A MODEL FOR EVALUATING THE EFFICACY OF ELEARNING IN HIGHER EDUCATIONAL INSTITUTIONS USING EDUCATIONAL DATA MINING

SUBMITTED BY:

GEORGE N. KANGETHE

REG NO: 20/00514

A RESEARCH PROJECT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE IN DATA ANALYTICS IN THE SCHOOL OF TECHNOLOGY AT KCA UNIVERSITY

2022

DECLARATION

I declare that this research project is my original work and has not been previously published or submitted elsewhere for award of a degree. I also declare that this contains no material written or published by other people except where due reference is made, and author duly acknowledged.

Student Name: George Njenga Kangethe Reg No: 20/00514

Sign:  Date: June 28th, 2022

This proposal has been submitted for examination with my approval as the appointed university supervisor.

Sign:  Date: 28/06/2022

Dr. Lucy W. Waruguru

ABSTRACT

Educational Data Mining (EDM) and Learning Analytics (LA) play a key role in developing methods for discovering student learning patterns and behaviors by interrogating this robust set of data now available in learning environments. The main objective of this study is to develop a model for evaluating efficacy of eLearning at Higher Educational Institutions (HEI's). To measure the efficacy of eLearning, data on student activity within eLearning LMS and student academic performance is analyzed. In this study, Orange data mining tool is used for the analysis of the data. Support Vector Machine, Random Forest, Decision Tree, Nave Bayes, Logistic Regression, and Neural Network are among the categorization techniques provided within Orange. These classifiers are compared based on their accuracy. The selected classifiers are evaluated against a k-fold cross validation, accuracy, precision, recall, and F-score. According to the empirical findings, the Support Vector Machine (SVM) algorithm was the best data mining model for estimating students' academic achievement.

Keywords: Educational Data Mining, eLearning, Data Mining, Learning Management Systems

ACKNOWLEDGMENT

First and foremost, I would like to thank God Almighty for giving me the strength, good health, will, and for sustaining me throughout my course of study.

I especially would like to thank my family, and specifically my wife Evalyn Njenga for her belief in me and her tireless encouragement, our daughter Ashley Njenga, and sons Austin Njenga and Azriel Njenga, who sacrificed my usual attention and parental care during my study. This work is dedicated to them for the sacrifice, care, love, concern, encouragement, support, and enthusiasm.

I want to thank my supervisor Dr. Lucy Mburu, who guided, encouraged, and led me though the journey to completing this dissertation. I also thank Mr. Dennis Munene who read and reviewed my work and provided constructive comments that helped to shape my work.

I want to express my gratitude to everyone who, in some way or another, helped me to attempt and complete this project.

ACRONYMS AND ABBREVIATIONS

Moodle	Modular Object-Oriented Dynamic Learning Environment
LMS	Learning Management System
EDM	Educational Data Mining
LA	Learning Analytics
ODL	Open and Distance Learning
SARS	Severe Acute Respiratory Syndrome
ICT	Information Communication Technology
DLP	Digital Literacy Programme
MOOC	Massive Open Online Course
ITS	Intelligent Tutoring System
ECLAT	Equivalence Class Clustering and bottom-up Lattice Traversal
KDD	Knowledge Discovery in Databases
CAL	Computer-Aided Learning
HEI	Higher Educational Institutions
SIS	Student Information System
SVM	Support Vector Machine
kNN	k-Nearest Neighbor
TML	Technology Mediate Learning
TAM	Technology Acceptance Model

GLOSSARY

Efficacy

1. The power to produce an effect. (Merriam-Webster Dictionary)
2. The ability, especially of a medicine or a method of achieving something, to produce the intended result (Cambridge dictionary)
3. The quality of being effective; effectiveness (Cambridge dictionary)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1. Background of The Study

The education sector in Kenya has over the years been shifting to digital learning. The government has rolled out projects that aim to bolster this shift with the electrification of rural schools and the Digital Literacy Programme (DLP). The fundamental goal of these programs is to standardize the use of ICT in teaching and learning across all schools. Ronoh, P. K. (2021). To support the digital learning initiative, the government has developed digital content, built capacity in teachers, and procured relevant ICT devices ("Digital Content – DigiSchool – ICT Authority", 2013).

According to Waema (2005), Kariuki (2009), the national ICT policy for Kenya in consideration of the development and utilization of eLearning, lays the pivotal structure for eLearning. Also, ICT has been identified as a critical instrument for teaching and learning by Kenya's Ministry of Education Policy Framework for Education and Training (2012). Priority areas that have been identified within the policy framework include open and distance learning (ODL) and eLearning. Strategies laid out in the policy framework include the establishment the Open University of Kenya and pushing for expansion of ODL and eLearning in existing universities. This expansion is envisaged to be driven through leveraging advanced ICT capabilities and by taking advantage of the improved ICT infrastructure within the country.

The ongoing COVID-19 pandemic has caused a major shift in how learning is delivered. Many of our schools were instantaneously forced to implement measures for students to continue learning remotely, distance, or online while at home due to the restrictions brought about by the pandemic, World Bank. (2020). Rather than trying to recreate school, the education sector had to

shift to emergency remote learning while leveraging the assets of home-based learning while at the same time adapting to and adopting new learning and teaching paradigms, Aseey et al. (2020). Opponents of eLearning hold that the fortuitous and sudden move to eLearning – with little or no resources – will ultimately give rise to substandard user experience, that will hamper any sustained growth, proponents on the other hand maintain that a new hybrid educational model will emerge with notable gains (Li & Lalani, 2020).

In a comparison of eLearning against traditional in class learning, Titthasiri (2013) found that difference between eLearning and traditional learning was not statistically significant. Rashty (2003) affirms this finding that eLearning is as good as traditional learning. A study by Bencheva (2010) acknowledged the advantages and disadvantages for both modes of learning and concluded that there is no finding that supports the superiority of traditional in classroom learning.

This recent push to eLearning has only accelerated the growth of student data which educational institutions hold and is largely due to the entrenchment of digital learning platforms popularly known as Learning Management Systems (LMS). These LMS's accumulate large amounts of data about student activities known as log data. Log data is a record of the activities a student is engaged in, such as taking assessments, downloading reading material, uploading assignments, or even communication with other students. LMSs also include a database that contains information about the system, users details, academic results, and important interaction data. These LMS's however lack in providing tools for analyzing this vast amount of data. This is where EDM becomes important. Use of EDM methods in analyzing educational data is a very promising research area.

While there have been numerous studies on the use of data mining in higher educational institutions in the global space, research in the local arena has largely focused on distance learning

and the general technology readiness. In the Kenyan context, there have been a lot of studies around technology enhanced learning, challenges of eLearning and eLearning but little on educational data mining of eLearning generated data. Kashorda et al. (2007) is one such example of studies addressing the general technology readiness of higher education institutions (HEI's) in Kenya. The case study of University of Nairobi by Oketch (2013) also focuses on eLearning readiness. Kibuku et al. (2020), conducted a literature review on the eLearning challenges faced by universities in Kenya.

A look at statistics provided by Moodle (Modular Object-Oriented Developmental Learning Environment) Registered sites (2021) suggest that this is the most widely adopted LMS within higher education institutions in Kenya. Moodle boasts a user base of 641 learning institutions in Kenya, this is largely driven by the fact that Moodle is free for use and is distributed as an open-source general public license platform.

1.2. Statement of the Problem

The benefits of eLearning are indisputable, these include customized learning experience, flexibility, efficient and effective communication, scalability, cost-effectiveness, ease of access to information, and extended geographical access to education. Whilst there have been an innumerable number of studies adequately addressing the adoption of eLearning in higher educational institutions in Kenya, there is little research on the efficacy of these eLearning systems once adopted.

One of the most desired characteristics of any learning environment is its effectiveness, efficiency, and capacity to engage the student. Efficacy involves more than just adoption of eLearning; it includes harnessing of the data generated by the eLearning systems. A conclusive

study on efficacy should look at how eLearning has been adopted, student engagement with the eLearning system itself and how students have performed while engaged in eLearning. It is imperative that we interrogate the data from eLearning systems to measure how the effective use of the system affects the efficacy of eLearning.

A look at previous studies that sought to investigate the efficacy of eLearning found that these studies implemented models that relied on primary data such as questionnaires without applying secondary data extracted from eLearning systems. Other studies have addressed the challenge of eLearning effectiveness by introducing context aware eLearning systems. Context aware eLearning systems focus on personalization and adaptation of the learning content based on some knowledge about the learner (personalization principle, Mayer and Mayer (2005)).

In recent years, context aware eLearning systems, also known as adaptive learning management systems, have become more popular. However, striking a balance of the two major approaches in adaptive learning systems has proven difficult for these adaptive LMSs. The two techniques differ in that one focuses on tailoring learning content to a learner's specific needs (learner directed), while the other focuses on delivering learning content in the most appropriate order based on the learners needs (system controlled), Yaghmaie et al. (2011). Current studies on context aware systems are so focused on adaptation quality, the researchers determined, that they result in systems that are designated for unique learning purposes and are not operable with other systems. It is therefore difficult to generalize the effectiveness of eLearning through the study of these purpose-built context aware systems.

Liaw (2008) investigated the effectiveness of eLearning by considering learners self-efficacy, multimedia formats, and interaction environments. This study measured effectiveness using Likert

questionnaires distributed to 560 university students. The questionnaires focused on the three (3) considerations mentioned earlier and did interrogate the raw data within the system. While it is a good study, the use of questionnaires to measure eLearning efficacy can be subjective because responses from individuals can vary over time. The study relied on learner perception of usefulness and satisfaction and did not implement machine learning approaches to measure how various variables affect efficacy.

Araka et al. (2019) proposed a conceptual model for measuring and promoting Self-Regulated Learning (SRL) that is based on EDM. Araka et al. (2019), developed a model that uses EDM to enhance personalization and strengthen learning and teaching theories. This model was applied on LMS datasets to draw out learner's patterns which can be used to buttress SRL. While this study interrogated raw data from eLearning systems, the focus was on variables that influence SRL, ways to measure SRL and create eLearning interventions that promote the development of SRL skills. However, the study did not implement machine learning approaches to measure how various variables affect efficacy.

Romero et al. (2008) used various methods to analyze Moodle usage data in a case study tutorial that included a survey of the application of data mining in LMS's. As a survey paper, they did not focus on solving a specific problem but to introduce both a theoretical and practical way to perform Educational Data Mining. They concluded that it is important in the future to have data mining tools that are oriented specifically to eLearning environments.

Ogwoka et al. (2015) used data mining to predict student's performance but did not include data from eLearning platforms. Their research applied data obtained from the student management system which hosts the student's academic records.

In literature review, we identified a gap in studies applying Educational Data Mining Techniques within the Kenyan context. Despite the numerous research studies conducted on the topic of eLearning, the review of literature indicates that such studies have not addressed the efficacy of eLearning within Kenya institutions of higher learning. Most studies focused either on adoption of eLearning platforms or predicting students' performance based on student academic records and did not analyze the raw data from eLearning systems. An interrogation of peer-reviewed databases such as the IST Africa repository ("IST-Africa", 2021) of conference papers from Africa was conducted by providing the keyword 'educational data mining', this query returned only one (1) paper addressing this topic. A similar search with the keyword 'eLearning' revealed over twenty (20) papers addressing this topic. This study will extend the application of previous research and apply EDM on data collected from eLearning platforms such as Moodle LMS platform in use at higher educational institutions. The research will focus on the educational data generated by students while utilizing an LMS with video integration.

1.3. Main Objective

The main objective of this study is to develop a model to evaluate the efficacy of eLearning in higher educational institutions using educational data mining on a Learning Management System (LMS).

1.4. Specific Objectives

The specific objectives of this study are:

- To examine the factors that affect the efficacy of eLearning.
- To develop a model for determine the efficacy of eLearning based on students' engagement and academic achievement.
- To test and validate the model developed above.

1.5. Research Questions

This study attempts to address the following questions:

- What factors affect the efficacy of eLearning?
- What is the appropriate model for measuring the efficacy of eLearning?
- How valid is the developed model for application in measuring the efficacy of eLearning?

1.6. Significance of the Study

This study will extract data from the eLearning system for analysis and insight extraction. The developed model will be important in giving insights into student interaction with the eLearning systems such as time spent during assessments, how often a student logs onto the platform etc. The findings of this study could help decision-makers and educational policymakers to make use of the available knowledge to transform learning experiences, develop new models and accelerate the digital push by integrating eLearning as a fundamental component of school education. Potential stakeholders and groups of people who can leverage this knowledge include students, instructors, administrators/management, and curriculum developers.

Mining of educational data will assist curriculum developers in the process of deciding what should be taught and how learning should look like. Mabić et al. (2017) applied decision tree algorithm in mining educational data and presented how information gathered from this process can be used in curriculum development decisions such as determining the sequence of courses within the curriculum. According to the findings, there is a tremendous opportunity to apply data mining approaches to improve the quality of curriculum development. The study demonstrated how pre-determined rules might be applied to specific acts, such as establishing the sequence of courses within a program.

Fortino et al. (2019) applied text mining techniques in correlating specific jobs to specific degrees and courses with the aim of assisting students and curriculum developers in addressing the question, “which degrees and courses are relevant for specific jobs” or “for given courses taken by a student, what jobs are they most likely to succeed in” and “how current and proposed degree programs are aligned to specific job groups”.

Gaining insights into student behavior and how they learn can help educational administrators and management to improve current study programs, and the general education practice. Administrators can see detailed data for the entire organization and make decisions based on what works and what does not. Policy makers can also understand how students learn with specific interventions and how improvements to these interventions could be implemented. According to Bienkowski et al. (2012), administrators can set or adapt policies, and implement programs to improve certain metrics by using data from eLearning systems.

Insights from the LMS activity data will provide recommendations to both instructors and students about eLearning courses based on behavior. Students get to understand activities that lead to successful completion of courses while instructors get to track the learning process and adjust their instructional actions where low performance is noted, e.g., specific chapters. Bienkowski et al. (2012) suggests that eLearning platforms should offer personalized learning experiences similar to what Netflix does in the entertainment industry. These personalized experiences will be based on the data garnered from these systems where students can get recommendations of courses to enroll in.

1.7. Motivation of the Study

eLearning environments generate vast amounts of data linked to learning and teaching exercise, which presents the possibility of acquiring precious information that may be utilized to support education related decision making. Based on the data accumulated thus far since the beginning of the COVID-19 pandemic, we can make use of existing data analytics capabilities to gain insights into the performance of students during this period of remote learning. These insights will go a long way in aiding decision-making at the policymaking level and build the case for continued remote and online learning in a post-COVID-19 world.

Veneri, D. (2011) sought to review of literature pertaining to the use and effectiveness of Computer-Assisted Learning (CAL) in physical therapy education. The study found that remarkably, there are few studies available that address CAL use and effectiveness. The study concluded that CAL can effectively communicate material when compared to traditional methods of learning. The researchers urged for future studies to include larger and broader representations of the educational field.

Means e al. (2013) undertook an empirical review to analyze the efficacy of online and blended learning. The systematic review results of this study found that typically, students undertaking online learning performed reasonably better that those undertaking face-face instruction. Blended learning also provided an advantage in that students had more learning time, instructional materials, and course components that encouraged learner interaction, according to the study. The researchers recommended future experimental research addressing different kind of learners.

Noesgaard & Ørngreen (2015) completed an exploratory and integrative assessment of definitions, methodologies, and factors that promote eLearning effectiveness. The study found that majority of studies implemented pre and post tests to measure effectiveness of eLearning. The

study also developed a model for discerning the correlation of the key elements that affect effectiveness. The researchers identified three factors that influence effectiveness, the setting in which the eLearning system is employed, the eLearning system itself and the students prior experience, motivation and interactions with the eLearning system itself. The study made a final call to learning designers and researchers to focus their measurement efforts towards counting what matters to them and their stakeholders.

Not many studies have addressed this issue adequately, yet it is of national importance for Kenya to meet its educational targets under vision 2030. Despite the continued adoption of eLearning and use of various LMS among higher learning institutions in Kenya, there is little evidence that research studies have attempted to mine the data generated by these LMS's. Many of the studies conducted on eLearning within higher education institutions in Africa focus on users' perceptions obtained through surveys, this results in highly subjective interpretations on the effectiveness of eLearning in African higher education institutions. Statistics provided by Moodle stats website, there are currently 651 institutions in Kenya using Moodle ("Registered sites", 2021). This calls for future works to further investigate effectiveness of eLearning as indicated by the literature above serves as the motivation for this study.

1.8. Scope of the Study

This study aims to study the effectiveness of eLearning, particularly in higher education institutions.

1.9. Structure of the Research

The next chapters of this paper constitute literature review and methodology. In literature review we address previous studies that have attempted to investigate the problems statement from a

global and local perspective. In methodology, the framework and model used in this study are laid out. The appendices lay out the project budget and expected timelines.

CHAPTER TWO

LITERATURE REVIEW

2.1. Introduction

In this section a review of previous and related works that have undertaken to study the underlying focus of this study is presented. The aim of reviewing related literature is to form a basis for showing the effectiveness of eLearning as presented by other studies, highlight the commonly accepted models of mining educational data, and identify the various variables that have been found to affect efficacy of eLearning.

2.2. Theoretical Review

The increased growth in large educational datasets has led to the emergence of a broader selection of methods to generate insights from this data. This has spawned two general communities in the research field, educational data mining and learning analytics. Learning analytics uses educational data mining methods to analyze these large datasets. Baker & Yacef (2009) defined educational data mining as the application of data mining methods to educational data. There are five common methods that are classified into five categories, as proposed by Baker (2010), they include relationship mining, clustering, prediction, discovery with models, and separation of data for use in the process of human judgment, Baker (2010), Baker & Yancef (2009), Ventura (2010).

Rastrollo-Guerrero et al. (2020) looked at a number of articles that attempted to predict student behavior in the learning environment. According to the findings, there is a strong inclination to predict student success at the university level, with 70 percent of publications focusing on this topic. They also noted that the study of student dropout at the early stages of their academic careers is fascinating since there are still opportunities to learn more about useful predictive techniques

that can assist reduce student dropout. According to the study's findings, supervised learning is the most frequently applied method for predicting student behavior. They also discovered that most authors employed the support vector machine (SVM) method the most and that it delivered the best accurate predictions.

Using data mining techniques in higher education has so far focused on mining student performance and enrollment data, with significant studies focusing on predicting student performance and researching learning in order to make improvements to present educational practice. The following are the key domains of data mining implementations in higher education.

- Data analysis and visualization is used to highlight important information and aid decision-making. It can aid in the analysis of students' course activities and the development of a broad picture of a student's learning.
- Predicting student performance assists in predicting a student's performance, i.e., his/her course success. To analyze educational data, several approaches, and models such as neural networks, Bayesian networks, rule-based systems, classification, regression, and correlation analysis are used.
- Students are divided into groups based on their unique qualities, personal characteristics, and other factors. The educator can use these clusters/groups of students to create a customized learning system that promotes effective group learning.
- The application of information technology, i.e., a mature strategic information system, is the process of managing and generating strategic information (SIS). In educational institutions, SIS can be used to help with academic and administrative tasks. The goal is to

propose a method for understanding students' perspectives, satisfactions, and dissatisfactions with each aspect of the educational process.

- Target marketing generates a target set using a data mining technique, which is then used by marketing agents to plan promotions and marketing campaigns. Case studies assist institutions in developing a cost-effective technique for identifying alumni who are most likely to make pledges.
- Enrolment management is a term used frequently in higher education to indicate well-planned plans and methods for shaping an institution's enrolment and meeting set targets. Marketing, admission policies, retention initiatives, and financial aid granting are all examples of such strategies.

In this study, our focus is on the prediction theory in order to infer the efficacy of eLearning from student activity data. The focus of prediction is on developing a model through which we can extrapolate a single component of the data (the predicted variable) from a single or set of feature elements of the data (the predictor variables). Classification algorithm is considered for the process of extracting useful patterns. Several studies analyzing LMS data have adopted classification with successful results.

Kika et al. (2019) uses data mining techniques to analyze LMS log data and classify student learning styles. They combined Moodle log data with questionnaire data to predict students learning style on visual/verbal dimension. This study is good and makes use of both LMS data and questionnaire data in the analysis, but the research problem in this study is focused on understanding students learning styles. The researchers did not look at the effectiveness of the LMS or how it factors on the student’s academic achievement.

Casey & Gibson (2010) applied classification algorithms on Moodle log data and finding correlations that indicate student performance. They concluded that Moodle data can provide a measure of student engagement and that future research could extend this to other unexplored applications. The research found that daily course views are a good indicator for student performance. Other measures include page views, course logins, resource views, off-campus, and on-campus. This study makes in-depth analysis of Moodle log data; however, the researchers reviewed the measures independent of the student's final grade. The work also looked at student activity based on whether students accessed the LMS while on-campus or while off-campus.

Kazanidis et al. (2012) used four educational data mining techniques that included classification on data from an online learning system, the results showed that students having higher number of activities with the LMS had better performance. The study analyzed data from the eLearning system 'Open eClass' and looked at measures that include number of user sessions per course, number of user visits per course, duration of user visits per course, number of different pages in the course, number of files in the course and size of the educational content files in the course. While the study analyzed LMS data, this study's main goal was to evaluate online course content and its usage, and the effect of LMS users' online conduct on their performance as measured by their course grade.

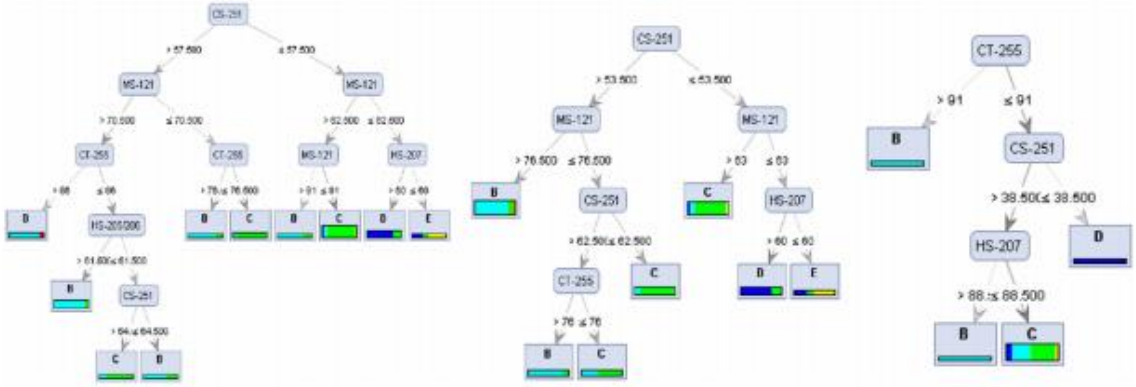
Classification algorithms have been used in mining educational data with successful results, albeit the focus has been on predicting student performance. The studies reviewed above indicate that further work is needed to extend the application of the prediction theory to other outcomes of eLearning.

2.3. Empirical Review

Asif et al. (2017) examined three research questions with the objective of equipping relevant stakeholders with information that would be helpful in improving educational programs at their institution. The researchers used data mining techniques to examine the performance of undergraduate students. They focused on two aspects of student performance: forecasting a student's academic achievement at the end of a four-year study program and merging usual progressions with prediction results. They found that it is feasible to predict the performance at graduation in a four-year study program using pre-university final grades obtained and the grades obtained in the first and second year with decent accuracy They also identified four course units that serve as effective indicators of good or poor performance. Lastly, they determined that the student's progressive academic performance over four-years of study remained relatively the same such that students who obtained high grades maintained high grades, and students who obtained low grades maintained low grades as they progressed through the four years.

FIGURE 2.1:

With feature selection, a decision tree was created with the Gini Index, Information Gain, and Accuracy.



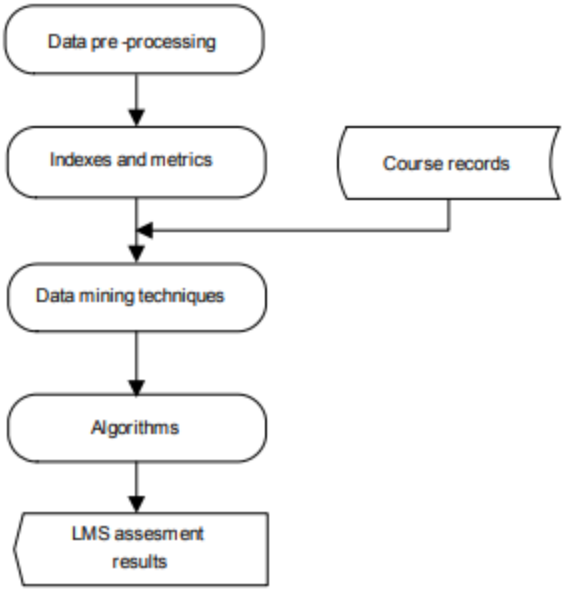
Source: Asif et al. (2017)

Alhassan et al. (2020) looked at how LMS assessment and activity aspects affected students' academic achievement. The researchers looked at five classification techniques for predicting student performance: Random Forest (RF), Sequential Minimum Optimization (SMO), Multilayer Perceptron (MLP), and Logistic Regression and Decision Tree (J48). Among the features collected from the LMS known as Blackboard were students' evaluation grades and metrics of their online activity. The Random Forest algorithm surpassed other classifiers in terms of accuracy in predicting student achievement. The study concluded that assessment data significantly influences student performance.

Kazanidis et al. (2012) applied existing techniques using a different approach to analyze log data from LMS. The researchers applied two course classification algorithms to investigate the link between LMS usage and the corresponding student performance in the exams. This was to discover the relations between students' performance, course characteristics and its usage. The research proposes new metrics and measures that include number of files, size of files, number of pages within each course on the online learning platform and the usage statistics for each course. The number of sessions, visits, and duration of each visit are the statistics. The focus of this study was to find measures that can assist educators in reviewing course utilization and locating weaknesses with online courses. A regression analysis was applied to determine if there are any interdependencies between the metrics of the course usage and student's exam marks.

FIGURE 2.2:

An approach for LMS assessment.



Source: Kazanidis et al. (2012)

Mwalumbwe and Mtebe (2017) created a learning analytics tool to determine the causal relationship between LMS use and student performance. They interrogated the LMS log data from two courses taught at Mbeya University of Science and Technology by subjecting the student's final results to linear regression analysis. The study found that the elements which influence student performance include, forum posts, peer interactions, and exercises. On the other hand, time spent on LMS, the number of downloads and the frequency with which a student logs in did not have a significant influence on student's performance.

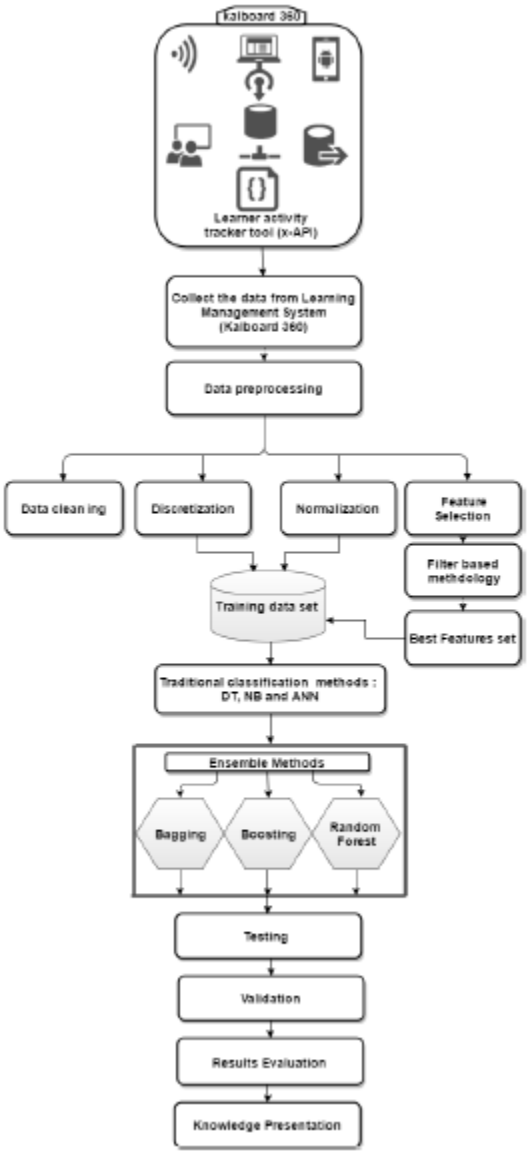
Damuluri et al. (2019) compared the performance of several classification algorithms by examining various elements of course material and students' online activity using data gathered from the Blackboard Learning Management System (LMS). Students' grades are influenced by a variety of criteria, including the amount of hours spent on the course each day, the number of hours spent on specific course content, lab assignments, homework assignments, the total number of

logins, and so on, according to their research. They believe that the Support Vector Machine is more accurate and effective than the other methods.

Amrieh et al. (2016) propose a new framework for predicting student performance, this model is based on incorporating new data features known as behavioral features into data mining techniques. The researchers applied combined methods to enhance the performance of a set of classifiers: Artificial Neural Network, Naïve Bayesian, Artificial Neural Network, and Decision Tree. The results obtained indicate a compelling relationship between students' behavior and their academic performance. The proposed model achieved an accuracy of 80% when applied to new newcomer student data.

FIGURE 2.2:

Student Performance Prediction Model Research Steps.



Amrieh et al. (2016)

Davies & Graff (2005) looked at how often 122 undergraduate students interacted online and measured this against the grades obtained at the end of the study year. The results led to the

conclusion that higher online interaction and participation did not necessarily result in higher grades but also indicated that students who had less frequent participation and interaction obtained lower grades overall. The results from this study indicate that there are other salient factors that contribute to the students' performance. Further research is needed to determine whether the online interactions and participation are important in providing the necessary student support and establish the quality and dynamics of these interactions.

Quinn & Gray (2019) investigated the use of Moodle LMS data in predicting the academic achievement of students in a blended learning environment. From Moodle logs of completed courses, the team created measurements of student activity. These were then used to predict whether a student would pass or fail a course and the alphabetic grade the student obtained. However, the focal point of this study was the classifiers that could predict the possibility of failing grades based on data obtained early in the semester. The results showed that classifier models based on the complete course data predicted the students grade moderately well with an accuracy of 60% and an accuracy of 92% on predictions of a student passing or failing. The classifier models based on the first six weeks of data did not accurately predict students who would fail. These models did show improvement as the data increased over the course of time with improved accuracy of prediction at 10-weeks.

2.4. Methods/Tools used to predict student performance

Predictive modelling is commonly used in educational data mining methods to predict student achievement. Several approaches are utilized to create predictive modelling, including classification, regression, and categorization. Classification is the most common approach used to

predict student success. In this study, several strategies for predicting student performance are used. The following are some of the algorithms that have been used:

2.4.1. Logistic Regression

Logistic Regression is a classification problem-solving Machine Learning method. It is a probability-based predictive analytic method. Logistic regression is a classification technique that assigns observations to a discrete set of classes, Kleinbaum & Klein, (2010). Classification challenges include email spam or not spam, online transaction fraud or not fraud, and tumor malignant or benign. This makes logistic regression a good fit for this study due to the discrete classes that are observed and predicted. The logistic sigmoid function converts a probability value from the output of logistic regression.

Advantage of Logistic Regression algorithm are:

- The training time of the logistic regression algorithm is significantly smaller than that of most complicated algorithms due to its straightforward probabilistic interpretation.
- It is very fast at classifying unknown records.
- It's easy to expand to a multi-class classification system.
- It is less inclined to over-fitting.

2.4.2. Naïve Bayes

The Bayes Theorem provides the basis for the Nave Bayes algorithm, which is utilized in a wide range of classification problems. It's a probabilistic classifier, meaning it makes predictions based on the probability of an object, Webb et al., (2010). The Naïve Bayes Classifier is a simple and efficient classification approach for building fast machine learning models that can make quick

predictions. When we have more than one class and are working with text categorization, Nave Bayes performs well.

Advantage of Naïve Bayes algorithm are:

- It is straightforward, and if the conditional independence assumption is correct, a Nave Bayes classifier will converge faster than discriminative models such as logistic regression, using less training data.
- Even if the NB assumption is incorrect, less model training time is required.

2.4.3. Random Forest

Random Forests is a Machine Learning approach that addresses one of the major concerns with Decision Trees A supervised learning algorithm, Random Forest is classified as such. Based on the predictions of the decision trees, the algorithm determines the outcome. It builds decision trees out of data samples, extracts predictions from each one, and then votes on the best solution, Liu et al., (2012). It's an ensemble method that's better than using a single decision tree because it averages the outcomes to avoid overfitting. Although it has two capabilities that is regression and classification, the dataset in this study will be treated as a classification problem because we have two options, that is defining if a student passed, or a student failed the course.

Advantage Random Forest algorithm are:

- Accuracy is higher
- Runs effectively on large datasets
- Can maintain accuracy with a large proportion of missing data

2.4.4. Support Vector Machine (SVM)

SVM is a type of supervised machine learning approach that can be used to address problems like classification and regression. Each data item is represented as a point in n-dimensional space (where n is the number of features), with the value of each feature being the SVM algorithm's value for a certain coordinate, Jakkula, (2006). Then we locate the hyper-plane that clearly separates the two classes to complete categorization.

Advantage SVM algorithm are:

- SVM operates well when there is a clear contrast between classes.
- SVM is more effective in high-dimensional spaces.
- SVM is effective when the number of dimensions exceeds the number of samples.

2.4.5. k-Nearest Neighbour

The supervised machine learning method k-nearest neighbours (kNN) is a simple and straightforward technique that may be applied to both classification and regression problems. kNN aims to predict the right class for the test data by computing the distance between the test data and all of the training points. The k number of points that are the most comparable to the test data are then chosen, Kramer, (2013). The kNN algorithm calculates the chance of test data belonging to each of the 'k' training data classes, then chooses the class with the highest probability.

Advantage kNN algorithm are:

- There is no training stage in kNN, making it a lazy algorithm. Because there is no training stage, you may begin classifying fresh occurrences straight away.

- Because kNN is non-parametric, you can let the data speak for itself rather than making a bunch of assumptions about it (e.g., linearity, conditional independence, etc.).
- The kNN algorithm is resistant to data with a high level of noise.

2.4.6. Gradient Boosting

Gradient boosting is a machine learning technique that can be used to a wide range of problems including regression and classification. It is based on the idea that combining numerous weak learners (e.g., shallow trees) can produce a more accurate prediction. Gradient boosting works by repeatedly developing simpler (weak) prediction models, with each model attempting to predict the error left over from the preceding model, Netekin & Knoll, (2013). Gradient boosting is a greedy strategy that can easily overfit a training dataset since it is greedy.

Advantage Gradient Boosting algorithm are:

- Train more quickly, especially with larger datasets.
- Support for categorical features is included, as well as the ability to handle missing data natively.

2.4.7. Neural Network

A neural network is a machine learning computational model. It's a collection of algorithms that authenticate the underlying relationship in a dataset in the same way as the human brain does, Wang, (2003). In this context, neural networks refer to neuronal systems that can be organic or artificial in nature. Deep learning uses artificial neural networks to perform sophisticated computations on massive amounts of data. Training data is used by neural networks to learn and improve their accuracy over time.

Advantages of Neural Network algorithm are:

- Neural networks have the ability to self-learn and produce output that is not limited by the input.
- They can run numerous jobs at the same time without slowing down the system.
- Ability to work with limited information. Even with limited information, the data can yield output after training.

2.5. Variables influencing the efficacy of eLearning

In this study, the emphasis is on the factors affecting eLearning and as such, the variables used are those most used in previous studies. The variables selected for use in this study have been identified from the literature reviewed in the previous section of this paper. These variables have been applied extensively in measuring the effectiveness of eLearning and have been shown to significantly affect the outcome of eLearning. Quinn & Gray (2020) observed there was a positive linear link between the frequency of activity on Moodle and the academic achievement. Other variables that showed high correlations include assignment views and assignment submissions.

In another study mining Moodle to understand student behavior, Casey & Gibson (2010) found that the variable 'daily views' was a surprisingly good indicator for good academic achievement. Although the work was in the early stages, the researchers concluded that Moodle activity had a positive correlation with the grades achieved. Mödritscher et al. (2013) examined 14 variables that were deemed to influence final grades and they found that 'online activities' within the Moodle platform had good correlations.

Table 2.3:

Variables employed in the conceptual framework's creation and development.

No.	Variables	Description	Source / Study
1.	Played	The number of times a video was seen by a student.	Elbadrawy et al. (2016) Sinha & Cassell (2015) Kim et al. (2014) Pardos et al. (2013)
2.	Paused	The number of times student paused a video	Elbadrawy et al. (2016) Kim et al. (2014) Pardos et al. (2013)
3.	Segments	The number of times student rewinded a video	Elbadrawy et al. (2016) Kim et al. (2014) Pardos et al. (2013)
4.	Online C	Online activity while on-campus	Casey & Gibson (2010)

5.	Online O	Online activity while off-campus	Casey & Gibson (2010)
6.	CGPA	Cumulative Grade Point Average	Abu, A. (2016) Elbadrawy et al. (2014)
7.	Number of Attempts	Number of past class failures	Ünal, F. (2021) Elbadrawy et al. (2014)
8.	Coursework 1	Grades earned by the student in their first coursework	Ünal, F. (2021)
9.	Coursework 2	Grades earned by the student in their second coursework	Ünal, F. (2021)
10.	End of Semester Exam	Grades earned by the student in their end of semester exam	Ünal, F. (2021)
11.	Academic Achievement	Final evaluation for course undertaken	Romero et al. (2008) Mödritscher et al. (2013)

2.6. Theoretical Framework

To measure the effectiveness of eLearning, we need understand the learning theories that are relevant to eLearning. Theoretical concepts serve as the foundation for every practical discipline, and so determine how well that discipline develops, Kibuku et al. (2018). The three classical learning theories are Constructivism, Cognitivism and Behaviorism, Kibuku et al. (2018). A fourth

new and emerging theory is Connectivism. Connectivism was proposed by George Siemens as a successor to behaviorism, cognitivism, and constructivism, Siemens, (2004).

2.6.1. Behaviorism Theory

Behaviorism theory is concerned with observable signs of learning. It focuses on measurable and observable activity as a learning indicator, O'Donohue, & Kitchener (1998). It is sometimes referred to as stimulus-response theory. This theory proposes that for learners to learn, they must be actively engaged and quickly rewarded for their efforts. The stimulus, the reaction, and the relationship between the two are all important factors in behaviorism. It's crucial to pay attention to how the stimulus-response relationship is formed, developed, and maintained.

If a student exhibits desirable behavior in class, the principles are reinforced. As a teacher, we should reinforce this conduct (stimuli) because the desired behavior will most likely grow more likely in the future (response). Learning, according to supporters of behaviorism, is simply the acquisition of new behaviors, Clark, (2018). Because such variables are not observable behavior, behaviorists do not place a high value on them as part of the learning process. They don't explain about memory or how new habits or changes in behaviors are remembered for later usage. The behaviorist turns his head and requests only that he be allowed to observe what his subjects are doing under certain stimuli, Watson, (1920).

According to behaviorism, learning at its best occurs when a teacher takes charge of the learning process and actively reinforces students in order to achieve the desired learning outcomes. The outcomes of learning are measurable/observable. Repetition and practice are essential for learning because they improve the link between the teacher's stimulus and the learner's desired response. Feedback is essential for learning because it encourages students to deliver the desired response

so that learning results can be measured. Positively reinforced behavior is more likely to be repeated, while negatively reinforced behavior is less likely to be repeated.

2.6.2. Cognitivism Theory

Cognitivism is a learning theory that focuses on how the mind receives, organizes, stores, and retrieves knowledge. It employs the mind as a data processor, much like a computer. As a result, cognitivism considers learning as an internal mental process rather than observable behavior, Clark (2018). In contrast to behaviorism, learning is viewed as internal mental processes by cognitivism, which goes beyond visible behavior. Learners, according to this perspective, are actively involved in the way they process knowledge. Knowledge, memory, reasoning, and problem solving are all areas that can be improved. As a result, cognitivists have concentrated on identifying mental processes — internal and conscious depictions of the world – that they believe are critical to human learning. According to the cognitive theory, in order to comprehend learning, we must look beyond observable behavior and consider the learner's ability to consciously reorganize his psychological sphere in response to experience.

Bloom's taxonomy of learning objectives Bloom et al., (1956) are related to the development of distinct kinds of learning skills, or ways of learning, and are the most extensively used cognitivism theories in education. They claimed that there are three major domains of learning, thinking (cognitive), feeling (affective) and doing (psycho-motor). Cognitive learning theory is underpinned by the principle that learning is the process of structuring information into models that can be understood. Instructions ought to be arranged, ordered, and presented in a way that the learner can understand and apply. It is critical to remember and recall information in order to

develop schemas in the brain. Organizing learning information helps with memory. Teachers must supply tools that aid in the processing of information by students' brains.

2.6.3. Connectivism Theory

In recent times, a new learning theory has emerged - Connectivism - which seeks to address learning in the Internet age. Connectivism argues that other learning paradigms focus on the process of learning rather than the usefulness of what is being taught, Siemens, (2004). Connectivism learning theory has been proposed as a theoretical framework for the digital era, Siemens, (2004), Goldie, (2016). According to Siemens (2005), connectivism asserts that knowledge is produced beyond the level of individual human participants, and that it is always modifying and changing. In connectivism, new types of knowledge develop through the shared connections between all the 'nodes' in a network. Although organizations can and should 'plugin' to this world of constant information flow and take meaning from it, knowledge in networks is neither controlled or created by any formal organization. As nodes come and go, and information flows across networks that are themselves interconnected with a plethora of other networks, knowledge in connectivism is an unpredictable, dynamic phenomena. Connectivism proponents contend that the Internet alters the fundamental nature of knowledge. A term like "constructing meaning" has no significance under connectivism. Connections are formed spontaneously, through a process of association, rather than being "built" through some kind of deliberate action.

2.6.4. Constructivism Theory

Constructivism is a teaching and learning philosophy that is founded on the idea that cognition (learning) is the product of "mental building". Constructivism is a learner-centered method in which learners actively create meaning from new information, while educators support learning

by providing in-depth feedback and posing directional questions. Clark (2018). According to this theory, learning occurs when an individual accumulates knowledge as a result of their experiences in the environment. People develop their own understanding and knowledge of the world in this way. Students attempt to make sense of new information by comparing it to what they already know and have encountered. They might revise their beliefs or write off new information. But in order to actively contribute to the creation of their knowledge, individuals must be able to inquire, investigate, and assess what they already know. A constructivist approach to learning in the classroom involves motivating students to use active learning techniques including experiments and real-world problem solving, preferably with authentic data, as well as to build knowledge and reflect on what they have learned.

Constructivism changes the teacher's role to one of assisting pupils in the construction of knowledge rather than simply reproducing a set of facts. To enable students to develop and verify their ideas, make conclusions and deductions, and express their knowledge, the constructivist educator offers tools like problem-solving and inquiry-based learning activities, such as in an eLearning setting. Before managing the exercises that will address and expand on the students' prior knowledge, the teacher must first become familiar with their preconceptions. Constructivist instructors encourage their pupils to assess the degree to which the activity is benefiting them in understanding. Students learn how to learn by challenging themselves and their learning methods while using computers either online or offline. The learners are then given the resources they need to continue learning throughout their lives.

Constructivism's educational theory is supposed to assist the teaching-learning strategy known as self-directed learning (SDL), which is used in eLearning. Constructivism theory states that

knowledge is created via personal experience, maturity, and engagement with one's environment, making eLearning an active information process. Contrary to objectivism, constructivism is a different educational philosophy since it sees the learner as a passive receiver of knowledge (Rovai, 2004).

A crammed educational style based on objectivist educational philosophy may result in learning performance that is inferior than that of eLearning, with the exception of a strategic approach pertaining to the efforts and studies for the pleasure of the self-learner. According to Lee et al. (2007), the SDL instructor serves as a learning helper and a mentor rather than a lone knowledge provider and messenger.

Through self-regulated learning, students take the lead in creating a whole learning process that includes problem perception, adoption, and evaluation of solutions (Lee, 2004). By structuring or rearranging knowledge, selecting knowledge, and putting it to practical use, learners do the same tasks as producers (Thatcher & Pamela, 2002).

eLearning is one of the SDL strategies that should be evaluated. This is justified by the fact that students only attend lectures to note the date, location, and topic of the lecture as well as to switch up the order in which they attend lectures. In comparison to currently practiced off-line education, adequate student monitoring is difficult due to changes in the evaluation method for learning progress as well as the removal of one-on-one contacts with the teacher from the process. As a result, it's critical to keep track of one's ability to arrange self-learning time, digest information, prepare data, and maintain data control.

eLearning Theory

In a study by Mödritscher, (2006), the results showed that behaviorism and constructivism are better suited for online learning compared to cognitivism. Constructivism outperformed behaviorism in a number of ways, including the ability to convey content from different perspectives, active knowledge production, and the development of meta-cognitive methods. eLearning theory is based on cognitive science concepts, which show how the usage and design of educational technology can improve learning effectiveness, David (2015); Wang (2012). It's made up of concepts that can be included into instructional design, showing "how educational technology might be used and structured to facilitate effective learning". Given the focus that learners are actively participating in the way they process information, and the inclusion of technology as a critical component, the Constructivism learning theory is most appropriate to the goals of this study. Because it underscores how technologies (the environment) can be used and developed to offer new learning opportunities and encourage effective learning, eLearning theory fits within the larger theory of Constructivism.

Mayer, Moreno, Sweller (2015), and their fellow researchers developed eLearning design concepts that emphasize limiting unnecessary cognitive load while introducing germane and intrinsic demands at relevant user levels. They proposed 11 design principles that guide the eLearning theory. They include.

- Multimedia Principle
- Modality Principle
- Coherence Principle
- Contiguity principle

- Segmenting Principle
- Signaling Principle
- Learner Control Principle
- Personalization Principle
- Pre-training Principle
- Expert Effect

For the purposes of this study, we give attention to three principles that are inherent to eLearning. According to the multimedia principle, combining any two of the auditory, visual, or text elements together facilitates deeper learning than just using one or using all three, Mayer, (1997). The principle of modality states that when visuals are complemented by audio narration rather than onscreen text, learning is more effective, Low & Sweller, (2005). Giving students the opportunity to set their own speed, according to the learner control principle, allows them to study more efficiently. It is preferable to merely have access to the play and pause buttons rather than having advanced controls such as fast forward, Scheiter, K. (2014). Given the nature of eLearning, where learners can control the pace of learning, learning is delivered in a combination of audio and visual elements, and visuals are enhanced by audio narration, the three principles outlined above are the most relevant to this study.

Factors affecting eLearning effectiveness

In their explorative and integrative review of the effectiveness of eLearning, Noesgaard et al. (2015) found that 56 percent of research papers defined eLearning effectiveness through 'learning outcome'. Studies dealing with higher education measured effectiveness by cognitive knowledge indicators. The research also found that learner motivation, prior experience with, and interaction

with the eLearning system all play a role in influencing effectiveness. This finding underpins the selection of connectivism theory for the purposes of this study. The researchers also found that a majority of these studies employed a quantitative research design to investigate and validate their findings.

When learners get new knowledge as a result of the eLearning effort, this is referred to as a "learning outcome". According to Boghikian-Whitby et al. (2015), formative assessments, summative assessments, and final letter grades are all used to assess students' learning. Therefore, to measure the effectiveness of eLearning, this study will apply student performance as the 'learning outcome'. Learner activity on the eLearning platform, such as accessing the platform on-campus or off-campus, will provide data on interaction level and learner motivation. It is therefore imperative to investigate what factors affect the efficacy of eLearning i.e., student performance.

Model for measuring effectiveness

Once the factors have been identified, the next step is to determine the appropriate model for measuring the efficacy of eLearning. Existing research has found that there is no simple or easy way to measure the effectiveness of eLearning, Kapounová, (2007). Liaw, (2008) proposed a conceptual model to measure the effectiveness of eLearning that considered three factors: multimedia formats, interaction environments and learners' self-efficacy. The researcher conducted questionnaire responses and applied regression analysis to check the effect the three factors on effectiveness of eLearning. The case study also investigated students' behavioral intention and perceived satisfaction.

In another study, Macgregor, & Turner, (2009) proposes a conceptual framework of e-learning effectiveness. The researchers considered system design, and usability, learner control, learning styles, social interaction, and information literacy as factors that affect eLearning effectiveness. The proposed framework consists of node A and node B. External elements impacting the learner and the efficacy of the student learning experience within the eLearning environment are the focus of Node A. Internal forces affecting eLearning efficacy are the focus of Node B. The 'learner control' component connects these nodes, indicating a two-way interaction provided by the student-system link, as well as a significant variable of efficacy as the degree of student learning control allowed.

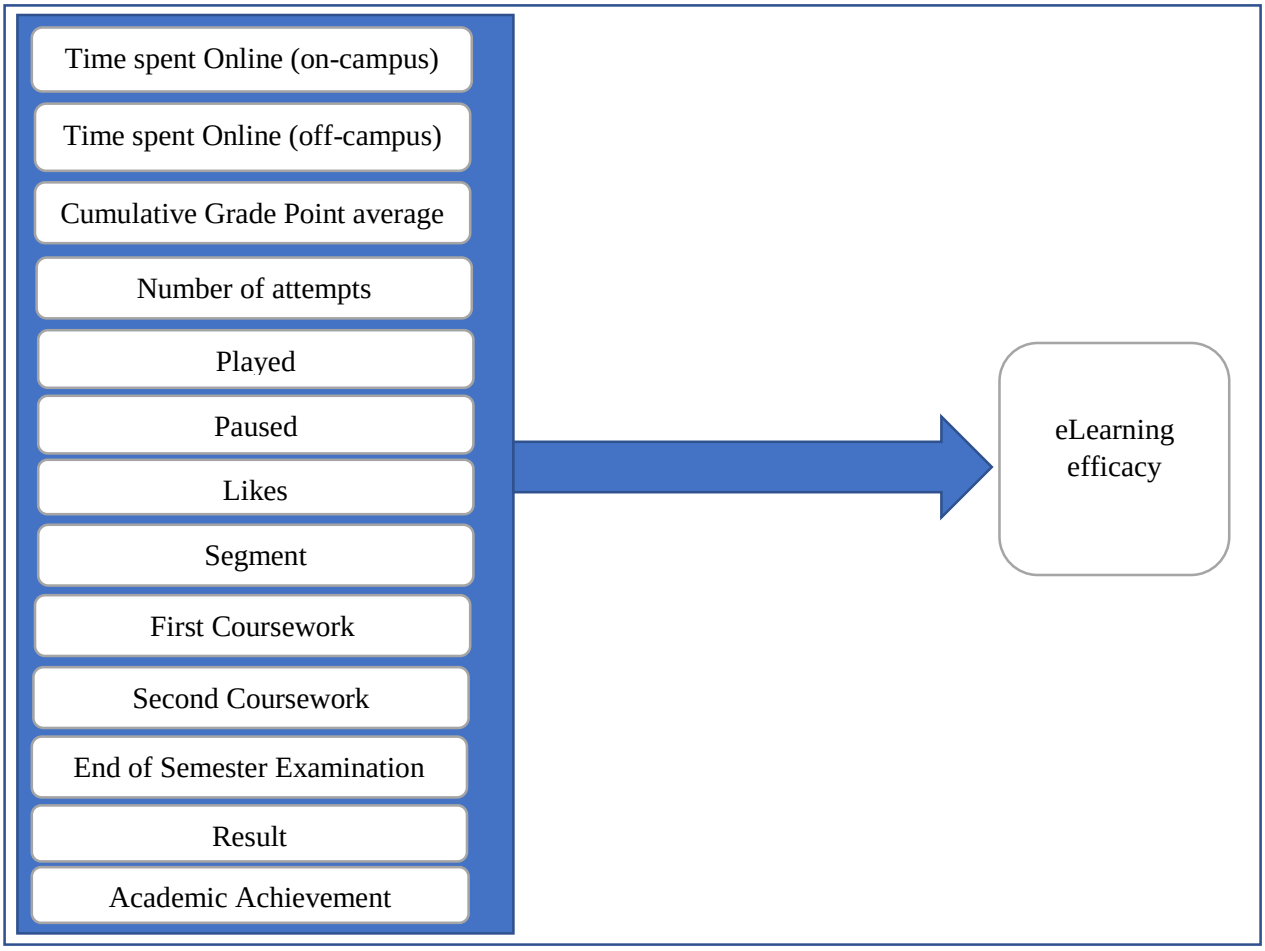
This study will use data from multimedia interactions, deliberate actions of learners, and academic performance to measure the efficacy of eLearning, based on the three design principles mentioned above.

2.7. Conceptual Framework

A conceptual framework, according to Mugenda & Mugenda (2003), is defined as a model that identifies the ideas under the investigation and their connections. The framework purposes to explain the relationship between variables and combine the idea in a methodical way to provide direction. It is assumed that an independent variable affects a dependent variable. The framework used in this study is enthralled on the comprehensive approach to measure the efficacy of eLearning LMS.

FIGURE 2.1:

Proposed framework



2.8. Operationalization of Variables

The selected variables must be defined in a measurable form in order to be used. The indicators and values expected from each variable that will be observed are summarized in Table 2.2 below. The indicators will be summations of the count of unique entries found in the database and each variable will have a numerical value that can be applied to an algorithms for data mining and machine learning. The expected value of the dependent variable 'efficacy of eLearning' is a binary value with 0 indicating no effect and 1 indicating an effect is present.

TABLE 2.2:
Operational definition of variables

Variable	Indicators	Values
Time spent Online (on-campus)	• On campus activities carried out by students (in minutes)	Discrete data type such as “25”
Time spent Online (off-campus).	• Off campus activities carried out by students (in minutes)	Discrete data type such as “25”
Cumulative Grade Point average (CGPA)	• The student's cumulative grade point average	Discrete data type such as “4.0”
Number of attempts	• A count of the total number of attempts made in the module	Discrete data type such as “1”
Played	• The number of times a video has been seen.	Discrete data type such as “4”
Paused	• The number of times student paused the video.	Discrete data type such as “4”
Likes	• The number of times the video has been liked by the student.	Discrete data type such as “4”
Segment	• The number of times a student has seen a certain video segment by using the slider	Discrete data type such as “4”
First Coursework	• Grades earned by the student in their first coursework	Discrete data type such as “86.5”
Second Coursework	• Grades earned by the student in their second coursework	Discrete data type such as “86.5”

End of Semester Examination	• Grades earned in the end semester examination	Discrete data type such as “86.5”
Result	• Student Performance	Pass or Fail
Efficacy of eLearning	• Comparing student performance while undertaking eLearning against while undertaking in-person learning	Binary i.e. 0 or 1

2.9. Summary

The literature above has shown that various studies have been completed on the adoption of eLearning platforms and as well as applying data mining techniques to evaluate student performance. From the literature reviewed it is observed that though a several studies have been conducted on the adoption of eLearning in Kenya and the application of data mining to evaluate student academic performance, the data used (surveys and student academic records) does not provide the necessary information to answer critical questions concerning online engagement and student outcomes. It is important to include data from learning management systems in evaluating student performance since these platforms have now become the de facto mode of student – teacher interactions in a post-COVID world.

CHAPTER THREE

METHODOLOGY

3.1. Introduction

In this section, we discuss data gathering and analysis methodology employed in this study. Areas discussed include research design, target population, the sample size and the sampling procedure used, research instrument, data collection techniques, and the data processing and analysis procedures used. The research schedule and budget are also proposed.

3.2. Research Design

This research's proposed design is to use a descriptive quantitative research methodology. A quantitative research strategy is essentially what descriptive quantitative research is. Quantifiable data from the population sample is obtained for statistical analysis in descriptive quantitative research. The descriptive research method is used to precisely describe a population or situation. Since descriptive designs do not change the environment or influence any parameters while collecting data, they cannot study cause and effect. Cross-sectional surveys, comparative designs, and correlations are all examples of descriptive designs, Baker (2017). This serves well the objectives, conceptual framework and research questions intended to be covered.

Additionally, exploratory research is utilized. The major purpose of exploratory research is to identify new patterns or problems Dietterich (1990). It seeks to enhance the basic knowledge of the theory and advance into the unknown realms of the subject. It encompasses collecting and analyzing numerical data then using the data to look for patterns and averages, to create forecasts, and to put causal linkages to the test, and extrapolate the outcomes to wider populations. Exploratory research is beneficial in machine learning (ML) because it's utilized to find new

problems that can't be handled with the help of available techniques. Exploratory research aims to define the problem precisely and determine what distinguishes it from problems that have already been solved.

In their research on an efficient eLearning model, Shen et al. (2007) used an exploratory research approach. In order to better understand the connection between emotion and learning, researchers created and designed an experimental prototype of an emotional aware learning system.

Alali and Xanthidis (2014) carried out exploratory research to examine the challenges and opportunities of eLearning in the Gulf Council Countries (GCC). By using questionnaires that were distributed both online and offline, the study's data were gathered from Saudi Arabian citizens and educational institutions.

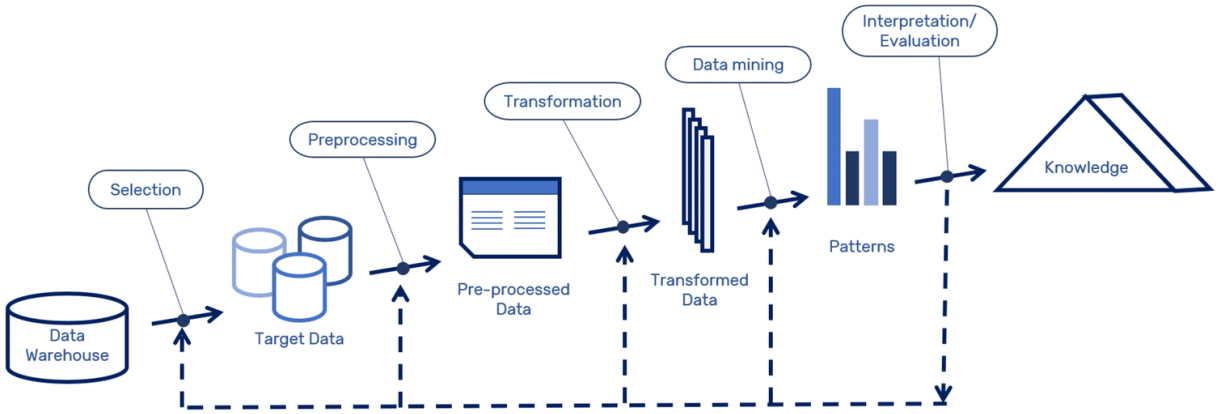
Ghosh's (2016) research used exploratory research to examine how learning outcomes were impacted by the acceptability of eLearning systems. The goal of the effort was to create and test an empirical model that used components from the Technology Mediate Learning (TML) framework (TAM). Using survey data acquired from an eLearning system designed to teach spreadsheet and database management software, the model was validated and the correlations were tested.

This study will make use of exploratory research in understanding the type of data that is best suited to measure the efficacy of eLearning. Since the required data will be collected through academic records and Learning Management System (LMS) through querying, it is key to determine whether the data will be useful in numerical form or in categorical form.

The purpose of the research is to find out if the introduction of eLearning during the Covid-19 period with support of an eLearning LMS platform has had a significant effect on students' performance. This research will analyze the primary indicators of student performance, i.e., course grades. For our analysis, we will use the effect of the dependent variables on the independent variable to determine the efficacy of learning through the LMS eLearning. Correlation analysis will be applied to explore the correlation between LMS usage variables and academic achievement.

FIGURE 3.1:

KDD Life Cycle



Source: Rotondo & Quilligan (2020)

Our model approach will align to the Knowledge Discovery in Databases (KDD) process as shown in Figure 3.1. KDD's objective is to provide methods and a process to extract knowledge from datasets. The main tasks will include data selection, data preprocessing, data transformation, and data mining. To develop the model, we will first export the data from the Moodle LMS database in csv format. Once the data is downloaded, we will then clean the data for noise, i.e., removing outliers and or missing values. The next phase will involve preprocessing the data into

a suitable form for the chosen model algorithm. This will be followed by specifying the test and training subsets. The next step is to train and model parameters from the training date set. We will then conduct model performance evaluation to check mode adequacy, after which we will be able to validate the model accuracy. Once satisfied with its performance, we will be able to make use of the model for the purpose's knowledge discovery.

The next section describes the various stages in the design approach.

3.2.1. Data

Moodle is a free, open-source learning management system designed to assist educators in forming efficient online learning communities. Moodle has been deployed at KCA university for the management of eLearning where students can access course materials, complete quizzes and assessments, submit assignments and exams papers. With Moodle, it is easy to create new courses and add content due to its modular design, Rice (2011).

Moodle keeps comprehensive logs of all student activities while logged into the system. Log files can be filtered by activity, user, day, and course. For items such as quizzes, we can obtain the score, elapsed time, and analysis of each student's responses. A detailed log of student involvement such as last login, number of reads, etc., is available in an activity report, Rice (2011). Moodle stores the logs in a relational database instead of text files. This means the data is stored in a single database. We use MySQL to extract data from the Moodle LMS since it is the most popular open-source database on the planet (MySQL, 2007).

3.2.2. Selection of Data

In educational data mining, the information required can be extracted from multiple sources such as Massive Open Online Course (MOOC), Intelligent Tutoring System (ITS) and Learning

Management Systems. In this study, the data is extracted from Moodle LMS while the independent variables for this research, as found in Table 2.2, are determined by the preprogrammed data collection module embedded in Moodle. The course grades will be collected as a dependent variable in this research.

3.2.3. Data Pre-processing and Transformation

To investigate the research questions for this study, the dataset is extracted from Moodle LMS. Moodle database structure allows for two options regarding data source. The first is an activity log, which Moodle utilizes to keep track of each student's activities. However, since Moodle is a web-based system, continuous tracking of usage is not possible because the system relies on http request and reply model. This makes it difficult to determine the amount of time spent in an activity. The second option is using data from sets of tables that Moodle creates for each and every module. These tables keep track of the major activities for the specific module, the tables provide data for activities such as start time, submission time for assignments, time taken, and final mark scored, for quizzes/assessments. The model will be based on the following variables as described in table 2.2.

- User-performed activities within campus
- User-performed activities outside of campus
- The number of times a student watched the video.
- The number of times a student paused the video
- The number of times a student expressed interest in the video by liking it
- The number of times a student watched a certain segment of a video.
- The number of attempts in a module
- Grades earned by the student in their first coursework
- Grades earned by the student in their second coursework
- Grades earned in the end semester examination

- The student's cumulative grade point average
- Final grade received by the student in the course.

3.2.4. Data Mining

This step entails feeding inputs into the model in order to train it on the expected input data and output expected. The excel sheets are first ingested into range and combined into a single sheet. After ingestion, the data is scaled in a [0,1] interval to guarantee efficiency while working with the data. The data is split using a ratio of 80:20, into training (80%) and test (20%) data. This is in following the pareto principle that indicates that a data-split of 80/20 ratio should be used Dunford et al. (2014). The training data is then fed into the classification model. The proposed framework will encompass classification techniques which include Random Forest, Decision Tree, Naïve Bayes, Support Vector Machine (SVM) and Neural Network. These algorithms are supervised learning algorithms that use the training data set to test the accuracy of the testing data. The classifiers are trained using the training data and a 10-fold cross validation technique is implemented.

3.2.5. Model Evaluation

Evaluation involves using the test data to check the model is properly trained by observing the actual model outputs compared to the expected model outputs. In order to assess the model's performance, this study will employ Confusion Matrix, and F1 Score metrics measurements. Binary classification data has two labels, positive (P) and negative (N). The outcomes are denoted as True Positive (TP) and True Negative (TN) for the correct positive and negative predictions, and False Positive (FP) and False Negative (FN) for incorrect positive and negative predictions.

Confusion Matrix

The confusion matrix is a *n x n* matrix with *n* being the number of expected classes. It's a term used to express how well a classification model performs.

TABLE 3.2:

Sample Confusion Matrix

		Predictions	
		Fail	Pass
Actual	Fail	a	b
	Pass	c	d

Confusion matrix provides information on;

Accuracy – proportion of the total number of predictions that were correct (Accuracy = (True Positive + True Negative)/total)

Precision – proportion of positive and negative cases that were correctly identified (Precision = True Positive/predicted yes)

Recall – proportion of actual positive cases which were correctly identified (True Positive Rate = True Positive/actual yes)

Specificity – proportion of actual negative cases that were correctly identified (True Negative Rate = True Negative/actual no)

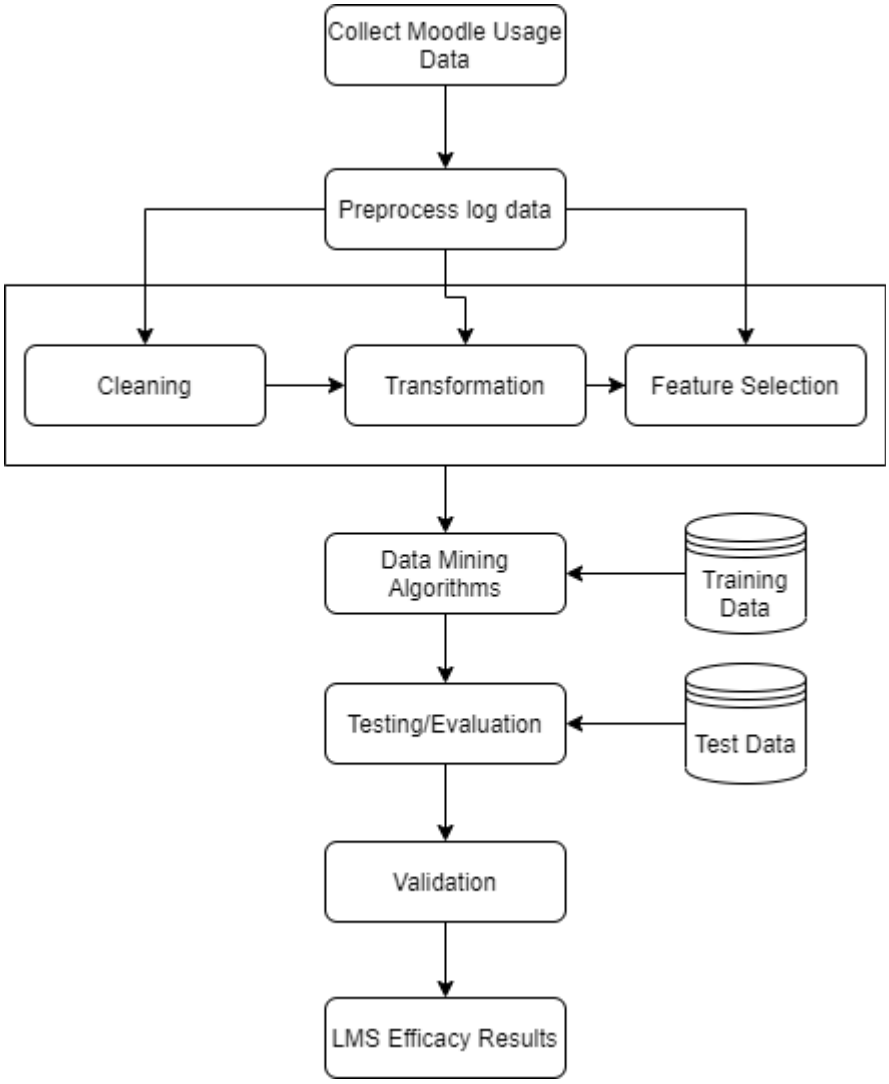
F1 Score

It's the harmonic mean of recall and precision and is calculated as per the formula.

$$F_1 = \left(\frac{\text{recall}^{-1} + \text{precision}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

FIGURE 3.2:

Flowchart of the proposed method



3.3. Target Population

The participants of this study are derived from students studying in a computing specialization and undertaking eLearning at the Middle East College (MEC), Muscat, Oman between Spring 2017 and Spring 2021. The sample population consists of 326 students out of which was found to be sufficient for the study.

3.4. Sampling and Sampling Procedure

Due to the sensitive nature of academic information, this study will employ a simple random sampling procedure that is relevant when sampling databases. Random sampling is used in instances when processing the entire database is either not required, is too costly in terms of resource usage or response time or for confidentiality purposes, Olken & Rotem (1986), Olken & Rotem (1995). In this study, random sampling is used as a measure of ensuring confidentiality of the student academic data used in this study. A sample is given on the basis that it does not expend much time, that it is cost effective, and it can be used to survey the entire study population, C.R. Kothari, (2004). This analysis will adopt a simple random sampling method for the student-based sample in which a systematic random sample of students is to be drawn using an equal probability selection method. Simple random sampling has been adopted to make certain that each sample of the given population has an equal chance of being chosen. Data from the computing department at MEC for the sixth and subsequent semesters will be used.

3.4.1. Simple Random Sampling Formula

Calculation of sample size

Formula

$$\text{Sample Size} = \frac{(\text{Z-score})^2 \times \text{Standard Deviation} \times 1 - \text{Standard Deviation}}{(\text{Margin of error})^2}$$

$$\mathbf{n = \frac{t^2 \times p(1-p)}{m^2}}$$

Where,

n = required sample size

t = confidence level at 95%

p = estimated prevalence of measure

m = margin of error at 5% (standard value of 0.05)

Based on the above formula, our ideal sample size will be 326 students enrolled in Information Technology programs within computing department at MEC.

3.5. Research Instrument

This study is based on data gathered from Moodle LMS platform and compares the performance of students while attending eLearning against the students' performance when attending in-person learning. The research instrument applied to this study will be the EDM techniques of relationship mining. To perform this analysis, we will apply statistical analysis techniques of correlation and regression analysis. To predict student's performance, we will apply machine learning techniques based on classification. Classification is deemed appropriate because we have two anticipated classes of the prediction output, a pass or a fail.

Rule induction is a data-mining method for finding frequent patterns, association, correlations, or casual structures by deducing if-then rules from a data set. The relationship between the attributes and class labels in the data set are shown using these symbolic decision rules. Rule induction CN2 algorithm will be applied to the data to mine information on how attributes in the data are related.

To implement the statistical and machine learning techniques mentioned above, we selected and adopted tools for data mining that are compatible with the Moodle platform. The tool adopted is, Orange Data Mining. Python general purpose programming language will be utilized to collect, analyze, and preprocess the data before analysis can be performed in the above tools. Python has been selected because of the research team familiarity with Python.

CHAPTER FOUR

DATA ANALYSIS, FINDINGS AND DISCUSSION

4.1. Introduction

This chapter deals with data analysis and interpretation in relation to the objectives and purpose. The aim of the research was to find out the efficacy of eLearning in Higher Education Institutions (HEI's). The chapter begins with a description of the dataset used. This chapter comprises of descriptive statistics on the dataset, visualizations of the data, data preprocessing applied, data analysis and the analysis results.

4.2. Dataset description

In this study, a publicly available dataset is used to predict student performance, Raza (2021). The data was gathered from the student information system (SIS), Moodle (the learning management system (LMS)), and eDify's applications video interactions. The data is from multiple systems in use at a higher educational institution (HEI) in the Sultanate of Oman. A student information system (SIS), Moodle LMS and eDify, the mobile application used for video content delivery. It comprises of data from five modules delivered in semesters between Spring of 2019 and Fall of 2019. There are 326 student records with a total of 40 features. 24 of the features come from the SIS, 10 from Moodle and 6 from the video interactions.

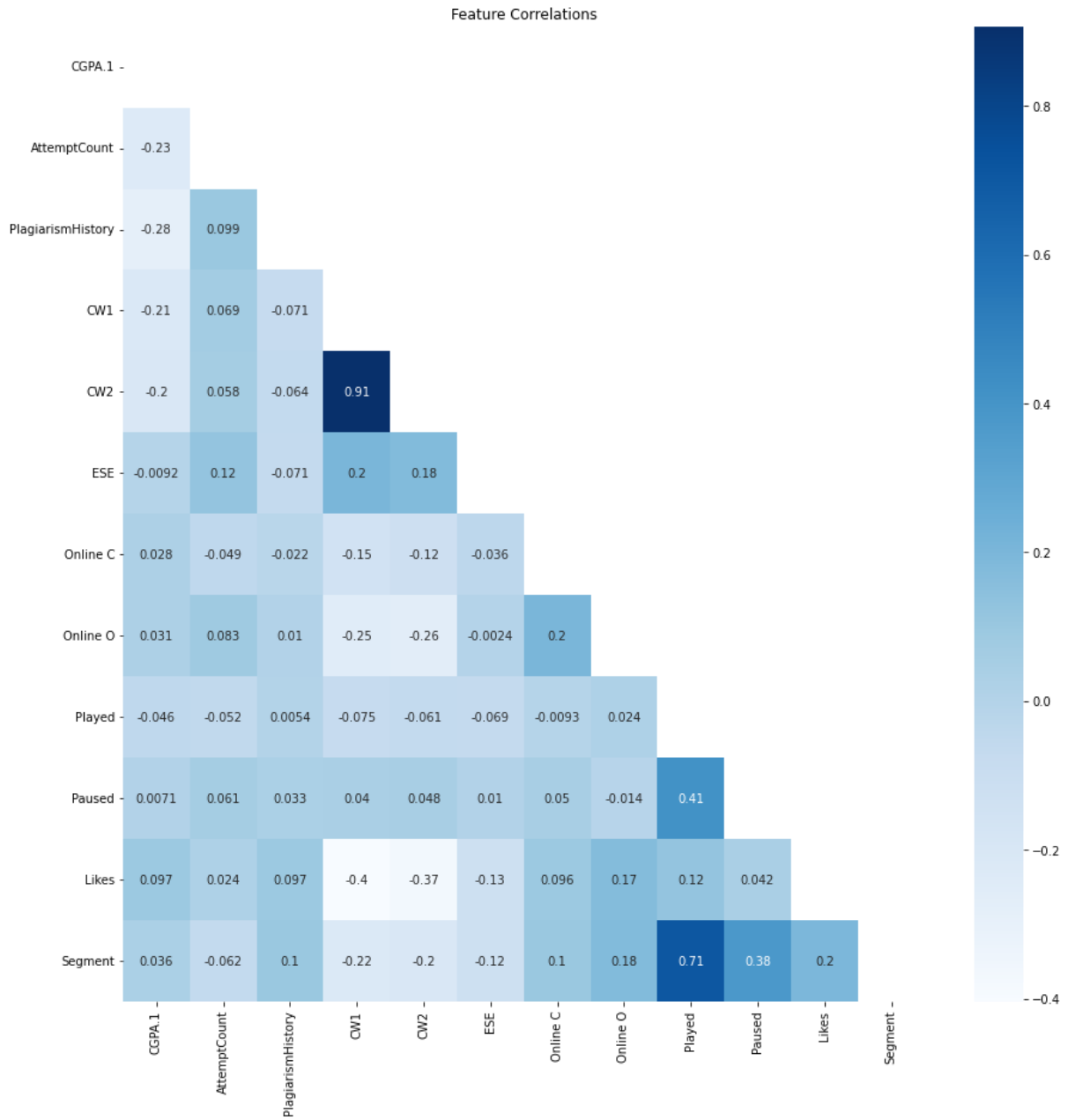
FIGURE 4.1:

Sample of dataset

ApplicantName	CGPA	AttemptCount	RemoteStudent	Probation	HighRisk	TermExceeded	AtRisk	AtRiskSSC	OtherModules	PlagiarismHistory	CW1	CW2	ESE	Online C
Student 322	Fair	Low	No	No	No	No	No	No	High	Low	Adequate	Fail	Fail	Very Good
Student 323	Adequate	Low	No	Yes	Yes	No	Yes	No	Low	Low	Adequate	Fail	Good	Poor
Student 324	Adequate	Low	No	No	No	No	No	No	High	Low	Adequate	Fail	Adequate	Excellent
Student 325	Good	Low	No	No	No	No	Yes	No	Low	Low	Adequate	Fail	Adequate	Good
Student 326	Adequate	Low	No	No	Yes	No	Yes	No	High	Low	Fail	Adequate	Fail	Excellent

FIGURE 4.2:

Correlation matrix of features



We can infer that there is a high correlation between coursework 1 and coursework 2, and a high correlation between the number of times a video is played and number of times segments re-winded

4.3. Descriptive Statistics

The dataset contained 19 categorical features and 12 numerical features.

TABLE 4.1:

Descriptive Statistics of the 12 Numerical Features

	CGPA.1	AttemptCount	PlagiarismHistory	CW1	CW2	ESE	OnlineC	OnlineO	Played	Paused	Likes	Segment
count	324	326	326	326	326	326	326	326	326	326	326	326
mean	2.47	1.12	0.18	51.10	50.08	56.38	208.35	194.98	2.13	2.18	1	1.48
std	0.61	0.41	0.39	22.48	23.50	17.62	124.02	131.13	1.76	2.60	0.97	1.88
min	0	1	0	11	0	0	2	0	0	0	0	0
25%	2	1	0	28.25	29	46	115.5	86	1	0	0	0
50%	2.25	1	0	50.58	44	56.75	188	184	1	1	1	0
75%	3	1	0	71.80	72.37	67.75	292	283	3	4	1.75	3
max	4	5	2	92.7	96	96	597	587	8	13	3	7

4.4. Data Preparation

Data pre-processing is a required step in preparing a dataset for classification procedures. It is vital to note that the quality and reliability of the available information have a direct impact on the outcome of this assignment. This task entails a thorough examination of variables and their corresponding values to rule out any anomalies. Three main pre-processing tasks were used in this investigation.

Feature selection. We examine our data meticulously to uncover attributes that have a stronger influence on our output variable. Even though our dataset does not have a huge number of attributes, some of them are unrelated to student achievement.

Orange has several different feature ranking algorithms. To choose acceptable attributes, we employed a ranking algorithm.

Missing data. Data that has not been captured for a variable for the observation in question is referred to as "missing data". The statistical power of the study is weakened by missing data causing the results to be skewed.

We chose to impute rather than delete in this investigation. For missing data, the imputation process generates credible predictions. When the percentage of missing data is low, it's the most beneficial. The missing observations were imputed using values obtained by calculating the mean.

Data transformation. The purpose of this pre-processing activity is to combine data into a single dataset from multiple sources. Then change the source data file's format to the destination data file's format.

4.4.1. Data Cleansing

In order to prepare the data for model training, the dataset was subjected to a thorough pre-processing procedure. Data cleansing was used at this stage to identify any missing or noisy data. Data cleansing involved separation of data and removal of information that was unrelated to the analysis. We checked for duplicates, null values, and correct data types. Unnecessary data was cleansed, and data was isolated from information that was not relevant to the study during this

process. The study took into account historical data for each student enrolled in the two modules. After data cleansing, a total of 17 features were identified as sufficient for the study.

Data preprocessing was completed using Python where the individual export files from the different systems were imported and aggregated into one data frame. During this activity, various attributes such as 'OtherModules', 'RemoteStudent', 'ModuleTitle', 'ModuleCode', 'SessionName', 'RollNumber', 'SpecialNeed', 'Advisor', 'PrerequisiteModules', 'User full name', 'Affected user', 'Event context', 'Component', 'Event name', 'Description', 'Time' which are either metadata, not useful or were anonymized to hide student identities and as such they were dropped from the final dataset. IP address was used to determine if students Moodle activity was on-campus or off-campus. The attribute 'ApplicantName' was however not dropped but tagged as a meta-attribute.

One column, CGPA was found to have missing values. We opted to fill the missing values using the average of the CGPA column as shown in figure 4.3 below

FIGURE 4.3:
Handling of missing values

```
data["CGPA.1"].fillna( data["CGPA.1"].mean().round(1), inplace = True)
data.isnull().sum(axis=0)
```

✓ 0.9s

```
CGPA.1              0
AttemptCount        0
```

This exercise resulted in the reduction of the variables/features from 31 down to 19 in the cleansed dataset. The cleansed dataset was then exported from the data frame as a comma separated file (CSV) for downstream use within the Orange data mining tool.

TABLE 4.2:

List of features and description

Feature	Values	Description
ApplicantName	Text + Numeric	Random generic student ID for identification and mapping
Cumulative Grade Point Average (CGPA)	0.00–4.00	Student cumulative grade point average of the student
AttemptCount	Low (1), Medium (2) and High (>2)	The number of attempts in the module
HighRisk	Yes/No	Shows students having a high failure rate in the same module
TermExceed	Yes/No	Students' progression rate in the degree plan
AtRisk	Yes/No	Student previously failed two or more modules
AtRiskSSC (Student Success Centre)	Yes/No	Student registered by the SSC for any educational deficiencies
PlagiarismHistory	Numeric Value	Students have been booked for any integrity violation
First Coursework (CW1)	1 - 100	Marks obtained in first coursework
First Coursework (CW2)	1 - 100	Marks obtained in second coursework
End Semester Examination (ESE)	1 - 100	Marks obtained in end semester examination
Students' online activity on Campus (Online C)	Time spent in minutes	Student Moodle activity performed within campus
Students' online activity off Campus (Online O)	Time spent in minutes	Student Moodle activity performed outside of campus
Played	No. of times video played	Accessed eDify and played video
Paused	No. of times video paused	Accessed eDify and paused video
Likes/Dislikes	Yes/No	Accessed eDify and either liked or disliked video
Segment	No. of times segments rewinded	Accessed eDify and rewinded a video segment
Result (Target Variable)	Pass/Fail	Overall marks obtained in the module

These 18 features were then ranked and selected by manually writing code on Python for feature extraction and ranking as shown on figure 4.4. We applied the Recursive Feature Extraction (RFE) approach to narrow down the relevant features in our dataset. Based on the output of RFE ranking, the final number of features proposed for use in the data mining model was 9 features.

FIGURE 4.4:

Feature Ranking

```
selected_rfe_features = pd.DataFrame({'Feature': list(X_train.columns), 'Ranking': rfe.ranking_})
selected_rfe_features.sort_values(by='Ranking')
```

5] ✓ 0.1s

	Feature	Ranking
0	CGPA	1
15	Paused	1
14	Played	1
13	Online O	1
11	ESE	1
10	CW2	1
9	CW1	1
16	Likes	1
17	Segment	1
12	Online C	2
6	AtRisk	3
4	HighRisk	4
1	AttemptCount	5
7	AtRiskSSC	6
5	TermExceeded	7
3	Probation	8
2	RemoteStudent	9
8	PlagiarismHistory	10

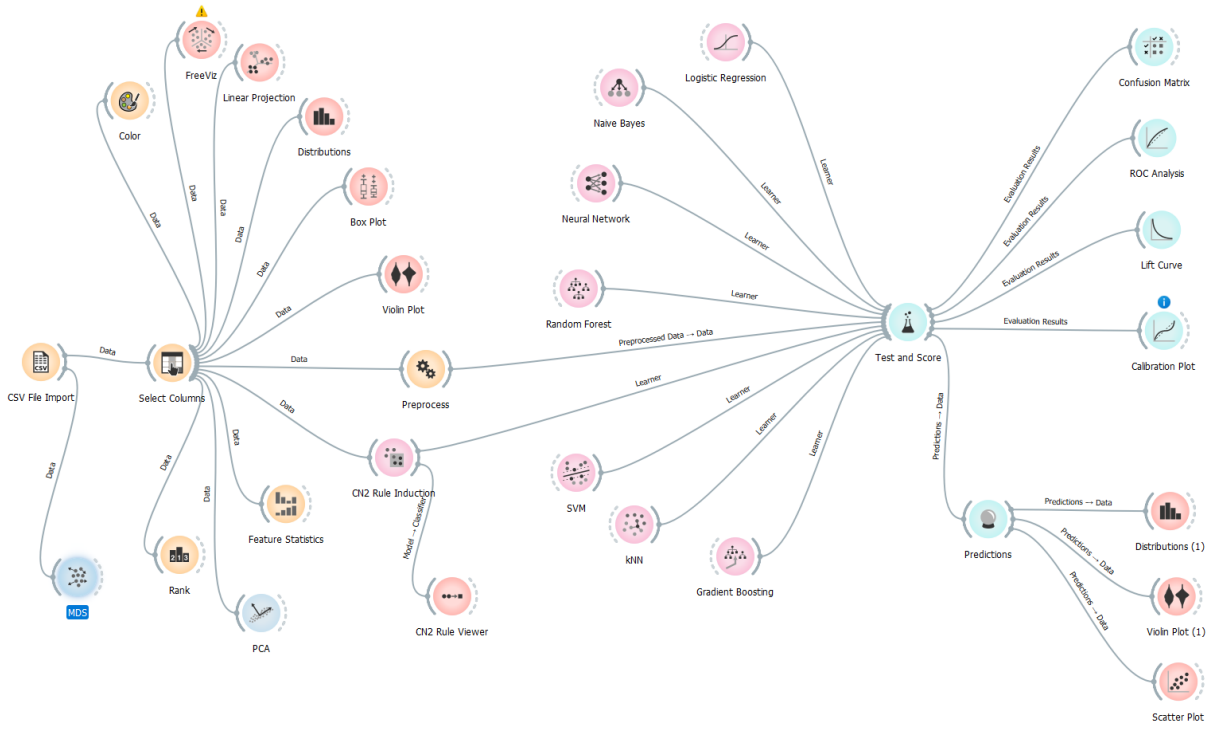
4.4.2. Data Preprocessing

The key rationale for testing multiple algorithms on the dataset was because their performance differed depending on which attributes were employed. According to research, algorithms behave differently depending on the dataset, as well as their efficiency and performance. It's much easier

to find the optimal algorithm for the dataset in terms of accuracy and performance using this strategy. This study used a similar approach to achieve its goal. The study used the Orange data mining tool, and the process is shown in Figure 4.5 below.

FIGURE 4.5:

Data Mining Process



Raw data acquired from these systems was transformed into a suitable format using pre-processing techniques. The CGPA was converted as excellent, very good, good, fair, adequate, or poor/fail; the plagiarism count was converted into low, medium, or high; the coursework (CW)1 was converted as excellent, very good, good, fair, adequate, or poor/fail; coursework (CW)2 was converted as excellent, very good, good, fair, adequate, or poor/fail; and the end-of-semester evaluation (ESE) was converted as excellent, very good, good, fair, adequate, or poor/fail.

FIGURE 4.6:

Transforming continuous data to categorical data

```
pd.cut(data.CW1,bins=[0,50,66,74,80,87,100],labels=['Fail', 'Adequate', 'Fair', 'Good', 'Very Good', 'Excellent'])
pd.cut(data.CW2,bins=[0,50,66,74,80,87,100],labels=['Fail', 'Adequate', 'Fair', 'Good', 'Very Good', 'Excellent'])
pd.cut(data.ESE,bins=[0,50,66,74,80,87,100],labels=['Fail', 'Adequate', 'Fair', 'Good', 'Very Good', 'Excellent'])
pd.cut(data['Online C'],bins=[0,30,50,100,200,300,500],labels=['Poor', 'Adequate', 'Fair', 'Good', 'Very Good', 'Excellent'])
pd.cut(data['Online O'],bins=[0,30,50,100,200,300,500],labels=['Poor', 'Adequate', 'Fair', 'Good', 'Very Good', 'Excellent'])
#data.insert(6,'CW1 Ranking',category)
```

✓ 0.7s

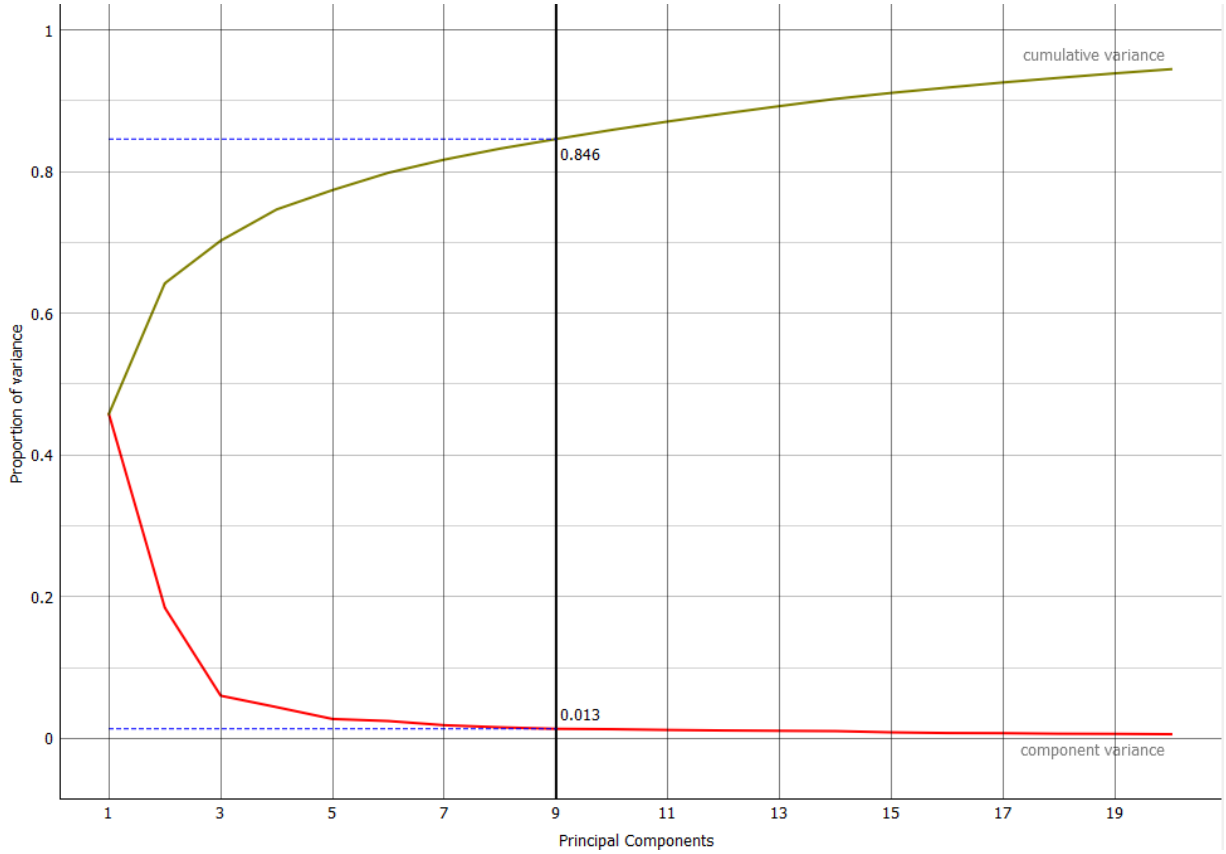
```
0      Excellent
1           Poor
2      Very Good
3           Fair
4      Very Good
         ...
321    Excellent
322    Excellent
323    Excellent
324    Very Good
325    Very Good
Name: Online O, Length: 326, dtype: category
Categories (6, object): ['Poor' < 'Adequate' < 'Fair' < 'Good' < 'Very Good' < 'Excellent']
```

Moodle keeps track of online activity in minutes, which were then transformed into nominal order as follows: into two categories: on-campus activity (excellent, very good, good, fair, adequate, or poor/fail); and off-campus activity (excellent, very good, good, fair, adequate, or poor/fail). AttemptCount and PlagiarismHistory were converted to (low, medium, or high). These features were converted into categorical data through discretization which was done to the continuous data. The data from eDify was not converted in any way, all four attributes were taken as numerical data.

Once the data was converted to categorical, the continuous data was discretized and discrete variables continuized, feature scaling was then applied. The data was standardized to $\mu = 0$ and $\sigma 2 = 1$. After standardization, we then applied the principal component analysis (PCA) as demonstrated in Figure 4.7. The PCA was used to minimize the number of variables from 19 to 9 components which explained 84.6 percent variance.

FIGURE 4.7:

Principal Component Analysis



4.5. Experimental Findings

We conducted some experiments to assess the accuracy and utility of various classification algorithms for predicting students' final grades based on information from their e-learning system usage data.

As a pre-processing operation, we experimented with using the numerical data as extracted without any discretization. We choose to classify the data using seven different algorithms that are cutting-edge and have shown to be effective in the field of EDM. The following classifiers were used:

- Support Vector Machine (SVM)

- k-Nearest Neighbour (kNN)
- Neural Network (NN)
- Gradient Boosting (GB)
- Logistic Regression (LR)
- Random Forest (RF)
- Naïve Bayes (NB)

The data was fed into designed model using the seven algorithms and used to predict the result attribute. The evaluation metrics are shown on table 4.3 below (ranked on accuracy descending).

TABLE 4.3:

Different classification algorithms' performance outcomes

Model	AUC	Accuracy	F1	Precision	Recall	LogLoss	Specificity
SVM	0.778	0.871	0.847	0.877	0.871	0.356	0.476
kNN	0.769	0.868	0.853	0.859	0.868	1.512	0.537
Neural Network	0.785	0.868	0.853	0.859	0.868	0.381	0.537
Gradient Boosting	0.773	0.862	0.844	0.852	0.862	0.404	0.511
Logistic Regression	0.755	0.844	0.820	0.826	0.844	0.411	0.445
Random Forest	0.744	0.837	0.815	0.817	0.837	0.892	0.443
Naive Bayes	0.727	0.801	0.799	0.797	0.801	0.453	0.534

The evaluation metrics show that Support Vector Machine (SVM) performed best with an accuracy of 87.1% and was followed by Neural Network and k-Nearest Neighbour (kNN) which both had an accuracy of 86.8%. The confusion matrices of each algorithm are discussed further in the following section.

Support Vector Machine (SVM) Confusion Matrix

The SVM algorithm was able to accurately predict 91.7% of failed students and 86.8% of passed students as seen in table 4.4 below. The algorithm also misclassified the predicted class 8.3% for fail and 13.2% for pass.

TABLE 4.4:

SVM Confusion Matrix

		Predicted		
		Fail	**Pass**	**Sum**
Actual	**Fail**	91.7%	13.2%	62
	Pass	8.3%	86.8%	264
	Sum	24	302	326

k-Nearest Neighbour (kNN) Confusion Matrix

The kNN algorithm was able to correctly predict 77.1% of failed students and 88.0% of passed students as seen in table 4.4 below. The algorithm also misclassified the predicted class 22.9% for fail and 12.0% for pass.

TABLE 4.5:

kNN Confusion Matrix

		Predicted		
		Fail	**Pass**	**Sum**
Actual	**Fail**	77.1%	12.0%	62
	Pass	22.9%	88.0%	264
	Sum	35	291	326

Neural Network Confusion Matrix

The Neural Network algorithm was able to correctly predict 77.1% of failed students and 88.0% of passed students as seen in table 4.4 below. The algorithm also misclassified the predicted class 22.9% for fail and 12.0% for pass.

TABLE 4.6:

Neural Network Confusion Matrix

		Predicted		
		Fail	**Pass**	**Sum**
Actual	**Fail**	77.1%	12.0%	62
	Pass	22.9%	88.0%	264
	Sum	35	291	326

Gradient Boosting Confusion Matrix

The Gradient Boosting algorithm was able to correctly predict 75.8% of failed students and 87.4% of passed students as seen in table 4.4 below. The algorithm also misclassified the predicted class 24.2% for fail and 12.6% for pass.

TABLE 4.7:

Gradient Boosting Confusion Matrix

		Predicted		
		Fail	**Pass**	**Sum**
Actual	**Fail**	75.8%	12.6%	62
	Pass	24.2%	87.4%	264
	Sum	33	293	326

Logistic Regression Confusion Matrix

The Logistic Regression algorithm was able to correctly predict 69.0% of failed students and 85.9% of passed students as seen in table 4.4 below. The algorithm also misclassified the predicted class 31.0% for fail and 14.1% for pass.

TABLE 4.8:

Logistic Regression Confusion Matrix

	Predicted		
	Fail	**Pass**	**Sum**

Actual	**Fail**	69.0%	14.1%	62
	Pass	31.0%	85.9%	264
	Sum	29	297	326

Random Forest Confusion Matrix

The Random Forest algorithm was able to correctly predict 64.5% of failed students and 85.8% of passed students as seen in table 4.4 below. The algorithm also misclassified the predicted class 35.5% for fail and 14.2% for pass.

TABLE 4.9:

Random Forest Confusion Matrix

		Predicted		
		Fail	**Pass**	**Sum**
Actual	**Fail**	64.5%	14.2%	62
	Pass	35.5%	85.8%	264
	Sum	31	295	326

Naïve Bayes Confusion Matrix

The Naïve Bayes algorithm was able to correctly predict 47.5% of failed students and 87.3% of passed students as seen in table 4.4 below. The algorithm also misclassified the predicted class 52.5% for fail and 12.7% for pass.

TABLE 4.10:

Naïve Bayes Confusion Matrix

	Predicted		
	Fail	**Pass**	**Sum**

Actual	**Fail**	47.5%	12.7%	62
	Pass	52.5%	87.3%	264
	Sum	59	267	326

4.6. Research Findings

To acquire a better grasp of our findings, we applied the CN2 rule induction algorithm on the non-processed data. The CN2 method is a classification strategy for efficiently inducing basic, intelligible rules of the type "if condition, then predict class," even in environments with noise. The domain of machine learning known as rule induction is concerned with extracting formal rules from a set of data. The derived rules could be a whole scientific model of the data, or they could just be local patterns in the data.

The CN2 Rule Inducer was chosen because it is easy to interpret for non-expert data mining users. In this situation, determining the probabilities of student interactions inside the learning environment systems would be easy for a faculty member.

Interpretations of the rules set by CN2 rule inducer is that a student who had a score of 49.0 or higher on their end of semester exam and had Moodle activity while off-campus for at least 235 minutes has a 99 percent chance of passing the module. If a student scored 29.5 or above on courserwork 2 and liked a video at least three times, they have a 91 percent chance of passing the module. A student who scored 44.76 or higher in their coursework 2 and accumulated more than 191 minutes of Moodle activity in-campus has a 92 percent probability of passing the module. Moodle activity of more than 189 minutes, having a cumulative GPA of greater than 2.46 and scoring 30.0 or higher in their coursework 2 gave the student a 96 percent probability of passing the module. While a student who has a cumulative GPA of greater than 3.0 and scored less than 19.0 in their coursework 1 has an 86 percent chance of failing the module. There is a 93 percent

probability that a student with a cumulative GPA of greater than 2.25 and an end of semester exam score of greater than 43.5 will pass the module. A student who has a 77 or greater in their coursework 1 result and played a video fewer than 2 times has a 94 percent probability of passing the module. Similarly, a student who has a cumulative GPA of 2.25 and played a video fewer than 2 times has a 95 percent probability of passing the module. Alternatively, a student who played a video fewer than 2 times and spent more than 328 minutes online in campus has an 83 percent probability of failing the module. A student who is not at risk and plays a video fewer than 2 times has a 94 percent probability of passing the module. If the student uses Moodle outside of the campus and likes a video that they have paused twice or more, they are likely to pass the module.

4.6.1. Objective one Results

The first objective was to investigate the factors that influence the efficacy of eLearning. Through the study, it was found that nine key factors played a significant role in the academic performance of a student at the end of a course. These are Moodle activity off-campus (Online O) or on-campus (Online C), student played video (Played), student paused video (Paused), student liked video (Liked), student replayed a segment of the video (Segment), the cumulative grade point average (CGPA), marks obtained in course work 1 (CW1), marks obtained in course work 2 (CW2), marks obtained in end of semester examination (ESE).

TABLE 4.11:

Factors affecting the efficacy of eLearning

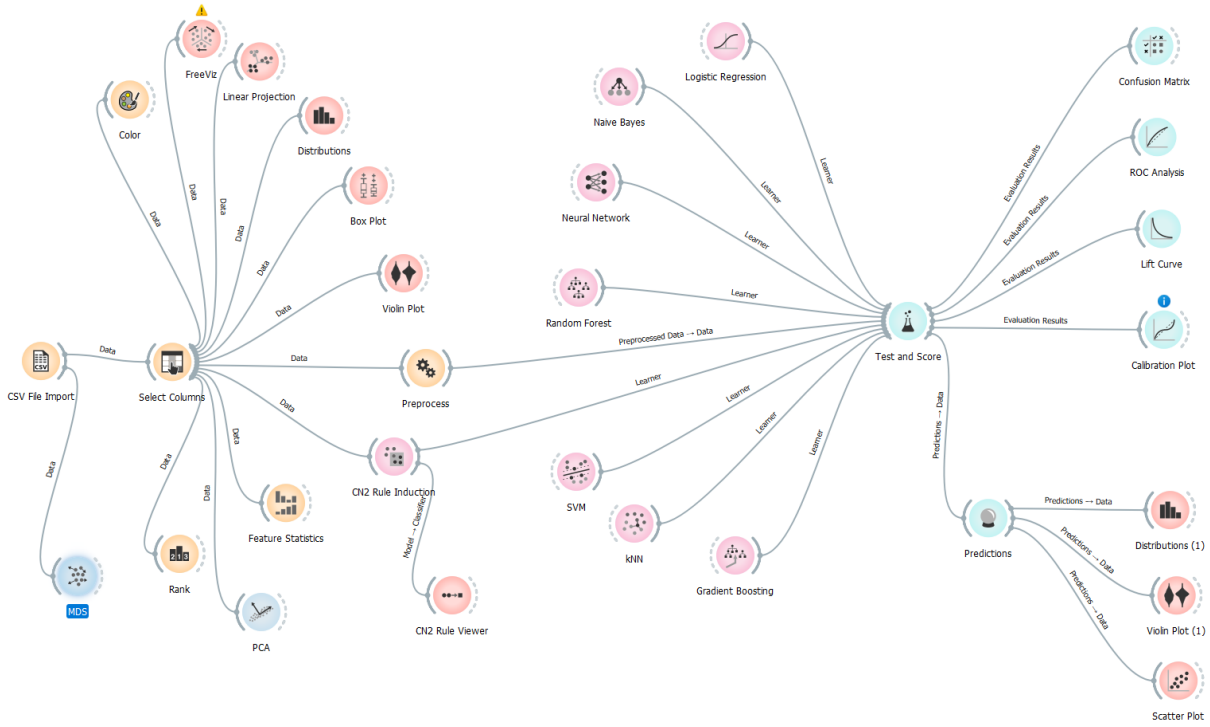
Variable	Information Gain
ESE	0.125
Played	0.062
CW1	0.055
CW2	0.048
Segment	0.046
Online C	0.018
CGPA	0.011
Likes	0.011
Online O	0.01
Paused	0.009

4.6.2. Objective two Results

The second objective was to develop a model to determine the efficacy of eLearning based on students' engagement and academic achievement. In order to find an appropriate model, we compared several algorithms i.e., Support Vector Machine (SVM), Logistic Regression, Neural Network, Random Forest, k-Nearest Neighbor (kNN), Naïve Bayes, and Gradient Boosting as shown in figure 4.8 below.

FIGURE 4.8:

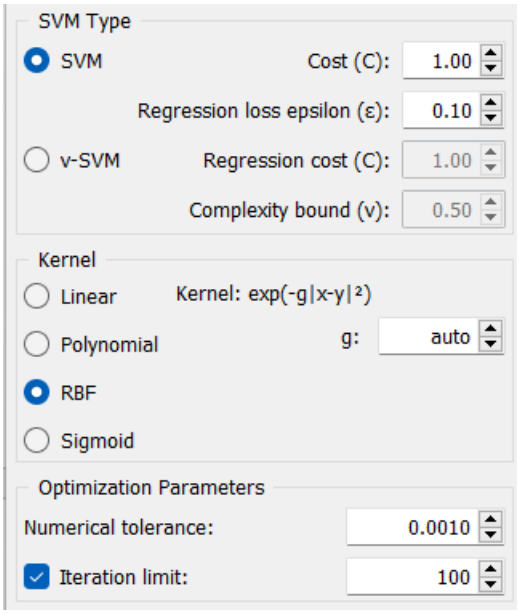
Model design on Orange data mining tool



Based on the performance and the prediction outputs from the selected seven models, this study found that the Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel calibrated with a cost of 1.00 and a regression loss of 0.10 and optimized with a numerical tolerance of 0.0010 is an adequate model to predict the student performance.

FIGURE 4.9:

Selected model: SVM



4.6.3. Objective three Results

The third objective was to test and validate the model developed above. The result of this objective is presented in table 4.5 and figure 4.10 below.

The performance of the seven models was then evaluated with the aim of identifying the best performing model. The results show that SVM performed best with an accuracy of 92% and was closely followed by Neural Network which had an accuracy of 91%.

TABLE 4.12:

Algorithm Evaluation Metrics

Model	AUC	Accuracy	F1	Precision	Recall	LogLoss	Specificity
SVM	0.835	0.920	0.916	0.919	0.920	0.289	0.722
Neural Network	0.835	0.911	0.907	0.908	0.911	0.332	0.720
kNN	0.840	0.899	0.896	0.895	0.899	1.650	0.729
Gradient Boosting	0.845	0.893	0.883	0.889	0.893	0.311	0.617
Random Forest	0.810	0.871	0.860	0.862	0.871	0.917	0.575
Logistic Regression	0.781	0.862	0.846	0.851	0.862	0.402	0.523
Naive Bayes	0.764	0.850	0.841	0.838	0.850	0.424	0.570

The prediction outputs from the seven models are shown below against the expected output as per the target variable ‘Result’.

FIGURE 4.10:

Expected vs Predicted outputs

Result	ApplicantName	Logistic Regression	Naive Bayes	kNN	SVM	Random Forest	Neural Network	Gradient Boosting
Pass	Student 1	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 10	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Fail	Student 100	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 101	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 102	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 103	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 104	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 105	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 106	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 107	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 108	Pass	Fail	Pass	Pass	Fail	Pass	Pass
Pass	Student 109	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 11	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 110	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 111	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Fail	Student 112	Pass	Pass	Fail	Fail	Pass	Fail	Fail
Fail	Student 113	Pass	Fail	Fail	Pass	Pass	Fail	Pass
Pass	Student 114	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 115	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 116	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Fail	Student 117	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Fail	Student 118	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 119	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 12	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 120	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 121	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 122	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 123	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 124	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Fail	Student 125	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 126	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Fail	Student 127	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 128	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Pass	Student 129	Pass	Pass	Pass	Pass	Pass	Pass	Pass

4.7. Discussion of Results

The Orange data mining software was used to analyze the dataset of 326 Middle East College (MEC) students in the Sultanate of Oman. The Orange software is an open-source program that uses a visual approach to machine learning for interactive data analysis, making it simple to build and configure workflows for various machine learning research. An Orange workflow was created in this investigation, as shown in Fig. 4.8. Seven data mining algorithms i.e., Support Vector

Machine (SVM), k-Nearest Neighbor (kNN), Neural Network (NN), Gradient Boosting (GB), Logistic Regression (LR), Random Forest (RF) and Naïve Bayes (NB) were applied to evaluate the predictive capabilities of the student features considered.

The workflow in Orange data mining allows for import of a dataset in csv format. A column selection widget provides the ability to select the feature columns and the target column in the imported data file. Columns can also be indicated as 'meta' variable that provides additional information on the dataset and are not considered for the analysis. Once the features and target variables are identified, the data is visualized in distribution plots, scatter plots, box plots and feature statistics. Feature ranking widget then applies the information gain ranking method to identify the most relevant features ranked by the information gain scores. The workflow branches to preprocessing and rule induction. In rule induction, "if–else–then", rules were extracted from a set of observations. The attributes and class labels in the data set have an underlying link that is explained by these symbolic decision rules. The rules and observations are discussed in detail in the research findings section of this paper (section 4.6). In preprocessing, the researcher handled the missing values through imputation using the average/mean value of the particular feature. The relevant features from the ranking were discretized in operation that transformed numeric variables to categorical, and feature scaling was then applied by way of standardization. Finally, the principal component analysis (PCA) was applied to the data. PCA was used to minimize the number of variables from 19 to 9 components which explained 84.6 percent variance. The seven selected algorithms were then fed with the data in the test and score dataflow section. Stratified sampling with a ten (10) fold cross validation was selected for this analysis. The output from test and score includes algorithm evaluation as shown on Table 4.12 above, and area under the curve (AUC) was used for model comparison. Evaluation results were visualized in a confusion matrix,

receiver operating characteristics (ROC) analysis graph, a lift curve, and a calibration plot. A table showing the predicted vs the expected for each classifier was output as shown in Figure 4.10 above. Further visualization of the prediction was done through a distribution plot, box plot and scatter plot.

Predicting student success and categorizing students are common activities in educational data mining, and both educators and students benefit from them. In this study, we offered a case study as an example in which these tasks were applied to online usage data acquired from the Middle East College of Muscat, Oman eLearning system.

The study was significant in understanding whether the adoption of eLearning influences the performance achievement of students. It demonstrated the value of data mining in higher education, notably in determining how the introduction of eLearning will affect students' performance and as an early warning system for identifying individuals who require help. There has been a wider adoption of eLearning and this trend is poised to grow even further with the fast-tracking of eLearning as enabled by the recent COVID-19 global pandemic. Universities and HEI's quickly adopted eLearning as a response to the pandemic, and many have opted to continue with this mode of learning for the foreseeable future. As a result, it was critical to conduct this research in order to determine how the transition to eLearning affects overall student performance.

This study looked at the efficacy of eLearning in Higher Educational Institutions (HEI's) by analyzing data that is available publicly and is sourced from a HEI in Oman. To discover knowledge, we used data mining techniques. The study evaluated the classification algorithms' performance using four common assessment metrics: accuracy, sensitivity, specificity, and f-measure as shown in table 4.12. For comparison with the baseline technique, a 10-fold cross-

validation was utilized, partitioning the data into 10 folds and utilizing nine folds for training and one-fold for testing. In this approach the dataset is partitioned into ten equal subgroups for training and testing. The subsets are run ten times each, with 90 percent of cases being trained and 10 percent being used to test the model each time. The tested instances in each iteration are unique. The final result is then computed as the average of the results. To evaluate mislabeling, confusion metrics were employed to analyze supervised learning, with each column of the matrix representing occurrences in a predicted class and each row representing instances in an actual class as shown in figure 4.11 below.

The Support Vector Machine (SVM) algorithm showed the best performance in predicting the students' performance. The classifier correctly predicted that 92.5 percent of the time, a student would pass the course and 89.1 percent of the time, a student would fail the course. The result was an overall accuracy of 92 percent and an F-1 score of 0.91.

FIGURE 4.11:
SVM Algorithm Evaluation Metrics

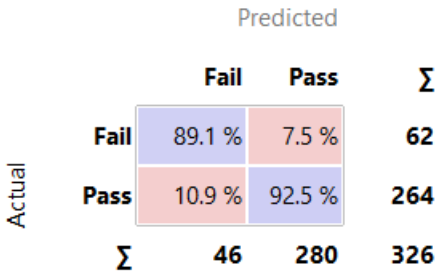
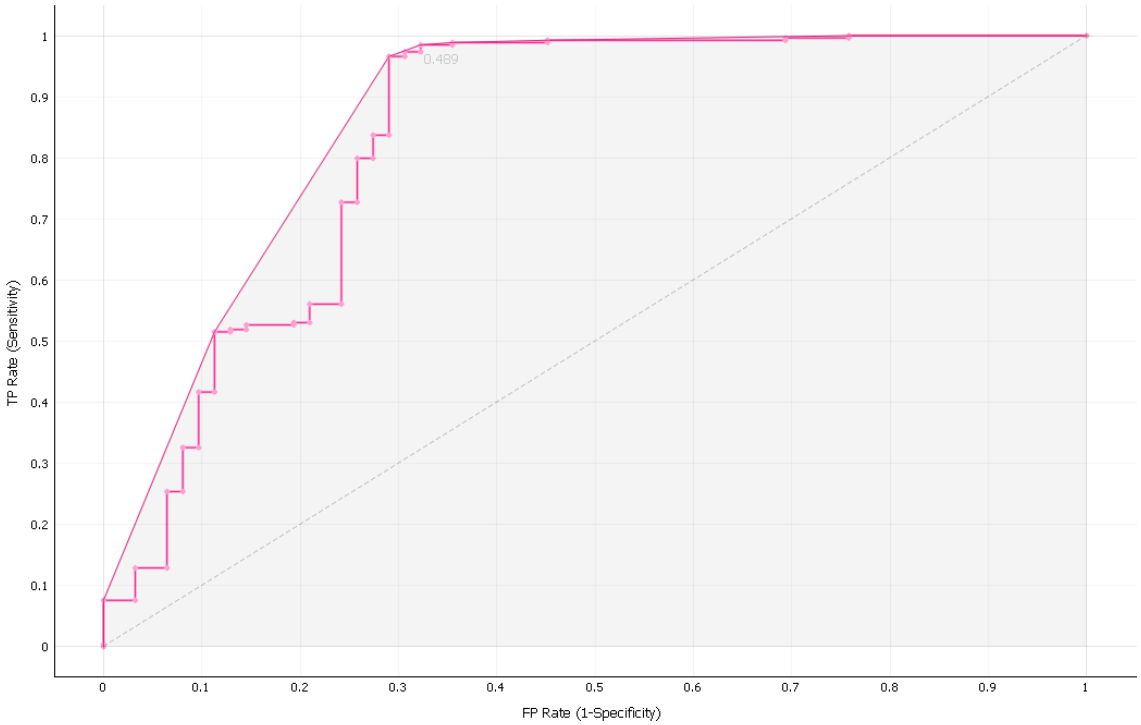


FIGURE 4.12:

SVM Algorithm ROC Curve



The optimum area under the curve (AUC) value should be somewhere between 0.5 and 1. The AUC for our SVM model is 0.489.

In the literature review and theoretical framework, it was shown from previous studies that multimedia interactions, deliberate actions of learners and academic performance are good measures of the effectiveness of eLearning. Research findings from this study support the conclusion that learners with high activity on the eLearning platform have a higher probability of passing the course while learners who showed low activity on the eLearning platform have a higher probability of failing the course. Learners who accessed the eLearning platform off-campus displayed deliberate actions and this too played a role in their academic performance. This is indicative of the efficacy of eLearning in higher educational institutions. Furthermore, the analysis

revealed that when learners have control of the pace of learning through integrated video delivery, learning is more effective. The learner control principle identified as part of the eleven design principles formulated by Mayer, Moreno & Sweller, (2015) comes to life when video delivery is incorporated in the eLearning platform. As a learner centered approach, constructivism is therefore well embodied in the three design principles assessed in this study.

A qualitative study by Xu, & Ebojoh, (2007) that made use of official record and interview sessions found that a gap still exists in the measurement of eLearning systems. The study discovered a number of variables, including assessment, benefits, constraints, and design delivery methods, that affect the efficacy of online learning programs. They proposed a model for effectiveness of online programs that addresses the delivery method, assessment, benefits, and constraints. The assessment, benefits, and constraints are all influenced by the design delivery, which has an impact on the effectiveness of the online education platform.

Blackburn, (2015) employed both qualitative and quantitative data analysis methods to investigate the effectiveness of pictures and stories in an eLearning course of statistics. The writers looked at empirical data on students' understanding and interpretation of the topic. According to the author, contextualizing theories inherent in eLearning systems might help students overcome barriers to learning by encouraging them to learn using scenarios that are more relevant to them. They claimed that the findings, as well as a noticeable improvement in concentration, enthusiasm, and participation, show a benefit over typical didactic teaching-centered techniques used in the authors' statistics classes.

Alwadei et al. (2020) studied the effectiveness of adaptive eLearning as compared to traditional learning for dental students. The authors investigated the impact of an Adaptive Learning Platform

(ALP) on learning by assessing learning effectiveness for both dental students who used the ALP summatively and formatively in a mixed learning environment versus students who studied in a face-to-face environment utilizing data from a period of 5-years as evaluated by the performance of the students on the final exam in a single preparatory course. The data was collected by use of pre-tests and post-tests. An adaptive learning intervention can have a considerable impact on student learning performance, according to the study. Any adaptive learning system's success, however, is largely dependent on good instructional design.

4.8. Summary

In this chapter, the study attempted to present a study analysis on the many elements such as descriptive and inferential analysis in this chapter. A description of the dataset used in the study which was sourced from publicly available datasets and originated from the Middle East College (MEC), Muscat Oman was provided. We further discussed the data cleansing and preprocessing steps taken to ensure the data was ready for machine learning. These steps included handling of missing entries, noise removal to make the data consistent, and feature extraction. Experimental results showing the performance of seven machine learning algorithms applied on the preprocessed data were used to inform the feature transformation approach that would work well with prediction. Patterns and associations were mined by applying CN2 rule induction on the unprocessed data. The results of the study objectives are provided and discussed. The next chapter summarizes the important research findings, draws conclusions, and makes recommendations for future research in relevant fields.

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.1. Introduction

The primary findings from the research of a model for evaluating the efficacy of eLearning in higher educational institutions using educational data mining are summarized in this chapter. The findings are used to develop conclusions and make recommendations.

5.2. Conclusions

One of the most important criteria for any college is student performance. Students' performance can be predicted based on their previous academic results. The conclusions of the study are that learning management systems can create new data on a student's conduct based on their digital profile. Students' eLearning activities and time spent on eLearning may be linked to their performance, according to the research. This form of analysis allows teachers to concentrate more on the students that require assistance. For starters, this data allows researchers to delve deeper into students' accomplishments in order to create and implement new forms of activities that result in favorable outcomes. Students who make use of the eLearning LMS platforms to their full extent, for example, obtain better grades, according to a study. The information gathered could be beneficial in identifying students who are struggling in a class early on. Teachers and students gain from this type of research because it enables teachers to identify ideal students for collaboration as well as students learning how to put in greater effort to achieve better results.

5.3. Contributions of the study

The current study provides significant contributions by attempting to fill several gaps. The study is useful in that it supports earlier studies while also revealing some fresh findings. First, the study

extends the limited research on the understanding of secondary data generated by learners while making use of LMS's and its impact on eLearning efficacy. Previous studies have investigated the efficacy of eLearning by interrogating factors derived from primary data mainly sourced from questionnaires which can be highly subjective. This is one of the earlier studies to consider factors from secondary data such as multimedia interactions, deliberate actions of learners and academic performance as important factors that affect the efficacy of eLearning.

Secondly, we assess the role of multimedia interactions, deliberate actions of learners and academic performance in eLearning environments. Thus, explaining the mechanism through which these factors can influence efficacy of eLearning. The factors identified in this study can play a role in the design of eLearning platforms that foster effective learning online.

Finally, this paper builds on and extends the body of work seeking to address the subject of eLearning efficacy in higher educational institutions from an African perspective.

The theoretical lens of the study is the constructivism learning theory that supports the explanation of a learner-centered approach in which students actively generate meaning from new material while instructors facilitate learning by providing comprehensive feedback and asking guided questions. The eLearning theory is anchored on three principles; the multimedia principle, the principle of modality and the learner control principle, as posited by previous researchers who proposed the theory based on constructivism. This model can be applied in study and practice in a number of different ways. This model, for instance, can help researchers better comprehend how to incorporate design ideas into training to encourage effective learning. The eLearning theory model can be used in research to describe the design principles in learning situations.

Hence, based on constructivism theory, the study intends to ascertain the importance of secondary eLearning data in shaping effective online learning. The study would add to the theoretical development by integrating eLearning theory with multimedia interactions, deliberate actions of learners, and academic and how they foster effective eLearning.

Only a few courses in one higher educational institution are included in the sample. Obviously, this study should be duplicated in other institutions, with different courses, and in different countries to see if the findings hold true in these different situations.

5.4. Recommendations for Future Research

Based on the findings of this study, we propose an LMS component that employs defined models which can be valuable in supplying educators with information obtained by using models defined in previous sub-sections. They can look through a list of students enrolled in their course and see what their chances are of succeeding. Educators may need to change their approach to dealing with children who are expected to be less than stellar and take corrective action sooner rather than later.

Further research is needed to model additional classification and clustering strategies. By boosting or fitting to eLearning data, the model has the potential to be further improved. Enriching student data with even more descriptors of their activity in the educational system (e.g., information gleaned via social network analysis) is also a worthwhile investment.

Further research focused on reinforcement learning can be useful in guiding learner's future course trajectory based on the strength of performance on current courses.

REFERENCES

Abu, A. (2016). Educational Data Mining & Students' Performance Prediction. *International Journal Of Advanced Computer Science And Applications, 7(5).* https://doi.org/10.14569/ijacsa.2016.070531.

Alali, A. S., & Xanthidis, D. (2014, January). An exploratory study of eLearning challenges and opportunities in the GCC. In 2014 World symposium on computer applications & research (WSCAR) (pp. 1-6). IEEE.

Alcala-Fdez, Jesus & Fernández, Alberto & Luengo, Julián & Derrac, J. & Garc'ia, S & Sanchez, Luciano & Herrera, Francisco. (2010). KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Journal of Multiple-Valued Logic and Soft Computing. 17. 255-287.

Alhassan, A., Zafar, B., & Mueen, A. (2020). Predict Students' Academic Performance based on their Assessment Grades and Online Activity Data. International Journal of Advanced Computer Science and Applications (IJACSA), 11(4).

Alwadei, A. H., Tekian, A. S., Brown, B. P., Alwadei, F. H., Park, Y. S., Alwadei, S. H., & Harris, I. B. (2020). Effectiveness of an adaptive eLearning intervention on dental students' learning in comparison to traditional instruction. Journal of Dental Education, 84(11), 1294-1302.

Amrieh, E., Hamtini, T., & Aljarah, I. (2016). Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. *International Journal Of Database Theory And Application, 9*(8), 119-136. doi: 10.14257/ijdta.2016.9.8.13

Araka, E., Maina, E., Gitonga, R., & Oboko, R. (2019). A Conceptual Model for Measuring and Supporting Self-Regulated Learning using Educational Data Mining on Learning Management Systems. *2019 IST-Africa Week Conference (IST-Africa).* doi: 10.23919/istafrica.2019.8764852

Aseey, A., 2020. Follow the Bridge: COVID-19 and Remote Learning in Kenya. [Blog] Research, Innovation and Enterprise Blog, Available at: <https://uonresearch.org/blog/follow-the-bridge-covid-19-and-remoteLearning-in-kenya/> [Accessed 29 December 2020].

Asif, R., Merceron, A., Ali, S., & Haider, N. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers &Amp; Education*, *113*, 177-194. doi: 10.1016/j.compedu.2017.05.007

Baker, C. (2017). Quantitative research designs: Experimental, quasi-experimental, and descriptive. Evidence-based practice: An integrative approach to research, administration, and practice, 155-183.

Baker, R.S.J.d., (2008). Data Mining for Education McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition). Oxford, UK: Elsevier.

Baker, S., & Inventado, P. S. (2016). Educational data mining and learning analytics: Potentials and possibilities for online education. In G. Veletsianos (Ed.), Emergence and Innovation in Digital Learning (83–98). doi:10.15215/aupress/9781771991490.01

Bencheva, N. (2010). Learning styles and eLearning face-to-face to the traditional learning. *Научни Трудове На Русенския Университет*, *49*(3.2), 63-67.

Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: An Issue Brief. *Office of Educational Technology, US Department of Education*.

Blackburn, G. (2015). Effectiveness of eLearning in statistics: Pictures and stories. E-Learning and Digital Media, 12(5-6), 459-480.

Boghikian-Whitby, S., & Mortagy, Y. (2008). The Effect of Student Background in E-Learning — Longitudinal Study. *Issues In Informing Science And Information Technology*, *5*, 107-126. doi: 10.28945/999

Casey, K., & Gibson, P. (2010). Mining Moodle to understand student behavior.

Clark, K. R. (2018). Learning theories: constructivism. Radiologic Technology, 90(2), 180-182.

C.R. Kothari, (2004). Research Methodology Methods and Techniques,2nd Edition, New Age International Limited, Publishers - ISBN: 978812241522

Damuluri, S., Ahmadi, P., & Islam, K. (2019, June). A Study of Several Classification Algorithms to Predict Students' Learning Performance. In 2019 *ASEE Annual Conference & Exposition.*

David, L. (2015, December). E-learning Theory (Mayer, Sweller, Moreno). Learning Theories. https://www.learning-theories.com/e-learning-theory-mayer-sweller-moreno.html.

Davies, J., & Graff, M. (2005). Performance in e-learning: online participation and student grades. *British Journal Of Educational Technology*, *36*(4), 657-663. doi: 10.1111/j.1467-8535.2005.00542.x

Dietterich, T. G. (1990). Editorial exploratory research in machine learning. *Machine Learning*, *5*(1), 5-9.

Digital Content – DigiSchool – ICT Authority. (2013). Retrieved 21 July 2021, from http://icta.go.ke/digischool/digital-content/

Dunford, R., Su, Q., & Tamang, E. (2014). The pareto principle.

Elbadrawy, A., Studham, R. S., & Karypis, G. (2014). Personalized multi-regression models for predicting students' performance in course activities.

Farrell, G. (2007). ICT in Education in Kenya. *Survey of ICT and education in Africa: Kenya Country Report.–April*.

Fortino, A., Zhong, Q., Huang, W. C., & Lowrance, R. (2019, March). Application of Text Data Mining To STEM Curriculum Selection and Development. In *2019 IEEE Integrated STEM Education Conference (ISEC)* (pp. 354-361). IEEE.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). How to design and evaluate research in education. New York, N.Y: McGraw-Hill Higher Education.

Goldie, J. (2016). Connectivism: A knowledge learning theory for the digital age?. *Medical Teacher*, *38*(10), 1064-1069. doi: 10.3109/0142159x.2016.1173661

Hugenholtz, N. I., De Croon, E. M., Smits, P. B., Van Dijk, F. J., & Nieuwenhuijsen, K. (2008). Effectiveness of e-learning in continuing medical education for occupational physicians. *Occupational Medicine*, *58*(5), 370-372.

IST-Africa. (2021). Retrieved 14 May 2021, from http://www.ist-africa.org/Conference2019/default.asp?page=paper-repository&fltyear=all&flttheme=pc%3Aelrn&flttype=all&flttitle=educational+data+mining&fltauthor=&pagesize=100&submit=Search

Jakkula, V. (2006). Tutorial on support vector machine (svm). School of EECS, Washington State University, 37(2.5), 3.

Kapounová, J. (2007). Approaches to the evaluation of elearning. In CBLIS Conference Proceedings 2007 Contemporary Perspective on new technologies in science and education. CY - Λευκωσία: University of Cyprus.

Kariuki, G. (2009). Growth and improvement of information communication technology in Kenya [Electronic Version]. *International Journal of Education and Development using ICT*.

Kashorda, M., Waema, T., Omosa, M., & Kyalo, V. (2007). E-readiness survey of higher education institutions in Kenya: a study funded by Partnership for Higher Education in Africa.

Kazanidis, I., Valsamidis, S., Kontogiannis, S., & Karakos, A. (2012, October). Measuring and mining LMS data. In 2012 16th Panhellenic Conference on Informatics (pp. 296-301). IEEE.

Kenya Ministry of Education Science and Technology. (2016). A Policy Framework For Education and Training: Reforming Education and Training in Kenya. Nairobi: Government of Kenya.

Kibuku, R., Ochieng, P., & Wausi, P. (2020). e-Learning Challenges Faced by Universities in Kenya: A Literature Review. *Electronic Journal Of E-Learning*, *18*(2). doi: 10.34190/ejel.20.18.2.004

Kika, A., Leka, L., Maxhelaku, S., & Ktona, A. (2019, July). Using data mining techniques on Moodle data for classification of student? S learning styles. In *Proceedings of International Academic Conferences* (No. 9211567). International Institute of Social and Economic Sciences.

Kim, J., Guo, P. J., Seaton, D. T., Mitros, P., Gajos, K. Z., & Miller, R. C. (2014, March). Understanding in-video dropouts and interaction peaks inonline lecture videos. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 31-40).

Kleinbaum, D. G., & Klein, M. (2010). Introduction to logistic regression. In Logistic regression (pp. 1-39). Springer, New York, NY.

Kombo, D. K., and Tromp D. L. A. (2006). Proposal and thesis writing: An introduction. Nairobi: Paulines Publications Africa.

Kramer, O. (2013). K-nearest neighbors. In Dimensionality reduction with unsupervised nearest neighbors (pp. 13-23). Springer, Berlin, Heidelberg.

Lee, J. K. (2004). The effect of e-Learning environmental quality and self-efficacy on effectiveness of an e-Learning. Business administration Doctoral dissertation, Daegu University.

Lee, J. K., and Lee W. K., 2007. The relationship of e-Learner's self-regulatory efficacy and perception of e-Learning environmental quality.

Li, C., & Lalani, F. (2020). The COVID-19 pandemic has changed education forever. This is how. Retrieved 21 July 2021, from https://www.weforum.org/agenda/2020/04/coronavirus-education-global-covid19-online-digital-learning/

Liaw, S. S. (2008). Investigating students' perceived satisfaction, behavioral intention, and effectiveness of eLearning: A case study of the Blackboard system. *Computers & education*, *51*(2), 864-873.

Liu, Y., Wang, Y., & Zhang, J. (2012, September). New machine learning algorithm: Random forest. *In International Conference on Information Computing and Applications (pp. 246-252).* Springer, Berlin, Heidelberg.

Low, R., & Sweller, J. (2005). The modality principle in multimedia learning. *The Cambridge handbook of multimedia learning, 147, 158*

Mabić, M., Dedić, F., Bijedić, N., & Gašpar, D. (2017). Data mining and curriculum development in higher education. In *International Conference on Information Technology and Development of Education–ITRO*.

Mayer, R. E. (1997). Multimedia learning: Are we asking the right questions?. Educational psychologist, 32(1), 1-19.

Mayer, R., & Mayer, R. E. (Eds.). (2005). The Cambridge handbook of multimedia learning. Cambridge university press.

Mayer, R. E., Moreno, R., & Sweller, J. (2015). E-learning theory. Learning Theories.

Means, B., Toyama, Y., Murphy, R., & Baki, M. (2013). The effectiveness of online and blended learning: A meta-analysis of the empirical literature. Teachers college record, 115(3), 1-47.

Mödritscher, F., Andergassen, M., & Neumann, G. (2013, September). Dependencies between eLearning usage patterns and learning results. *In Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies (pp. 1-8).*

Moreno, R., & Mayer, R. (2007). Interactive multimodal learning environments. *Educational Psychology Review, 19(3), 309-326.*

Mwalumbwe, I., & Mtebe, J. S. (2017). Using learning analytics to predict students' performance in Moodle learning management system: A case of Mbeya University of Science and Technology. *The Electronic Journal of Information Systems in Developing Countries*, *79*(1), 1-13.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, 21.

Noesgaard, S. S., & Ørngreen, R. (2015). The Effectiveness of E-Learning: An Explorative and Integrative Review of the Definitions, Methodologies and Factors that Promote e-Learning Effectiveness. *Electronic Journal of ELearning, 13(4), pp277-289.*

Nunn, S., Avella, J., Kanai, T., & Kebritchi, M. (2016). Learning Analytics Methods, Benefits, and Challenges in Higher Education: A Systematic Literature Review. *Online Learning*, *20*(2). doi: 10.24059/olj.v20i2.790

O M Mugenda, A G Mugenda (2003), Research methods: Quantitative and qualitative. Nairobi: Africa Center for Technology Studies

O'Donohue, W., & Kitchener, R. (Eds.). (1998). Handbook of behaviorism. Elsevier.

Ogwoka, T., Cheruiyot, W., & Okeyo, G. (2015). A Model for Predicting Students' Academic Performance using a Hybrid of K-means and Decision tree Algorithms. *International Journal Of Computer Applications Technology And Research*, *4*(9), 693-697. doi: 10.7753/ijcatr0409.1009

Oketch, H. A. (2013). E-learning readiness assessment model in Kenyas' higher education institutions: A case study of University of Nairobi (Doctoral dissertation, University of Nairobi).

Olken, F., & Rotem, D. (1986). Simple random sampling from relational databases.

Olken, F., & Rotem, D. (1995). Random sampling from databases: a survey. *Statistics and Computing*, *5*(1), 25-42.

Pardos, Z., Bergner, Y., Seaton, D., & Pritchard, D. (2013, July). Adapting bayesian knowledge tracing to a massive open online course in edx. In *Educational Data Mining 2013*.

Quinn, R. J., & Gray, G. (2020). Prediction of student academic performance using Moodle data from a Further Education setting. *Irish Journal of Technology Enhanced Learning*, *5*(1).

Raga Jr, R. C., & Raga, J. D. (2017). Monitoring Class Activity and Predicting Student Performance Using Moodle Action Log Data. *International Journal of Computing Sciences Research*, *1*(3), 1-16.

Rashty, D. (2003). Traditional learning vs. elearning. *Retrieved May*, *10*, 2011.

Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. Applied sciences, 10(3), 1042.

Raza Hasan. (2021). Dataset of Student's Performance using Student Information System, Moodle and Mobile Application 'eDify'. https://doi.org/10.5281/zenodo.5591907

Registered sites. (2021). Retrieved 11 May 2021, from https://stats.moodle.org/sites/index.php?country=KE

Rice, W. (2011). Moodle ELearning Course Development: a complete guide to successful learning using Moodle. *Birmingham–Mumbai: Packt Publishing*.

Ronoh, P. K. (2021). *Transformational Leadership and Implementation of Digital Literacy Programme in Kenya* (Doctoral dissertation, JKUAT-COHRED).

Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions On Systems, Man, And Cybernetics, Part C (Applications And Reviews)*, 40(6), 601-618. doi: 10.1109/tsmcc.2010.2053532

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, *51*(1), 368-384.

Romero, C., Ventura, S., Espejo, P.G., & Hervás-Martínez, C. (2008). Data Mining Algorithms to Classify Students. EDM.

Rotondo, A., & Quilligan, F. (2020). Evolution Paths for Knowledge Discovery and Data Mining Process Models. SN Computer Science, 1(2), 1-19.

Rovai, A. P. (2004). A constructivist approach to online college learning. The internet and higher Education, 7(2), 79-93.

Salkind, N. (2010). Encyclopedia of Research Design. Retrieved 16 May 2021, from http://dx.doi.org/10.4135/9781412961288

Scheiter, K. (2014). The Learner Control Principle in Multimedia Learning. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning (Cambridge Handbooks in Psychology, pp. 487-512).* Cambridge: Cambridge University Press. doi:10.1017/CBO9781139547369.025

Shen, L., Leon, E., Callaghan, V., & Shen, R. (2007, August). Exploratory research on an affective elearning model. *In International Workshop on Blended Learning (pp. 15-17).*

Siemens, G. (2004). Connectivism: A learning theory for the digital age.

Siemens, G. (2005). Connectivism: A learning theory for the digital age. *International Journal of Instructional Technology and Distance Learning. Online]* retrieved from: http://www. idtl. org/Journal/Jam _05/article01. html.

Siemens, George & Baker, Ryan. (2012). Learning analytics and educational data mining: Towards communication and collaboration. *ACM International Conference Proceeding Series.* Doi: 10.1145/2330601.2330661.

Sinha T., Cassell J. (2015, March), "Connecting the Dots: Predicting Student Grade Sequences from Bursty MOOC Interactions over Time*" In Proceedings of 2nd ACM conference on Learning@Scale.*

Thatcher, J. B., & Perrewe, P. L. (2002). An empirical examination of individual traits as antecedents to computer anxiety and computer self-efficacy. MIS quarterly, 381-396.

Titthasiri, W. (2013, November). A comparison of eLearning and traditional learning: Experimental approach. In *International Conference on Mobile Learning, E-Society and ELearning Technology (ICMLEET)–Singapore on November* (pp. 6-7).

Ünal, F. (2021). Data Mining for Student Performance Prediction in Education. *Data Mining - Methods, Applications And Systems*. doi: 10.5772/intechopen.91449

Veneri, D. (2011). The role and effectiveness of computer-assisted learning in physical therapy education: a systematic review. Physiotherapy Theory and Practice, 27(4), 287-298.

Waema, M. T. (2005). A brief history of the development of an ICT policy in Kenya. In E. F. Etta & L. Elder (Eds.), At the crossroads: ICT policy making in East Africa (pp. 25-43). Nairobi: East African Educational Publishers Ltd.

Wang, V. C. (2012). Understanding and promoting learning theories. *International Journal of Multidisciplinary Research and Modern Education, 8(2), 343-347.*

Wang, S. C. (2003). Artificial neural network. In Interdisciplinary computing in java programming (pp. 81-100). Springer, Boston, MA.

Watson, J. B. (1920). Is thinking merely the action of language mechanisms?. British Journal of Psychology, 11, 87-104.

Webb, G. I., Keogh, E., & Miikkulainen, R. (2010). Naïve Bayes. Encyclopedia of machine learning, 15, 713-714.

Witten, Ian & Hall, Mark & Frank, Eibe & Holmes, Geoffrey & Pfahringer, Bernhard & Reutemann, Peter. (2009). The WEKA data mining software: An update. SIGKDD Explorations. 11. 10-18. 10.1145/1656274.1656278.

World Bank. (2020). The COVID-19 pandemic: Shocks to education and policy responses.

Xu, H., & Ebojoh, O. (2007). Effectiveness of online learning program: a case study of A higher education institution. Issues in Information systems, 8(1), 160.

Yaghmaie, Mahkameh & Bahreinineja, Ardeshir. (2011). A context-aware adaptive learning system using agents. Expert Syst. Appl.. 38. 3280-3286. 10.1016/j.eswa.2010.08.113.

Yu, T., & Jo, I. H. (2014, March). Educational technology approach toward learning analytics: Relationship between student online behavior and learning performance in higher education. In *Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 269-270).

Appendix 1: Research Schedule

Table A.1:

Proposed Research Schedule

2021	APRIL	MAY	JUNE	JULY	AUGUST	SEPTEMBER	OCTOBER
Ideation							
Draft Research Questions							
Writing Research Proposal							
Review of Literature							
Proposal Presentation							
Data Collection							
Model Formulation							
Data Analysis							
Model Validation							
Compiling the work							
Final defense							
Document submission							

Appendix 2: Resources and Budget

Table A.2:

Proposed Research Budget

No.	Item	Unit	Price	Total
1	Computer / Laptop	1	50,000	50,000
2	Internet connection for carrying out research, software downloads, etc.	5	5,000	25,000
3	Travel cost to meet Supervisor	30	1,500	45,000
4	Data Collection and cleaning	1	2,000	2,000
5	Final Dissertation preparation (i.e., printing, binding, etc.)	500	20 per page	10,000
6	Flash Disk	1	1,500	1,500
7	Miscellaneous expenses	1		5,000
8	Total			138,500