA and ML approaches, and interventions could be designed to minimise the number of students dropping out of higher educational institutions. The data gathered from educational institutions at all levels can serve as a great source for identifying students' learning habits, improving student learning habits, identifying students' problems early, and providing interventions. With the needed support and access to anonymised educational data by higher educational institutions, this study could be implemented institutionally to provide a solution to the higher education sector. The findings of this study can be used in higher education institutions to create tailored dropout prevention strategies, such as EWS, study assistance, mentorship programs and recommender systems that run throughout the educational levels.

6.1 Limitations of study

The research focus was to obtain student academic data and behavioural data from the GCTU in Ghana, but the author was unable to obtain such data from the authorities in the institution due to the sensitive nature of students' data. The researcher also reached out to different Universities in Ghana, but none was willing to give out student data for the research. Hence the research utilized the publicly available OULAD to answer the study's research questions. Due to the lack of qualitative data on the OULAD, this study faced constraints similar to any purely quantitative investigation. The dataset lacks explanations of reasons for students' withdrawal which could have been a basis for designing interventions to prevent students from failing. Many factors, such as motivation, student readiness for academic rigour, and economic and social status, were identified in the literature that could lead to student dropouts; however, students in the dataset were not defined in such features. Many of these factors could have been used to predict student dropout, which could assist educational stakeholders in addressing the student dropout problems in higher institutions.

6.2 Recommendations for future studies

The design of an EWS in a higher educational institution requires efforts and strict measures from educational stakeholders. The integration of the accurate and precise results of the dropout predictive model requires key decisions from stakeholders. Firstly, the educational stakeholder must promote the efficiency of educational data and make it available to the research administration to design and implement predictive systems that can detect students at risk of dropout early before it happens. Secondly, educational stakeholders are obligated to institute and/ or strengthen the existing advising and mentoring processes within higher educational institutions. This can serve as an intervention for students and a platform for them to open up and discuss their problems. Lastly, the available data stored in student databases can be utilised to identify students' behaviours and performance trends that can be monitored to suggest various student interventions. The early identification of students at risk of dropout can facilitate the educational stakeholders to design strategies such as student support and guidance committees in

the educational community. The deployment of predictive systems in education has shown promising results in improving students' performance in higher education institutions. To further enhance these results, techniques such as deep learning and Internet-of-Things (IoT) can be integrated into the design of predictive systems to achieve robust and high-performance systems. In addition, future studies must focus on implementing intervention systems, such as recommendation systems, to support ARS and prevent them from dropping out.

7 References

Abdulkareem, SA, Foh, CH, Lee, H, Carrez, F & Moessner, K. 2022. IoT Network Intrusion Detection with Ensemble Learners. *International Conference on ICT Convergence*. 2022-Octob:510–514. doi.org/10.1109/ICTC55196.2022.9952376.

Acquah, A. 2021. Acquah, A. Higher Education Finance between Ghana and the U.S. *Current Issues in Comparative Education (CICE)*. 23(1):90–108.

Adam, S, Adom, D & Bediako, AB. 2016. The Major Factors That Influence Basic School Dropout in Rural Ghana: The Case of Asunafo South District in the Brong Ahafo Region of Ghana. *Journal of Education and Practice*. 7(28):1–8. Available from: https://search.proquest.com/docview/1871574449?accountid=13042.

Adarkwah, MA. 2021. "I'm not against online teaching, but what about us?": ICT in Ghana post Covid-19. *Education and Information Technologies*. 26(2):1665–1685. doi.org/10.1007/s10639-020-10331-z.

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

Akçap, G, Altun, A & Petek, A. 2019. Using learning analytics to develop early- warning system for at-risk students.

Al-shehri, H, Al-qarni, A, Al-saati, L, Batoaq, A, Badukhen, H, Alrashed, S, Alhiyafi, J & Olatunji, SO. 2017. Student Performance Prediction Using Support Vector Machine and K-Nearest Neighbor. 17–20.

Alboaneen, D, Almelihi, M, Alsubaie, R, Alghamdi, R, Alshehri, L & Alharthi, R. 2022. doi.org/10.3390/data7020021.

Ali, W. 2020. Online and Remote Learning in Higher Education Institutes: A Necessity in light of COVID-19 Pandemic. *Higher Education Studies*. 10(3):16. doi.org/10.5539/hes.v10n3p16.

Aljohani, NR, Fayoumi, A & Hassan, SU. 2019. Predicting at-risk students using clickstream

data in the virtual learning environment. *Sustainability (Switzerland)*. 11(24):1–12. doi.org/10.3390/su11247238.

Alkhasawneh, R & Hobson, R. 2011. Modeling Student Retention in Science and Engineering Disciplines Using Neural Networks. 660–663.

Alokluk, JA. 2018. The Effectiveness of Blackboard System, Uses and Limitations in Information Management. *Intelligent Information Management*. 10(06):133–149. doi.org/10.4236/iim.2018.106012.

Alqurashi, E. 2019. Predicting student satisfaction and perceived learning within online learning environments. *Distance Education*. 40(1):133–148. doi.org/10.1080/01587919.2018.1553562.

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). doi.org/10.1186/s41239-020-0177-7.

Amara Atif, DR, Ayse, B & Mauricio, M. 2013. Learning Analytics in Higher Education:A summary of Tools and Approaches. In: *10th ascilite Conference*. V. 163. Macquarie University,Sydney. 667–667. doi.org/10.1093/toxsci/kfy118.

Van Appel, V & Durandt, R. 2019. INVESTIGATING POSSIBILITIES of PREDICTIVE MATHEMATICAL MODELS to IDENTIFY at RISK STUDENTS in the SOUTH AFRICAN HIGHER EDUCATION CON TEXT. *Perspectives in Education*. 37(2):1–15. doi.org/10.18820/2519593X/pie.v37i2.1.

Apusigah, AA. 2009. Over fifty years of higher education in Ghana: What has happened to Equity? *Contemporary social problems in Ghana*. (September):35–54.

Arnold, KE, Hall, Y, Street, SG, Lafayette, W, Pistilli, MD, Hall, Y, Street, SG & Lafayette, W. 2012. course signals in Pudu_using learning analytics to enhance leanring .pdf. (May):267–270.

Atif, A, Richards, D & Bilgin, A. 2015. Student preferences and attitudes to the use of early alerts. *2015 Americas Conference on Information Systems, AMCIS 2015*. (August 2016).

Atuahene, F & Owusu-Ansah, A. 2013. doi.org/10.1177/2158244013497725.

Awiagah, R, Kang, J & Lim, JI. 2016. Factors affecting e-commerce adoption among SMEs in Ghana. *Information Development*. doi.org/10.1177/0266666915571427.

Ayisi, AE. 2018. Higher Education Institutions' Impacts on the Socio-Economic Growth of Ghana. *International Journal of Education and Research*. 6(9):145–162. Available from: www.ijern.com.

Baars, GJAA, Stijnen, T & Splinter, TAWW. 2017. A Model to Predict Student Failure in the First Year of the Undergraduate Medical Curriculum. *Health Professions Education*. 3(1):5–14. doi.org/10.1016/j.hpe.2017.01.001.

Bainbridge, J, Melitski, J, Zahradnik, A, Lauría, EJM, Jayaprakash, S & Baron, J. 2015. Using Learning Analytics to Predict At-Risk Students in Online Graduate Public Affairs and Administration Education. *Journal of Public Affairs Education*. 21(2):247–262. doi.org/10.1080/15236803.2015.12001831.

Baker, RSJD. 2010. Mining data for student models. *Studies in Computational Intelligence*. 308:323–337. doi.org/10.1007/978-3-642-14363-2_16.

Baranauskas, JA, Netto, OP, Nozawa, SR & Macedo, AA. 2018. A tree-based algorithm for attribute selection. *Applied Intelligence*. 48(4):821–833. doi.org/10.1007/s10489-017-1008-y.

Bawakyillenuo, S, Osei-Akoto, I, Ahiadeke, C, Aryeetey, B & Agbe, K. 2013. Tertiary education and industrial development in Ghana. *International Growth Centre*. (August):53.

Behr, A, Giese, M & K, HDT. 2020. Early Prediction of University Dropouts – A Random Forest Approach.

Berk, RA. 2005. An Introduction to Ensemble Methods for Data Analysis.

Berland, M, Baker, RS & Blikstein, P. 2014. Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and Learning*. 19(1–2):205–220. doi.org/10.1007/s10758-014-9223-7.

Biau, G & Scornet, E. 2016. A random forest guided tour. *Test*. 25(2):197–227. doi.org/10.1007/s11749-016-0481-7.

Blunch, N-H. 2020. Learning from Ghana Recent Human Capital Improvements. *International Bank for Reconstruction and Development/The World Bank*. 282. Available from: www.worldbank.org.

Bravo-Agapito, J, Romero, SJ & Pamplona, S. 2021. Early prediction of undergraduate Student's

academic performance in completely online learning: A five-year study. *Computers in Human Behavior*. 115(October 2020). doi.org/10.1016/j.chb.2020.106595.

Burman, I & Som, S. 2019. Predicting Students Academic Performance Using Support Vector Machine. 13–16.

Campbell, JP & Oblinger, DG. 2007. Academic analytics. *EDUCAUSE review*. 42(4):40–57.

Chen, Y & Yang, H. 2016. A Novel Information-Theoretic Approach for Variable Clustering and Predictive Modeling Using Dirichlet Process Mixtures. *Scientific Reports*. 6(July):1–13. doi.org/10.1038/srep38913.

Chen, X, Xie, H, Zou, D & Hwang, GJ. 2020. Application and theory gaps during the rise of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*. 1(August):100002. doi.org/10.1016/j.caeai.2020.100002.

Cho, MH & Tobias, S. 2016. Should instructors require discussion in online courses? Effects of online discussion on community of inquiry, learner time, satisfaction, and achievement. *International Review of Research in Open and Distance Learning*. 17(2):123–140. doi.org/10.19173/irrodl.v17i2.2342.

Christie, ST, Jarratt, DC, Olson, LA & Taijala, TT. 2019. Machine-learned school dropout early warning at Scale. *EDM 2019 - Proceedings of the 12th International Conference on Educational Data Mining*. (Edm):726–731.

Chui, KT, Fung, DCL, Lytras, MD, Lam, TM, Tai, K, Chun, D, Fung, L, Lytras, MD, et al. 2020. Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. *Computers in Human Behavior*. 107(December 2017):105584. doi.org/10.1016/j.chb.2018.06.032.

Chung, JY & Lee, S. 2019. doi.org/10.1016/j.childyouth.2018.11.030.

Cisneros, L. 2020. Tertiary Education, Career Preparation, and Job Prospects: An International Perspective. 40–49.

Coldwell, J, Craig, a, Paterson, T & Mustard, J. 2008. Online students: Relationships between participation, demographics and academic performance. *The Electronic Journal of e-Learning*. 6(1):19–28. Available from: http://iucontent.iu.edu.sa/Scholars/Information Technology/Online

Students Relationships between Participation, Demographics and Academic Performance.pdf.

Coleman, CJ. 2021. Exploring a Generalizable Machine Learned Solution for Early Prediction of Student At-Risk Status. *ProQuest Dissertations and Theses*. (April). doi.org/10.7916/d8-5scb-n214.

Costa, LA, Souza, MV das S, Salvador, L do N & Amorim, RJR. 2019. Monitoring Students Performance in E-learning based on Learning Analytics and Learning Educational Objectives. In: *2019 IEEE 19th International Conference on Advanced Learning Tectnologies*. IEEE. doi.org/10.1109/ICALT.2019.00067.

Dass, S, Gary, K & Cunningham, J. 2021. Predicting Student Dropout in Self-Paced MOOC Course Using Random Forest Model.

Dawson, S, Jovanovic, J, Gašević, D & Pardo, A. 2017. From prediction to impact: Evaluation of a learning analytics retention program. *ACM International Conference Proceeding Series*. 474–478. doi.org/10.1145/3027385.3027405.

De-Arteaga, M, Herlands, W, Neill, DB & Dubrawski, A. 2018. Machine learning for the developing world. *ACM Transactions on Management Information Systems*. 9(2). doi.org/10.1145/3210548.

Deepak, KC. 2017. Evaluation of Moodle Features at Kajaani University of Applied Sciences-Case Study. *Procedia Computer Science*. 116:121–128. doi.org/10.1016/j.procs.2017.10.021.

Djan, J & George, B. 2016. Standardization or localization: A study of online learning programmes by tertiary institutions in Ghana. *European Journal of Contemporary Education*. 18(4):430–437. doi.org/10.13187/ejced.2016.18.430.

Doneva, R, Gaftandzhieva, S & Bandeva, S. 2021. Data Analytics Tools in Higher Education. *CEUR Workshop Proceedings*. 3061(January):91–99.

Dvorak, T & Jia, M. 2016. Do the Timeliness, Regularity, and Intensity of Online Work Habits and Academic Performance? *Journal of Learning Analytics*. 3(3):318–330.

Dyckhoff, AL. 2014. Action research and learning analytics in higher education. *E-Learning and Education : Eleed*. 1(10).

Edumadze, JKE, Barfi, KA, Arkorful, V & Baffour Jnr, NO. 2022.

doi.org/10.1080/10494820.2021.2018618.

Er, E. 2012. Identifying At-Risk Students Using Machine Learning Techniques: A Case Study with IS 100. *International Journal of Machine Learning and Computing*. 2(4):476–480. doi.org/10.7763/ijmlc.2012.v2.171.

Fairos, W, Yaacob, W & Nasir, S. 2019. Supervised data mining approach for predicting student performance. (December):1584–1592. doi.org/10.11591/ijeecs.v16.i3.pp1584-1592.

Farrahi, V, Niemelä, M, Tjurin, P, Kangas, M, Korpelainen, R & Member, S. 2019. Evaluating and Enhancing the Generalization Performance of Machine Learning Models for Physical Activity Intensity Prediction from Raw Acceleration Data. 1–12. doi.org/10.1109/JBHI.2019.2917565.

Faulconer, E, Griffith, JC & Frank, H. 2021. If at first you do not succeed: student behavior when provided feedforward with multiple trials for online summative assessments. *Teaching in Higher Education*. 26(4):586–601. doi.org/10.1080/13562517.2019.1664454.

Ferguson, R, Clow, D, Griffiths, D & Brasher, A. 2019. Moving forward with learning analytics: Expert views. *Journal of Learning Analytics*. 6(3):43–59. doi.org/10.18608/jla.2019.63.8.

Fisher, A, Rudin, C & Dominici, F. 2019. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*. 20:1–81.

Flanagan, B & Flanagan, B. 2018. Learning analytics platform in higher education in Japan Recommended citation : Flanagan , B ., & Ogata , H . (2018). Learning analytics platform in higher Learning analytics platform in higher education in Japan Hiroaki Ogata *. *Japan*. 10(4):469–484.

Gardner, J & Brooks, C. 2018. Student success prediction in MOOCs. *User Modeling and User-Adapted Interaction*. 28(2):127–203. doi.org/10.1007/s11257-018-9203-z.

Ghana National Council for Tertiary Education. 2018. Statistical Report on Tertiary Education for 2015 / 2016 Academic Year. *National Council For Tertiary Education*. 1(1):1–48.

Ghorbani, R & Ghousi, R. 2020. Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques. *IEEE Access*. 8:67899–67911.

doi.org/10.1109/ACCESS.2020.2986809.

Girdwood, A. 1999. Tertiary Education Policy in Ghana. An Assessment: 1988-1998. (May):1–84. Available from: http://eric.ed.gov/ERICWebPortal/recordDetail?accno=ED453719.

Gómez-Ramírez, J, Ávila-Villanueva, M & Fernández-Blázquez, MÁ. 2020. Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutation-based methods. *Scientific Reports*. 10(1):1–15. doi.org/10.1038/s41598-020-77296-4.

Gray, CC & Perkins, D. 2019. Utilizing early engagement and machine learning to predict student outcomes. *Computers and Education*. 131(December 2018):22–32. doi.org/10.1016/j.compedu.2018.12.006.

Hansen, LK & Salamon, P. 1990. Neural Network Ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 12(10):993–1001. doi.org/10.1109/34.58871.

Hasan, R, Palaniappan, S, Mahmood, S, Abbas, A, Sarker, KU & Sattar, MU. 2020. Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. *Applied Sciences (Switzerland)*. 10(11). doi.org/10.3390/app10113894.

Hashim, AS, Awadh, WA & Hamoud, AK. 2020a. Student Performance Prediction Model based on Supervised Machine Learning Algorithms Student Performance Prediction Model based on Supervised Machine Learning Algorithms. *IOP conference series:Material Science and Engineering*. doi.org/10.1088/1757-899X/928/3/032019.

Hashim, AS, Awadh, WA & Hamoud, AK. 2020b. Student Performance Prediction Model based on Supervised Machine Learning Algorithms. *IOP Conference Series: Materials Science and Engineering*. 928(3). doi.org/10.1088/1757-899X/928/3/032019.

Hassan, SU, Waheed, H, Aljohani, NR, Ali, M, Ventura, S & Herrera, F. 2019. Virtual learning environment to predict withdrawal by leveraging deep learning. *International Journal of Intelligent Systems*. 34(8):1935–1952. doi.org/10.1002/int.22129.

He, Y, Chen, R, Li, X, Hao, C, Liu, S & Zhang, G. 2020. Online At-Risk Student Identification Using RNN-GRU Joint Neural Networks _ Enhanced Reader.pdf. *Information*. 11(10):474.

Heinze, G, Wallisch, C & Dunkler, D. 2018. Variable selection – A review and

recommendations for the practicing statistician. *Biometrical Journal*. 60(3):431–449. doi.org/10.1002/bimj.201700067.

Hlosta, M, Zdrahal, Z & Zendulka, J. 2017. Ouroboros: Early identification of at-risk students without models based on legacy data. *ACM International Conference Proceeding Series*. 6–15. doi.org/10.1145/3027385.3027449.

Howard, E, Meehan, M & Parnell, A. 2018. Contrasting prediction methods for early warning systems at undergraduate level. *Internet and Higher Education*. 37(February):66–75. doi.org/10.1016/j.iheduc.2018.02.001.

Hussain, M, Zhu, W, Zhang, W, Abidi, SMR & Ali, S. 2019. Using machine learning to predict student difficulties from learning session data. *Artificial Intelligence Review*. 52(1):381–407. doi.org/10.1007/s10462-018-9620-8.

Ifenthaler, D & Yau, JYK. 2020. Utilising learning analytics to support study success in higher education: a systematic review. *Educational Technology Research and Development*. 68(4):1961–1990. doi.org/10.1007/s11423-020-09788-z.

Ihantola, P, Vihavainen, A, Ahadi, A, Butler, M, Börstler, J, Edwards, SH, Isohanni, E, Korhonen, A, et al. 2015. Educational data mining and learning analytics in programming: Literature review and case studies. *ITiCSE-WGP 2015 - Proceedings of the 2015 ITiCSE Conference on Working Group Reports*. 41–63. doi.org/10.1145/2858796.2858798.

Islam Sarker, MN, Wu, M & Hossin, MA. 2019. Economic effect of school dropout in Bangladesh. *International Journal of Information and Education Technology*. 9(2):136–142. doi.org/10.18178/ijiet.2019.9.2.1188.

Jayaprakash, SM, Moody, EW, Lauría, EJM, Regan, JR & Baron, JD. 2014. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. *Journal of Learning Analytics*. 1(1):6–47. doi.org/10.18608/jla.2014.11.3.

Joe Adu-Agyem and Patrick Osei-Poku. 2012. Quality Education In Ghana:The way Forward. *International Journal of Innovative Research and Development*. 1(9):164–177.

Jokhan, A, Sharma, B & Singh, S. 2019. Early warning system as a predictor for student performance in higher education blended courses. *Studies in Higher Education*. 44(11):1900–

1911. doi.org/10.1080/03075079.2018.1466872.

Kabari, LG & Onwuka, UC. 2019. Comparison of Bagging and Voting Ensemble Machine Learning Algorithm as a Classifier. *International Journals of Advanced Research in Computer Science and Software Engineering*. 9(3):19–23. Available from: www.ijarcsse.com,.

Kabathova, J & Drlik, M. 2021. Towards predicting student's dropout in university courses using different machine learning techniques. *Applied Sciences (Switzerland)*. 11(7). doi.org/10.3390/app11073130.

Kent, C, Laslo, E & Rafaeli, S. 2016. Interactivity in online discussions and learning outcomes. *Computers and Education*. 97:116–128. doi.org/10.1016/j.compedu.2016.03.002.

Kika, CA. 2018. Title : Supporting student experience management with learning analytics in the UK higher education sector.

Kim, Y, Heider, PM & Meystre, S. 2018.

Knowles, J. 2014. Of Needles and Haystacks: Building an Accurate Statewide Dropout Early Warning System in Wisconsin. *JEDM - Journal of Educational Data Mining*. 7(3):1–52. Available from: http://figshare.com/articles/Of_Needles_and_Haystacks_Building_an_Accurate_Statewide_Dropout_Early_Warning_System_in_Wisconsin/1142580.

Kokoç, M & Altun, A. 2021. Effects of learner interaction with learning dashboards on academic performance in an e-learning environment. *Behaviour and Information Technology*. 40(2):161–175. doi.org/10.1080/0144929X.2019.1680731.

Kostopoulos, G, Karlos, S & Kotsiantis, S. 2019. Multiview Learning for Early Prognosis of Academic Performance: A Case Study. *IEEE Transactions on Learning Technologies*. 12(2):212–224. doi.org/10.1109/TLT.2019.2911581.

Kubayi, SC, Jadhav, A & Ajoodha, R. n.d. A Machine Learning Approach for Predicting Students ' Second Year Outcomes.

Kumi-Yeboah, A. 2010. A Look at the Trend of Distance and Adult Education in Ghana. (Undetermined). *International Forum of Teaching & Studies*. 6(1):19–67. Available from: http://search.ebscohost.com/login.aspx?direct=true&db=eue&AN=508129525&site=ehost-

live&scope=site.

Kuranchie, A, Okyere, M & Larbi, E. 2021. A Non-state University's Contribution to the Tertiary Education Landscape in Ghana. *International Journal of Academic Research in Progressive Education and Development*. 10(1):99–113. doi.org/10.6007/ijarped/v10-i1/8327.

Kurilovas, E. 2019. Advanced machine learning approaches to personalise learning: learning analytics and decision making. *Behaviour and Information Technology*. 38(4):410–421. doi.org/10.1080/0144929X.2018.1539517.

Kuzilek, J, Hlosta, M & Zdrahal, Z. 2017. Data Descriptor: Open University Learning Analytics dataset. *Scientific Data*. 4:1–8. doi.org/10.1038/sdata.2017.171.

Lacave, C, Molina, AI & Cruz-Lemus, JA. 2018. Learning Analytics to identify dropout factors of Computer Science studies through Bayesian networks. *Behaviour and Information Technology*. 37(10–11):993–1007. doi.org/10.1080/0144929X.2018.1485053.

Lakkaraju, H, Aguiar, E, Shan, C, Miller, D, Bhanpuri, N, Ghani, R & Addison, KL. 2015. A machine learning framework to identify students at risk of adverse academic outcomes. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015-Augus:1909–1918. doi.org/10.1145/2783258.2788620.

Lee, S & Chung, JY. 2019. The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences (Switzerland)*. 9(15). doi.org/10.3390/app9153093.

Lewis, D, Madison-Harris, R, Muoneke, A & Times, C. 2010. Using Data to Guide Instruction and Improve Student Learning. *SEDL Letter*. XXII(Number 2):10–12. Available from: https://sedl.org/pubs/sedl-letter/v22n02/using-data.html.

Liang, J, Li, C & Zheng, L. 2016. Machine learning application in MOOCs: Dropout prediction. *ICCSE 2016 - 11th International Conference on Computer Science and Education*. (Iccse):52–57. doi.org/10.1109/ICCSE.2016.7581554.

Liao, SN, Zingaro, D, Thai, K, Alvarado, C, Griswold, WG & Porter, L. 2019. A robust machine learning technique to predict low-performing students. *ACM Transactions on Computing Education*. 19(3):1–19. doi.org/10.1145/3277569.

Liñán, LC & Pérez, ÁAJ. 2015. Educational data mining and learning analytics: Differences, similarities, and time evolution. *RUSC Universities and Knowledge Society Journal*. 12(3):98–112. doi.org/10.7238/rusc.v12i3.2515.

Liu, DYT, Atif, A, Froissard, JC & Richards, D. 2019. An enhanced learning analytics plugin for Moodle: Student engagement and personalised intervention. *ASCILITE 2015 - Australasian Society for Computers in Learning and Tertiary Education, Conference Proceedings*. (December):180–189.

Liz-Domínguez, M, Caeiro-Rodríguez, M, Llamas-Nistal, M & Mikic-Fonte, F. 2019. Predictors and early warning systems in higher education — A systematic literature review. *CEUR Workshop Proceedings*. 2415:84–99.

Lu, O, Huang, A, Huang, J, Lin, A, Ogata, H & Yang, S. 2018. International Forum of Educational Technology & Society Applying Learning Analytics for the Early Prediction of Students' Academic Performance in Blended Learning. *Source: Journal of Educational Technology & Society*. 21(2):220–232. Available from: https://www.jstor.org/stable/26388400?seq=1&cid=pdf-reference#references_tab_contents.

Macfadyen, LP & Dawson, S. 2010. Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers and Education*. 54(2):588–599. doi.org/10.1016/j.compedu.2009.09.008.

Mainali, S, Garzon, M, Venugopal, D, Jana, K, Yang, CC, Kumar, N, Bowman, D & Deng, LY. 2021. An Information-theoretic approach to dimensionality reduction in data science. *International Journal of Data Science and Analytics*. 12(3):185–203. doi.org/10.1007/s41060-021-00272-2.

Mamcenko, J. 2018. Comparative Analysis of Personalisation Approaches and Tools To Improve Learning Used in Different Learning Systems. *INTED2018 Proceedings*. 1(March):4674–4684. doi.org/10.21125/inted.2018.0917.

Mamcenko, J & Kurilovas, E. 2017. On using learning analytics to personalise learning in virtual learning environments. *Proceedings of the European Conference on e-Learning, ECEL*. 2010-Octob:353–361.

Marapelli, B. 2020. AApplication of Ensemble Machine Learning Methods to Improve Effort Prediction Accuracy. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME*. 9(02). doi.org/10.1016/j.procs.2020.02.168.

Marbouti, F, Diefes-dux, HA & Madhavan, K. 2016. Computers & Education Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education*. 103:1–15. doi.org/10.1016/j.compedu.2016.09.005.

Marbouti, F, Diefes-Dux, HA & Madhavan, K. 2016. Models for early prediction of at-risk students in a course using standards-based grading. *Computers and Education*. 103:1–15. doi.org/10.1016/j.compedu.2016.09.005.

Marques, J, Hobbs, D & Graf, S. 2014. Integrating an at-risk student model into learning management systems. *Nuevas Ideas en Informática Educativa TISE*. 120–124.

Márquez-Vera, C, Cano, A, Romero, C, Noaman, AYM, Mousa Fardoun, H & Ventura, S. 2016. Early dropout prediction using data mining: A case study with high school students. *Expert Systems*. 33(1):107–124. doi.org/10.1111/exsy.12135.

Martin, F & Bolliger, DU. 2018. Engagement matters: Student perceptions on the importance of engagement strategies in the online learning environment. *Online Learning Journal*. 22(1):205–222. doi.org/10.24059/olj.v22i1.1092.

Martin, F & Ndoye, A. 2016. Using Learning Analytics to Assess Student Learning in Online Courses. *Journal of Theoretical and Applied Information Technology*. 13(7).

Matarirano, O, Jere, NR, Sibanda, HS & Panicker, M. 2020. Antecedents of Blackboard Adoption by Lecturers at a South African Higher Education Institution – Extending GETAMEL. *International Journal of Emerging Technologies in Learning*. 16(1):60–79. doi.org/10.3991/IJET.V16I01.16821.

Matcha, W, Gašević, D, Uzir, NA, Jovanović, J & Pardo, A. 2019. Analytics of learning strategies: Associations with academic performance and feedback. *ACM International Conference Proceeding Series*. 461–470. doi.org/10.1145/3303772.3303787.

Mgala, M & Mbogho, A. 2015. Data-Driven Intervention-Level Prediction Modeling for Academic Performance. *ACM International Conference Proceeding Series*. 15(May).

doi.org/10.1145/2737856.2738012.

Mingyu, Z, Sutong, W, Yanzhang, W & Dujuan, W. 2022. An interpretable prediction method for university student academic crisis warning. *Complex and Intelligent Systems*. 8(1):323–336. doi.org/10.1007/s40747-021-00383-0.

Mothukuri, UK, Reddy, BV, Reddy, PN, Gutti, S, Mandula, K, Parupalli, R, Murty, CHAS & Magesh, E. 2017. Improvisation of learning experience using learning analytics in eLearning. *Proceedings - 2017 5th National Conference on E-Learning and E-Learning Technologies, ELELTECH 2017*. doi.org/10.1109/ELELTECH.2017.8074995.

Moubayed, A, Injadat, M, Nassif, AB, Lutfiyya, H & Shami, A. 2018. E-Learning: Challenges and Research Opportunities Using Machine Learning Data Analytics. *IEEE Access*. 6:39117–39138. doi.org/10.1109/ACCESS.2018.2851790.

Nácher, MJ, Badenes-Ribera, L, Torrijos, C, Ballesteros, MA & Cebadera, E. 2021. The effectiveness of the GoKoan e-learning platform in improving university students' academic performance. *Studies in Educational Evaluation*. 70. doi.org/10.1016/j.stueduc.2021.101026.

Nagy, M & Molontay, R. 2018. Predicting Dropout in Higher Education Based on Secondary School Performance. *INES 2018 - IEEE 22nd International Conference on Intelligent Engineering Systems, Proceedings*. 000389–000394. doi.org/10.1109/INES.2018.8523888.

Newland, B & Trueman, P. 2017. Learning Analytics in UK HE 2017. *HE Report on Learning Analytics*.

Nguyen, A, Wandabwa, H, Rasco, A & Le, LA. 2021. A Framework for Designing Learning Analytics Information Systems. *Proceedings of the 54th Hawaii International Conference on System Sciences*. (January). doi.org/10.24251/hicss.2021.002.

Nitu, M, Dascalu, M-I, Lazarou, E, Trifan, EL & Bodea, C-N. 2018. The 14 th International Scientific Conference eLearning and Software for Education Intelligent Education Assistant Powered by Chatbots. 12753.

Odhiambo Omuya, E, Onyango Okeyo, G & Waema Kimwele, M. 2021. Feature Selection for Classification using Principal Component Analysis and Information Gain. *Expert Systems with Applications*. 174(February):114765. doi.org/10.1016/j.eswa.2021.114765.

OECD. 2019. *Education at a Glance 2019 (Summary in Spanish)*. doi.org/10.1787/f6dc8198-es.

Oh, S. 2022. Predictive case-based feature importance and interaction. *Information Sciences*. 593:155–176. doi.org/10.1016/j.ins.2022.02.003.

Opitz, D & Maclin, R. 1999. Popular Ensemble Methods : An Empirical Study. *Journal of artificial intelligence research*. 11(July):169–198.

Oreshin, S, Filchenkov, A, Petrusha, P, Krasheninnikov, E, Panfilov, A, Glukhov, I, Kaliberda, Y, Masalskiy, D, et al. 2020. Implementing a Machine Learning Approach to Predicting Students Academic Outcomes. *ACM International Conference Proceeding Series*. 78–83. doi.org/10.1145/3437802.3437816.

Ornelas, F & Ordonez, C. 2017. Predicting Student Success: A Naïve Bayesian Application to Community College Data. *Technology, Knowledge and Learning*. 22(3):299–315. doi.org/10.1007/s10758-017-9334-z.

Osei, CK. 2010. Perceptions of students towards use of distance learning: The case in an executive masters business program in Ghana. *Online Journal of Distance Learning Administration*. 13(2):26–31.

Pang, Y, Judd, N, O'Brien, J & Ben-Avie, M. 2017. Predicting students' graduation outcomes through support vector machines. *Proceedings - Frontiers in Education Conference, FIE*. 2017-Octob:1–8. doi.org/10.1109/FIE.2017.8190666.

Pellagatti, M, Ieva, F & Paganoni, AM. 2021. Generalized mixed-effects random forest : A flexible approach to predict university student dropout. (December 2020):241–257. doi.org/10.1002/sam.11505.

Pudjihartono, N, Fadason, T, Kempa-Liehr, AW & O'Sullivan, JM. 2022. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Frontiers in Bioinformatics*. 2(June):1–17. doi.org/10.3389/fbinf.2022.927312.

Rajendran, S, Sinha, AA & Chamundeswari, S. 2021. Predicting Factors Impacting Student Academic Performance using Machine Learning Algorithms. *SSRN 3898302*.

Raju, D. 2012.

Ramaswami, G, Susnjak, T & Mathrani, A. 2022. On Developing Generic Models for Predicting

Student Outcomes in Educational Data Mining. *Big Data and Cognitive Computing*. 6(1). doi.org/10.3390/bdcc6010006.

Romero, C & Ventura, S. 2020. Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 10(3):1–21. doi.org/10.1002/widm.1355.

Romero, C, Espejo, PG, Zafra, A, Romero, JR & Ventura, S. 2010. Web Usage Mining for Predicting Final Marks of Students That Use Moodle Courses. *Computer Applications in Engineering Education*. 21(1):135–146. doi.org/10.1002/cae.20456.

Sangodiah, A, Tunku, U, Rahman, A, Beleya, P, Tunku, U, Rahman, A, Muniandy, M, Tunku, U, et al. 2015. Minimizing student attrition in higher learning institutions in Malaysia using support vector machine MINIMIZING STUDENT ATTRITION IN HIGHER LEARNING INSTITUTIONS IN MALAYSIA USING SUPPORT VECTOR MACHINE. (October).

Sciarrone, F. 2018. Machine learning and learning analytics: Integrating data with learning. *2018 17th International Conference on Information Technology Based Higher Education and Training, ITHET 2018*. doi.org/10.1109/ITHET.2018.8424780.

Siemens, G & Latour, B. 2013. Learning Analytics: The Emergence of a Discipline The slightest move in the virtual landscape has to be paid for in lines of code. *American Behavioral Scientist*. 57(10):1380–1400. doi.org/10.1177/0002764213498851.

Siri, A. 2015. Predicting S tudents ' Dropout at University Using Artificial Neural Networks Predicting S tudents ' Dropout at University Using Artificial Neural Networks. 7(June):225–247.

Stapel, M, Zheng, Z & Pinkwart, N. 2016. An Ensemble Method to Predict Student Performance in an Online Math Learning Environment. International Educational Data Mining Society. 231–238.

Tahiru, F & Agbesi, S. 2021. The Future of Artificial Intelligence in Education. In: *Digital Technology Advancements in Knowledge Management*. IGI Global. 187–194. doi.org/10.4018/978-1-7998-6792-0.ch010.

Taskin, G, Kaya, H & Bruzzone, L. 2017. Feature selection based on high dimensional model representation for hyperspectral images. *IEEE Transactions on Image Processing*. 26(6):2918–

2928. doi.org/10.1109/TIP.2017.2687128.

“Technical Specification Document”. n.d. 0–73.

Tenpipat, W & Akkarajitsakul, K. 2020. Student Dropout Prediction: A KMUTT Case Study. *2020 1st International Conference on Big Data Analytics and Practices, IBDAP 2020*. doi.org/10.1109/IBDAP50342.2020.9245457.

Tsai, YS, Rates, D, Moreno-Marcos, PM, Muñoz-Merino, PJ, Jivet, I, Scheffel, M, Drachsler, H, Delgado Kloos, C, et al. 2020. Learning analytics in European higher education—Trends and barriers. *Computers and Education*. 155(February):103933. doi.org/10.1016/j.compedu.2020.103933.

UNESCO. 2012. Available from: http://www.unesco.org/new/en/member-states/single-view/news/42_of_african_school_children_will_drop_out_before_the_end/.

UNICEF. 2020. Ghana Education Fact Sheets 2020 Analyses for learning and equity. 1–56.

Upoalkpajor, J-LN & Upoalkpajor, CB. 2020. The Impact of COVID-19 on Education in Ghana. *Asian Journal of Education and Social Studies*. 23–33. doi.org/10.9734/ajess/2020/v9i130238.

Uzir, NAA, Gaševic, D, Jovanovic, J, Matcha, W, Lim, LA & Fudge, A. 2020. Analytics of time management and learning strategies for effective online learning in blended environments. *ACM International Conference Proceeding Series*. 392–401. doi.org/10.1145/3375462.3375493.

Valecha, H. 2018. Prediction of Consumer Behaviour using Random Forest Algorithm.

Valle, N, Antonenko, P, Valle, D, Dawson, K, Huggins-Manley, AC & Baiser, B. 2021. The influence of task-value scaffolding in a predictive learning analytics dashboard on learners’ statistics anxiety, motivation, and performance. *Computers and Education*. 173(November 2020):104288. doi.org/10.1016/j.compedu.2021.104288.

Viberg, O, Hatakka, M, Bälter, O & Mavroudi, A. 2018. The current landscape of learning analytics in higher education. *Computers in Human Behavior*. 89(October 2017):98–110. doi.org/10.1016/j.chb.2018.07.027.

Wandera, H, Marivate, V & Sengeh, MD. 2019. Predicting school performance using a combination of traditional and non-traditional education data from South Africa. *Knowledge 4 All*. Available from: https://www.k4all.org/wp-content/uploads/2019/07/Predicting-school-

performance-using-a-combination-of-traditional-and-non-traditional-education-data-from-South-Africa.pdf.

Wang, Z, Zhu, C, Ying, Z, Zhang, Y, Wang, B, Jin, X & Yang, H. 2019. Design and Implementation of Early Warning System Based on Educational Big Data. *2018 5th International Conference on Systems and Informatics, ICSAI 2018*. (Icsai):549–553. doi.org/10.1109/ICSAI.2018.8599357.

Webster, R, Andre, J & Giang, TTT. 2019. Industry 4.0 and higher education: Combining learning analytics and learning science to transform the undergraduate learning experience in Vietnam. *International Conference: Leadership and Management on Higher Education: Driving Change with Global Trends*. (July 2019):1–12.

Whitten, LS, Clarksville, T, Sanders, AR & Stewart, JG. 2013. Degree Compass: The Preferred Choice Approach. *Journal of Academic Administration in Higher Education*. 39.

Xu, X, Wang, J, Peng, H & Wu, R. 2019. Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior*. 98(January):166–173. doi.org/10.1016/j.chb.2019.04.015.

Yerel, R, Dagli, G, Altinay, F, Ossiannilsson, E, Altinay, M & Altinay, Z. 2021. Sustainability in education: A scale on perceptions of organisational discipline related to the covid-19 period. *Sustainability (Switzerland)*. 13(15):1–25. doi.org/10.3390/su13158343.

Yu, CH, Lee, HS, Lara, E & Gan, S. 2018. The ensemble and model comparison approaches for big data analytics in social sciences. *Practical Assessment, Research and Evaluation*. 23(17).

Zafra, A & Ventura, S. 2009. Predicting student grades in learning management systems with multiple instance genetic programming. *EDM'09 - Educational Data Mining 2009: 2nd International Conference on Educational Data Mining*. (Mil):307–314.

Zemel, R. 2014. Ensemble Methods from University of Toronto CSC411 Machine Learning & Data Mining. 41. Available from: http://www.cs.toronto.edu/~rsalakhu/CSC411/notes/lecture_ensemble1.pdf.

7.1 APPENDIX A

Table of the dataset variable.

The table below describes variables in the various tables

7.1.1 List of Tables features

Table 10:List of features in courses table

Course Table contains the list of all available modules and their presentations. It consists of 22 rows with the following columns:

Column name	Datatype	features	sample feature	description
Code_module	object	categorical	AAA-GGG	course/module labels
Code_presentation	object	categorical	2013B, 2013J	code representing when a student registered for the module
length	int64	numerical	260	the duration(start to end) for the course represented in days

Table 11:List of features in the Assessment table

Assessment Table contains information about assessments in module presentations, and every presentation has several assessments followed by the final exam**.** The table consists of 206 rows with the following columns:

Column name	Datatype	features	sample feature	description
code_module	object	categorical	AAA-GGG	course label
code_presentation	object	categorical	2013B, 2013J	code representing when a student registered for the module.
id_assessment	int64	numerical	1752	assessment ID num
assessment_type	object	categorical	TMA	the type of assessment (TMA-Tutor marked assessment, CMA-Computer marked assessment, Exam -Final Exam)
date	float64	numerical	10	represent the cut-off day of the assessment.
Weight	float64	numerical	100	the weight of the assessment. Exams have a weight equal to 100%, and the sum of all other assessments is also 100%.

Table 12:List of features in the Vle table

Vle Table contains information about the materials available in the VLE. It consists of 6,364 rows with the following columns.

Column name	Datatype	features	sample feature	description
id_site	int64	numerical	546652	an ID number of the material
code_module	object	categorical	AAA-GGG	course label
code_presentation	object	categorical	203B, 2013J	code representing when a student registered for the module
activity_type	object	categorical	homepage	the type of materials associated with the module.
week_from	float64	numerical	NaN	the week from which the material is planned to be used.
week_to	float64	numerical	NaN	the week until which the material is planned to be used.

Table 13:List of features in the student Information table

Student Information Table contains student demographic information and also their results in each module they studied. It consists of 32,593 rows with the following columns:

Column name	Datatype	features	sample feature	description
code_module	object	categorical	AAA-GGG	course label
code_presentation	object	categorical	2013B,2013J	code representing when a student registered for the module
id_student	int64	numerical	24213	unique ID
gender	object	categorical	F/M	student's gender
region	object	categorical	Scotland	geographical region of students when taking the course
highest_education	object	categorical	A-level or Equivalent	highest level of education before enrolling in the course
imd_band	object	categorical	0-10%	place student lived whiles taking the course.
age_band	object	categorical	35-55	student age-band (range)
num_of_prev_attempts	int64	numerical	0	the number of times student's have attempted the module.
studied_credits	int64	numerical	240	total number credits of module student studied.
disability	object	categorical	Y/N	represents whether a student has a disability or not.
final_result	object	categorical	Withdrawn	student's final result in the module presentation.

Table 14:List of features in Student Assessment table

Student Assessment Table The table contains the results of students' assessments. If the student does not submit the assessment, no result is recorded. Results of the final exam are usually missing (since they are scored and used for the final marking immediately at the end of the module). It consists of 173912 rows with the following columns.

Column name	Datatype	features	sample feature	description
id_assessment	int64	numerical	1752	the assessment ID number
id_student	int64	numerical	24213	unique student ID
date_submitted	int64	numerical	18	the day of assessment submission.
is_banked	int64	numerical	0	a status indicating assessment transferred from the previous presentation.
score	float64	numerical	78	student score in the assessment. Scores range from 0 -100 and a score below 40 is Fail.

Table 15:List of features in student Vle table

studentVle Table contains information about students' interactions with the VLE. It consists of 10,655,280 rows with the following columns.

Column name	Datatype	features	sample feature	description
code_module	object	categorical	AAA-GGG	course label
code_presentation	object	categorical	2013B, 2013J	code representing when a student registered for the module.
id_student	int64	numerical	24213	unique student number
id_site	int64	numerical	546652	Vle identification number
date	int64	numerical	-10	the day student interacted with vle.
sum_click	int64	numerical	656	the number of times the student interacted with the material.

Table 16:List of features in Registration Table

Student Registration Table consists of 32,593 rows with the following columns

Column name	Datatype	features	sample feature	description
code_module	object	categorical	AAA-GGG	course labels
code_presentation	object	categorical	2013B,2013J	code representing when a student registered for the module
id_student	int64	numerical	24213	unique student number
date_registration	float64	numerical	-53	day student registered for the module presentation.
date_unregistration	float64	numerical	12	day student unregistered for the module.

	gender_F	gender_M	imd_band_0-10%	imd_band_10-20	imd_band_20-30%	imd_band_30-40%	imd_band_40-50%	imd_band_50-60%	imd_band_60-70%	imd_band_70-80%	imd_band_80-90%
)	0	1	0	0	0	0	0	0	0	0	1
I	0	1	0	0	0	1	0	0	0	0	0
3	0	1	0	0	0	0	0	0	0	0	0
4	1	0	0	0	1	0	0	0	0	0	0
5	1	0	0	0	0	0	0	1	0	0	0

Figure 35: Feature encoding

Out[31]:

	date_registered	date_unregistered	date_submitted	score	id_student	sum_click	date	code_presentation	studied_credits	num_of_prev_attempts	...	i
0	-52.0	76.257935	112.0	309.0	6516.0	2791.0	110.0	3	60.0	0.0	...	
1	-88.0	68.000000	23.0	609.0	8462.0	656.0	37.0	1	90.0	0.0	...	
3	-47.0	76.257935	56.0	330.0	11391.0	934.0	102.0	1	240.0	0.0	...	
5	-27.0	76.257935	120.0	1033.0	23698.0	910.0	86.0	3	120.0	0.0	...	
6	-25.0	76.257935	161.0	500.0	23798.0	590.0	110.0	1	60.0	0.0	...	
...	...	...	...	...	...	...	...	...	...	...	...	
26068	-38.0	76.257935	169.0	944.0	694380.0	253.0	14.0	3	60.0	0.0	...	
26069	-23.0	76.257935	121.0	407.0	694384.0	324.0	104.0	3	60.0	0.0	...	
26071	-115.0	180.000000	90.0	314.0	694416.0	26.0	65.0	3	30.0	0.0	...	
26072	16.0	76.257935	106.0	322.0	694428.0	2014.0	130.0	3	60.0	0.0	...	
26073	4.0	76.257935	107.0	462.0	694448.0	81.0	15.0	3	60.0	0.0	...	

17488 rows × 31 columns

Figure 36: Total input for Model development

7.1.2 Summary of Modules and withdrawn

Table 17: List of student code presentations

Class	Module	No. of students	Number of withdrawn students	Withdrawn %
Social Science	AAA	748	126	16.84
	BBB	7909	2388	30.19
	GGG	2534	292	11.52
STEM	CCC	4434	1975	44.54
	DDD	6272	2250	35.87
	EEE	2934	722	24.60
	FFF	7762	2403	30.95
		32593	10156 (31.16%)	

7.1.3 Summary of Module and Presentation Information

Table 18:Modules and code presentation in dataset

Module	Presentations		Total number	
AAA	2	2013J	383	748
		2014J	365	
BBB	4	2013B	1767	7909
		2013J	2237	
		2014B	1613	
		2014J	2292	
CCC	2	2014B	1936	4434
		2014J	2498	
DDD	4	2013B	1303	6272
		2013J	1938	
		2014B	1228	
		2014J	1803	
EEE	3	2013J	1052	2934
		2014B	694	
		2014J	1188	
FFF	4	2013B	1614	7762
		2013J	2283	
		2014B	1500	
		2014J	2365	
GGG	3	2013J	952	2534
		2014B	833	
		2014J	749	

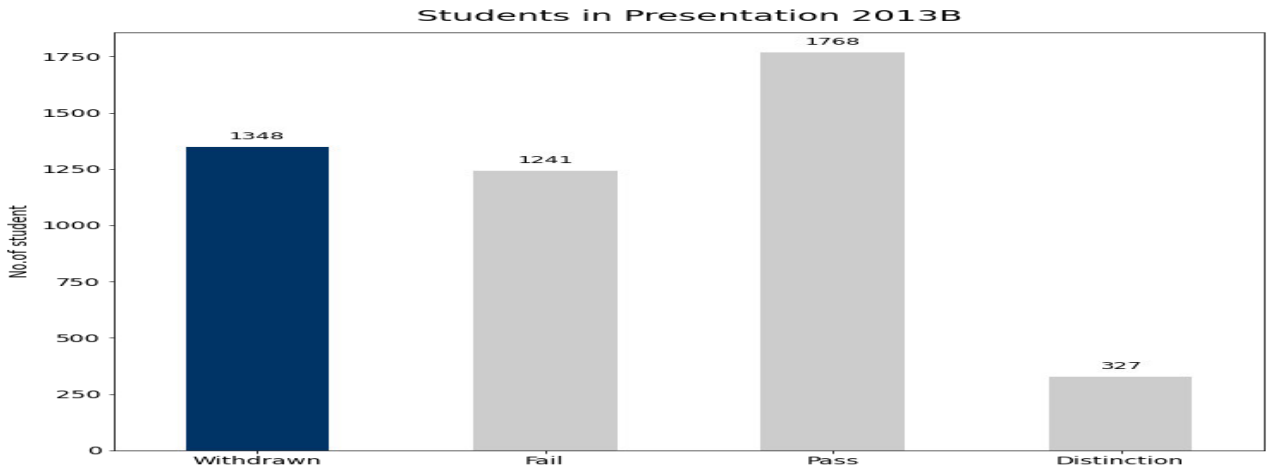

Figure 37:Final results of students in code presentation 2013B

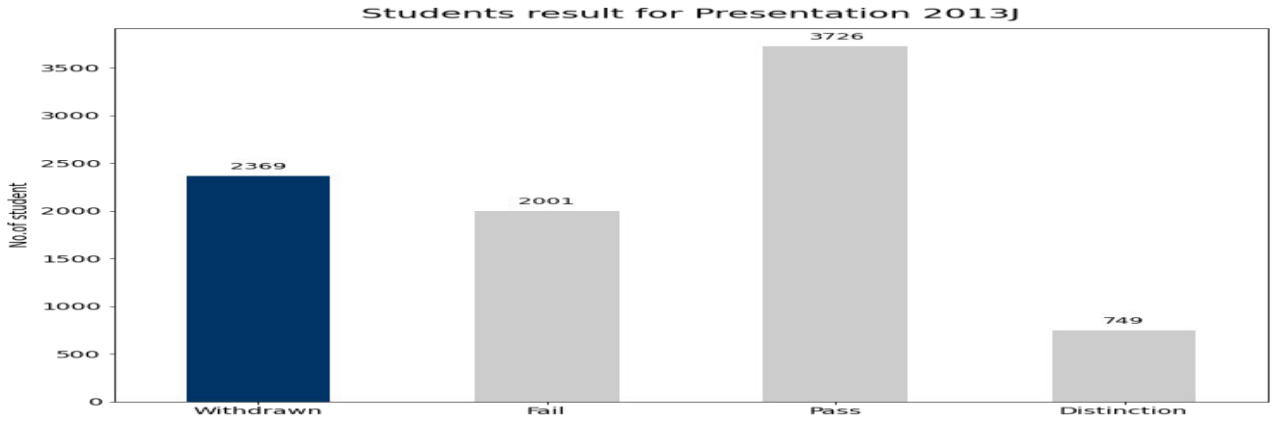

Figure 38:Final results of students in code presentation 2013J

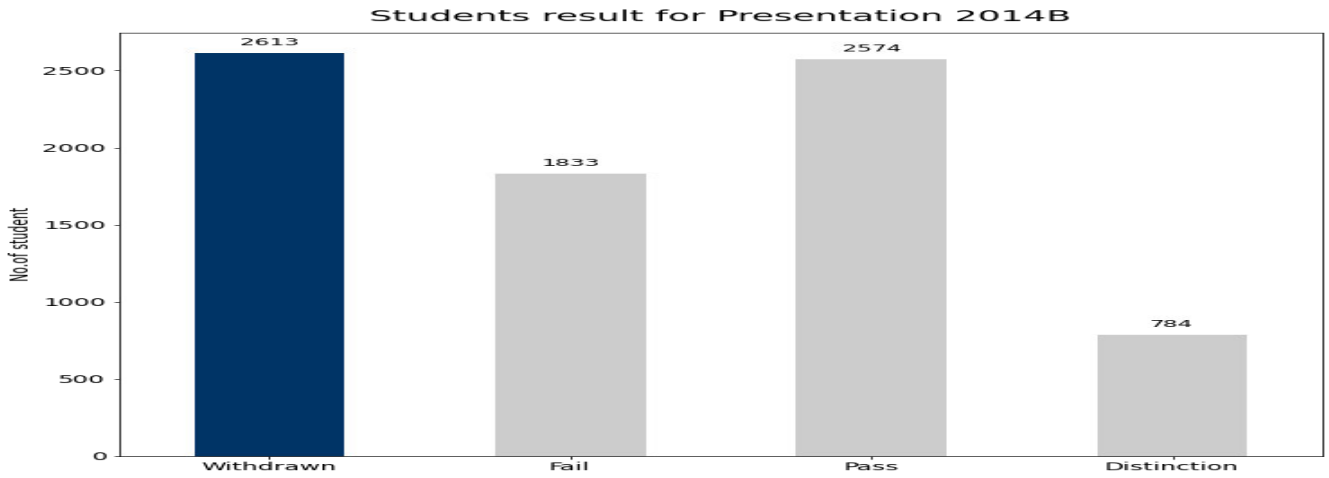

Table 19:Final results of students in code presentation 2014B

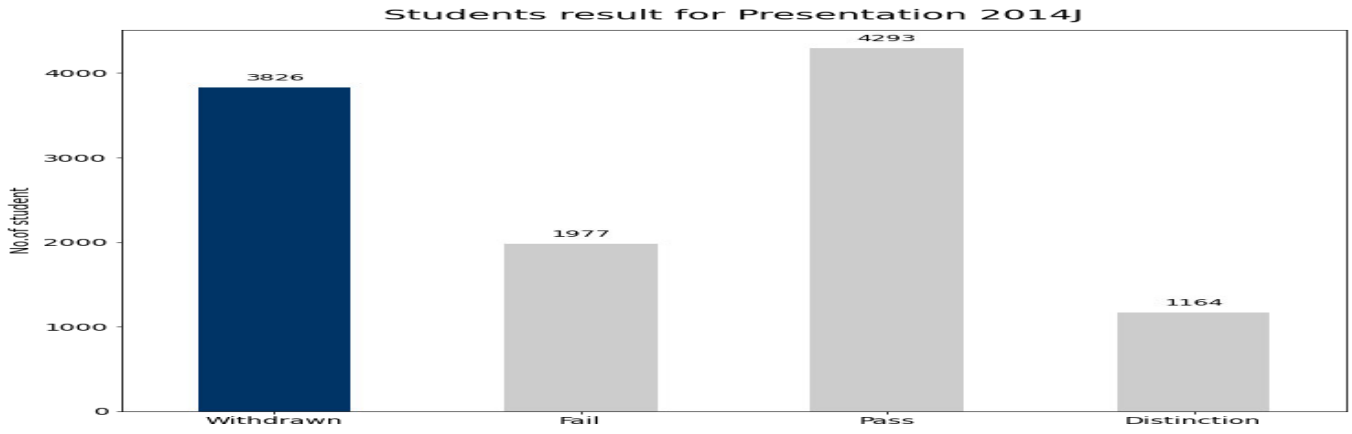

Figure 39:Final results of students in code presentation 2014J

7.2 APPENDIX B

Journal Paper (Full paper enclosed)

- Machine Learning-based predictive systems in higher education: A bibliometric analysis.

Conference paper (Full paper enclosed)

- Using an exploratory analytical approach to distinguish the habits of graduating and non-graduating students in a virtual learning environment.