AKETEN APPIAH-MENKA UNIVERSITY OF SKILLS TRAINING AND ENTREPRENEURIAL DEVELOPMENT

FACULTY OF APPLIED SCIENCES AND MATHEMATICS EDUCATION

DEPARTMENT OF INFORMATION TECHNOLOGY EDUCATION

USING MACHINE LEARNING TECHNIQUES FOR EARLY IDENTIFICATION OF AT-RISK STUDENTS IN A COURSE AT THE INTERMEDIARY LEVELS OF EDUCATION

ANANE GABRIEL

MAY, 2023

AKETEN APPIAH-MENKA UNIVERSITY OF SKILLS TRAINING AND ENTREPRENEURIAL DEVELOPMENT

FACULTY OF APPLIED SCIENCES AND MATHEMATICS EDUCATION

DEPARTMENT OF INFORMATION TECHNOLOGY EDUCATION

USING MACHINE LEARNING TECHNIQUES FOR EARLY IDENTIFICATION OF AT-RISK STUDENTS IN A COURSE AT THE INTERMEDIARY LEVELS OF EDUCATION

ANANE GABRIEL

(7211460003)

A PROJECT WORK SUBMITTED TO THE DEPARTMENT OF INFORMATION TECHNOLOGY EDUCATION OF THE AKETEN APPIAH-MENKA UNIVERSITY OF SKILLS TRAINING AND ENTREPRENEURIAL DEVELOPMENT IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF A MASTER OF SCIENCE DEGREE IN INFORMATION TECHNOLOGY EDUCATION

MAY, 2023

DECLARATION

STUDENT'S DECLARATION

I, Anane Gabriel, declare that this thesis, except for quotations and references contained in published works which have all been identified and duly acknowledged, is entirely my own original work, and it has not been submitted, either in part or whole, for another degree elsewhere.

Signature:……. Date:…….

SUPERVISOR'S DECLARATION

I hereby declare that the preparation and presentation of this work were supervised following the guidelines for supervision of thesis/dissertation/project as laid down by the Akenten Appiah-Menka University of Skills Training and Entrepreneurial Development.

Dr. William Asiedu (Supervisor)

Signature:…… Date:……

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my supervisor Dr. William Asiedu, for his invaluable guidance, unwavering support, and valuable insights throughout this thesis process. His encouragement and constructive feedback have been instrumental in shaping my research and academic development. I would also like to thank the Bia West Secondary School directorate for their time, expertise, and critical feedback. Their insightful comments and suggestions have greatly enhanced the quality of this thesis. I am also grateful to my colleagues and friends, who have provided me with moral support, encouragement, and technical assistance throughout this journey. Finally, I would like to express my heartfelt appreciation to my family for their love, encouragement, and support throughout my academic journey. Their unwavering belief in me has been my guiding light and main source of inspiration. Thank you all for your valuable contributions and support.

DEDICATION

I dedicate this thesis to my parent and wife, whose unwavering support, guidance and encouragement have been the foundation of my academic journey. Without them, I would not have been able to achieve this milestone. This thesis is a testament to their love and dedication.

ABSTRACT

This thesis focuses on the use of machine learning techniques to identify at-risk students at intermediary levels of education. The goal of this project is to develop a predictive model that can identify students who may be struggling in the course before they reach a critical point that leads to lower grades or withdrawal. Based on the analysis of related literature, this study selected students' personal characteristics and academic performance as input attributions at the intermediary level of education. Prediction models were developed using Artificial Neural Network (ANN), Decision Tree (DT) and Linear regression. A sample of 785 students was utilized in the procedures of model training and testing. The results of each model were presented in a confusion matrix and were analyzed by calculating the rates of accuracy, precision, recall, and F-measure. The results suggested all three machine learning methods were effective for student at-risk prediction, but DT presented a better performance. By identifying at-risk students early, teachers can provide additional support and resources to prevent students from falling behind and dropping out. Ultimately, this research could have significant implications for improving student retention rates in intermediary education and helping students achieve their academic goals.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.0 Background of the study

All students are expected to take general education classes in the intermediate levels of education to make sure they have the multidisciplinary background knowledge, analytical skills, and communication abilities needed to succeed in their chosen degrees and careers. Failing a course in general education could start a chain reaction of negative consequences that can range in severity from small (having to repeat a course) to severe (e.g., expulsion from school, delayed degree completion, and lost eligibility for financial help). As a result, a student's performance in these courses might be seen as being crucial to their academic achievement, including retention and graduation (Makombe & Lall, 2020). It is common knowledge that the implementation of corrective interventions depends significantly on the timeliness of the identification of students who are at risk. Yet, teachers typically have minimal knowledge of their student's performance during the first semester, which can make it difficult to identify at-risk students and have broad repercussions if the wrong assumptions are drawn (Pilotti et al., 2022). One of the tactics that can be utilized to increase completion rates is identifying at-risk students. Early identification of at-risk students may enable teachers to conduct instructional interventions and enhance course design. Instructors could give students instant feedback with a quick intervention solution, and retention rates may increase (Al-Shabandar et al., 2019).

Despite the requirement for accurate early forecasts of students' academic performance, which depend on limited data, the majority of research on machine learning intended to aid educators in performance prognostications has referred to substantially more data that has been gathered over a much longer time frame and has frequently involved discipline-specific subject matters. Examples include forecasting students' final course grades in a particular area

of study based on their cumulative grade point average (CGPA), grades in prerequisite courses, or more simply, on their academic history as demonstrated by their performance in previous courses(Pilotti et al., 2022). Yet, a wide range of creative stand-alone or hybrid solutions that emerge frequently in the body of existing literature, together with the algorithms that provide the best results, tend to vary greatly. Because of this, choosing and using a good method to identify at-risk learners may prove to be so difficult and overwhelming for a teacher whose professional field is not computer science. The most likely course of action is to disregard potentially useful technological solutions.

Early forecasts are undoubtedly more impactful than later ones, but early in a course, the teacher has access to limited data, making it difficult to predict student's difficulties (e.g., early poor performance a sign of a temporary problem, perhaps connected to the peculiarities of an assignment, or a reliable indicator of severe problems). It must be emphasized that by early identification of a student at risk of dropping out or retention, the required assistance and interventions may be given to the student, reducing dropouts and increasing retention, performance, and completion rates.

Numerous methods are being suggested right now to assess students' academic achievement. One of the most used methods to evaluate student performance is data mining. Recently, data mining has been utilized extensively in the educational field. Data mining in education is what it is termed. To determine important facts and trends from a sizable educational database, a procedure known as " Educational data mining" is implemented (Shahiri et al., 2015).

Prediction is among the most popular kinds of Educational Data Mining techniques. When making predictions, the objective is to create a model that can infer just one feature of the data (the predicted variables, which are comparable to the dependent variables in

conventional statistical analysis) from a combination of other aspects of the data (predictor variables, similar to independent variables in traditional statistical analysis). Knowing the anticipated variable for a small collection of data is essential when creating a prediction model; a model is then developed for this small set of data and statistically validated so that it may be used on a larger scale (Baker & Siemens, n.d.). For example, one could gather information on 500 learners who may be at risk of dropping out of school, create a prediction model to determine the extent to which a particular student will do so, validate it using data from segments of the 500 students who were left out when developing the prediction model, and then use the resulting model to make predictions about new students. Models for predictions are frequently used for predicting either future occurrences or factors that are difficult to measure immediately in the present time.

1.1 Statement of the problem

In recent years, a brand-new field of research known as "educational data mining" has emerged as a result of the development of numerous statistical approaches to analyze data in the context of education. One such use of educational data mining is early student outcome forecasting. Identification of the "weak" student is required at all educational levels in order to plan some sort of remedial for them (Baek & Doleck, 2021). Although computer systems are not as adept at acquiring knowledge as humans are, numerous machine-learning methods have been created that are useful for particular learning tasks. They are particularly helpful in poorly understood fields where people might lack the information necessary to create efficient experiences and understand algorithms. In general, machine learning investigates techniques that draw conclusions from data provided (the input set) to generate broad assumptions that make predictions about future instances to come. Several situations involving intermediary learning could benefit from the capacity to forecast a student's success. A supervised machine learning system can train on the key demographic traits of

students as well as their grades and a few written assignments. The learning algorithm might then be able to anticipate how prospective students will perform, making it a useful tool for identifying expected underachievers.

Due to the advancement of artificial intelligence, machine learning, and deep learning techniques, researchers have been able to design a number of prediction models to uncover oblique research trends that explain the strengths and weaknesses of students. Researchers can utilize machine learning approaches to examine several factors that have a big impact on student at-risk rates in order to lower dropout rates. By accurately identifying students who are likely to quit their educational pursuits, predictive models driven by Machine Learning techniques let teachers develop preventative strategies before dropout behavior starts. As far as the researcher is aware, there has not been a publication like this in the literature. By utilizing machine learning approaches to comprehend characteristics connected with students' learning behavior and how they connect with academic results, this research project's primary objective is the earliest possible detection of students who are in danger of dropping out.

1.2 Purpose of the study

To reduce student dropout in a course, it is important to comprehend the fundamental elements of student dropout and which students are at risk of dropping out at the intermediary level of education. Goverment officials are working harder in the area of education to reduce the number of student dropouts at the intermediary level of education. In Ghana, there are numerous initiatives aimed at lowering the number of students dropping out, including free senior high school initiatives and school feeding programs. At the intermediate level of education, those programs are available to all students and are not specifically aimed at the group of students who are in danger of dropping out. An effective program to prevent student dropout must be economical and should focus on at-risk students. At-risk students must first be identified using school data that is currently available, then they must interact with a

pertinent outreach program, and finally, the intervention must be assessed. In line with this, the study aims to use machine learning techniques for the early identification of at-risk students in a course at the intermediary levels of education.

1.2.1 Specific purpose of the study

The research set out to accomplish the following goals;

1. To assess students' performance using a machine learning algorithm.
2. To use regression methods to predict students' marks.
3. The effect of students' demopraphy on their academic performance.

1.3 Research questions

The following questions were the focus of the research.

1. Which machine learning algorithm performs best in predicting students performance, considering different factors such as previous academic achievements and socio-economic indicators?
2. How will linear regression, artificial neural network (ANN), and decision tree methods be used to predict students' marks?
3. How does demographic information impact the academic performance of students'?

1.4 Importance of the study

To solve institutional teaching and learning issues pertaining to the early detection of at-risk students of attrition or academic failure, the topic of learning analytics has received a lot of attention. The study examines how educational settings affect a model that predicts student success to address this issue. Through this approach, the research hopes to empirically show how crucial it is to comprehend the course and disciplinary context when creating and analyzing models that predict academic success and attrition (Gašević et al., 2016). This

study uses machine learning techniques to analyze these perspectives in order to help the teachers and administrators develop educational pedagogy interventions, improve the student's academic achievement, and identify the students at risk of low performance and graduate from the institute late. The work closes the gap between machine learning techniques currently in use and empirical predictions of student performance. The study's findings are pertinent to academia, researchers, and educational institutions since they advance our understanding of how to anticipate student achievement.

1.5 Organization of the Study

There are five chapters in the research. The study background, problem statement, purpose, specific goals, research questions, the significance of the study, and study organization are all presented in chapter one. A review of the literature is presented in chapter two on peer instruction, educational data mining, at-risk learners, students at-risk of low performance, and machine learning. The research on the application of machine learning algorithms for the early detection of at-risk students in a course is reviewed in more detail in Chapter 2. The research methodology and machine learning approach are presented in chapter three. The study's analysis, findings, and discussion are presented in chapter four. The study's summary, conclusion, and recommendation are presented in Chapter 5.

1.6 Delimitation of the Study

Even though it was recognized that there are other senior high schools in the Western North region, the study only included Bia Senior High School. The research is restricted to the self-reported perspectives of the information gathered from the school management. Additionally, the study's population is restricted to the students at the educational institution where the study was carried out. As a result, only the schools in the study area could be generalized from the findings of this study.

1.7 Limitations of the Study

The study's shortcomings were noted in a number of ways. The first is related to the failure to gather data at the planned rate of 90% return. The second has to do with the scant information regarding the demographics of students that the school management provides.

1.8 Definition of terms

Artificial Neural Network: This models the neuronal network that makes up the human brain in an effort to aid the computer in comprehending information and making judgments in a manner similar to that of a human. Configuring common computers to function like networked brain cells is the first step in the creation of ANNs.

Educational Data Mining: It entails gathering, processing, and analyzing data from educational systems to find trends or connections that could enhance educational outcomes.

Decision Tree: Employs a tree-like framework of choices and likely outcomes, such as chance event outcomes, resource costs, and utility, to aid in decision-making. This method may be used to demonstrate an algorithm that solely employs conditional control statements.

Linear Regression: A dependent variable's relationship to one or more independent variables is modeled using a statistical technique called linear regression. It bases its analysis on the supposition that the variables are inversely proportional, meaning that when the value of one variable changes, so does the value of the other.

At-risk student: A student who needs short-term or continuing support to flourish academically is referred. At-risk learners often exhibit psychological or behavioral issues, exhibit absenteeism, underperform in class, show little enthusiasm for learning, and seem disconnected from the school environment.

Predictive Modelling: This is a computational function for forecasting future occurrences or results by looking for patterns in a collection of input data.

CHAPTER TWO

LITERATURE REVIEW

2.0 Introduction

The study seeks to use machine learning techniques for the early detection of at-risk students in a course at the intermediate levels of education. A review of the literature is presented in chapter two on peer instruction, educational data mining, at-risk learners, students at-risk of low performance, and machine learning. The research on the application of machine learning algorithms for the early detection of at-risk students in a course is reviewed in more detail in this Chapter.

2.1 Machine Learning

The assumption that robots may learn on their own how to tackle a particular problem by being given access to the proper data is advanced by the artificial intelligence branch known as machine learning. Machine learning enables machines to execute independently intellectual tasks that have historically been addressed by humans by utilizing sophisticated statistical and mathematical techniques (Musumeci et al., 2019). After learning a set of rules from examples, machine learning is the process of creating a classifier that can be used to generalize from new cases (Bagaa et al., 2020). There are two steps involved in classifier generation. A provided training dataset is used to build the classification model in the first stage. This process is known as training (Al-Shabandar et al., 2019). Construction of categorization rules takes place in this step. The accuracy of the classification criteria is evaluated in the second phase, testing. If the classifier's accuracy is greater than the desired threshold, the classifier model developed in the first phase may be used for the classification of new data records (Er, 2012). Machine learning algorithms are required for training the classifier. Because these algorithms are frequently reliant on forecast accuracy, choosing the right one is crucial. Machine learning approaches incorporate certain features of the human

mind that allow us to solve exceedingly complex issues quicker than even the quickest computers (Numerous challenging issues, including speech recognition from text adaptive control), and mark-up estimate in the building sector has been effectively solved using machine learning approaches(Mair et al., 2000).

Machine learning is important because it helps companies identify customer behavior trends and simplifies the process of developing new items (Yarkoni & Westfall, 2017). Many of the most successful companies operating today, including Instagram, Amazon, and Uber, heavily rely on machine learning. Machine learning technology has become a crucial differentiator and competitive advantage for many firms. How an algorithm learns to improve its prediction accuracy is how traditional machine learning is typically classified (Bagaa et al., 2020). There are four fundamental strategies: unsupervised learning, supervised learning, semi-supervised learning, and reinforcement learning. One may use a number of different algorithms depending on the type of data that data scientists are attempting to forecast (Nielsen, 2019).

2.1.1 Supervised learning

With this sort of machine learning, data analysts offer the algorithms labeled training data and indicate the variables they want the computer to look for relationships between. Both the algorithm's input and output are described. Data analysts have to use labeled inputs and desired outputs to train the algorithm in supervised machine learning (Gajane & Pechenizkiy, 2018). For the following tasks, supervised learning approaches are effective: dividing information into two groups using a binary system. select from a variety of more than two response categories. The purpose of regression modeling is to forecast both continuous values and Assembling: combining the output from different machine learning models to produce an accurate forecast (Mair et al., 2000).

2.1.2 Unsupervised learning

The machine learning algorithms employed in this type of analysis are trained on unlabeled data. The software searches the data sets for any significant connections. The data utilized to train algorithms as well as the predictions they make are both predetermined (Rose, 2018). Machine learning algorithms that are unsupervised don't need labels on the input data. Unlabeled data is examined by analysts to identify tendencies that can be used to segment it (Goto et al., 2019). Unsupervised algorithms are used by neural networks and the great majority of deep learning methods (Zoabi et al., 2021). Unsupervised learning techniques perform well for the following tasks: Clustering is the process of separating a dataset into groups that have a similar appearance. Anomaly detection is the method of finding out-of-the-ordinary data points in a data collection. Association mining is the method of locating clusters of objects that commonly occur together in a data collection. Dimensionality reduction is the method of minimizing the number of variables included in a data source.

2.1.3 Semi-supervised learning

This strategy integrates two different kinds of machine learning. Although an algorithm may be fed primarily labeled training data, the algorithm is still free to independently explore the data and gain an understanding of the data set. Semi-supervised learning is carried out by data analysts using a small set of tagged training data (Mackenzie, 2015). In order for the algorithm to analyze fresh, unlabeled data, it has to know the dimensions of the data set. In general, algorithms work better when they are trained on sets of data containing labels (Seko et al., 2017). Semi-supervised learning satisfies both the performance of supervised learning and the efficacy of unsupervised learning. Machine translation, which trains algorithms to translate languages using less than a full dictionary of words, is one use of semi-supervised learning. Fraud identification is the process of finding instances of fraudulent activity when

there aren't many successful ones. Data labeling: After being educated on tiny data sets, algorithms can learn to automatically apply data labels to bigger sets.

2.1.4 Reinforcement learning

This is frequently used by data analysts to teach a machine to perform a multi-step process with precise criteria (Acharya & Sinha, 2014). An algorithm is programmed by data analysts to accomplish an outcome, and when it decides how to do so, it receives good or negative input. Yet, the algorithm typically decides for itself what to do at each level. (Weiss & Indurkhya, 1995). A few industries that frequently use reinforcement learning include robotics, where robots have been taught to perform tasks in the real world, video games, where robots have been taught to play a variety of video games, and resource management, where businesses can use reinforcement learning to assist them figure out how to allocate resources.

Nowadays, many different applications employ machine learning. The news feed on Instagram is powered by a recommendation algorithm, which is arguably one of the most well-known applications of machine learning (Pojon, n.d.). Instagram uses machine learning to customize each user's feed specifically. The predictive algorithm will begin to display more of that group's activity in the feed more quickly if a member regularly takes the time to halt and read the postings in that group. The algorithm is actively working in the background to reinforce recognized similarities in the person's online behavior. The news feed will update itself if a user's reading tastes alter and they don't check out posts from that group in days (Rawat & Malhan, 2019). Several scholars have employed various machine learning to carry out categorization in various knowledge domains over the years (Papamitsiou & Economides, 2014). Three machine learning algorithm classes have been selected for this research, and a representative algorithm from each class has been used for categorization (Yarkoni & Westfall, 2017). Three classes of machine learning algorithms are chosen to add some variety

to the classifiers, so they shouldn't all make the same or related mistakes. The rationale for selecting these machine learning algorithms is then addressed. Without making significant model assumptions, DT demonstrates the ability to model complicated interactions between variables. They can reduce modeling time when the data set is large because they do not require extensive training (Lotter et al., 2020). The algorithm has been selected to serve as a representative of this class. Because of its exceptional capacity for self-learning and self-adaptation, ANN has been employed in many challenging applications, including predicting share value and breast cancer, to mention a few (Żytkow & Sanjeev, 1998). They are frequently discovered to be more accurate and efficient than other classifier methods. The Multi-Layer Perceptron (MLP) is selected as the class's representative member.

2.2 At-risk Learners

At–risk students show less persistence and receive uneven and patchy praise for their successes. Delay is yet another significant problem with online classes. As a result, students who are enrolled in online learning frequently put off and rush through their homework. These students exhibit lower achievement and long-term retention, which is not surprising (Al-Mobayed et al., 2020; Lotter et al., 2020). Dyslexia, low self-esteem, poor use of ICT (information and communication technology) skills, a lack of tutor assistance and encouragement, infrequent online logins or communication, repeating a module after failing, and a lack of comprehensive formative assessment are additional factors associated with at-risk online learners. Further, (Lotter et al., 2020) discovered that at-risk students frequently switch up their study locations and exhibit poorer motivation, a severe internal locus of control, a lack of computer comfort, and little incentive to enroll in the course. These elements are the main cause of at-risk online learners' significantly higher rate of withdrawal from videoconferencing and web-based distance education, according to (Lizzio & Wilson, 2013). In light of these findings and the rising popularity of online learning, it is crucial to

develop an effective screener that enables teachers or educators to identify at-risk learners based on psychological perspectives early on. Additionally, early identification gives at-risk online learners enough time for possible guidance and even intervention, which can help them succeed in the virtual school setting.

2.3 Student At-Risk of Low Performance

The studies done in the field of learning analytics to forecast a student's academic success in terms of award-gap results, grade prediction, or overall average grades are examined in this area. Data on daily student behavior from virtual learning environments to forecast whether students will succeed in an educational program or fail. (Heuer & Breiter, 2018) According to the researchers, binary data had the same predictive value as the precise number of clicks. They compared the outcomes of four different supervised machine-learning algorithms. They additionally grouped learners based on their regular activities using K-means clustering. (Al-Shabandar et al., 2019) examined the elements of student participation that were closely associated with academic success. They noted that frequent course logins, the length of resources pay attention to, and repeated resource watch were significant determinants of student performance. The Logistic Regression model classified learners as good or not based on these characteristics. They discovered that the student's natural curiosity affected the success of the certification. In this area, their "learning effectiveness" approach performs better than alternatives. (Peña-Ayala, 2014) shown that the C4.5 classifier outperforms previous approaches in predicting whether a student would complete a course of study by taking into account household expenses and students' personal information in addition to academic performance indicators. An overview of the methods utilized in the field of learning analytics to determine the final score of a learner was presented by (Shahiri et al., 2015). They noticed that the most often utilized datasets are cumulative grade point averages and internal evaluations. Decision trees and neural networks were the most often used machine

learning algorithms. Logistic regression was employed by (Daimiel et al., 2020) to forecast whether a learner will obtain their online certificate. The authors estimated the likelihood that students would receive a certificate using the results of the homework and student interaction.

(Okubo et al., 2017) used student click stream information with the online institute platform to evaluate the effectiveness of Recurrent Neural Networks to conventional regression techniques when predicting final student grades. In comparison to other models, the usage of recurrent neural networks demonstrated superior predictive power. Early identification of students in ongoing classes is a crucial component of anticipating at-risk students so that intervention techniques can be applied to enhance their academic success in that course. Principal Component Regression was utilized by (Haiyang et al., 2018) to estimate student performance by utilizing data from the students, such as quiz and assignment scores. Only one-third of the semester had passed when their model began to predict students' performance. This first performance assessment can be used to identify students who require prompt attention in order to possibly turn around their circumstances. (Willging & Johnson, 2009) conducted in-depth research to pinpoint the elements that contribute to learning abandonment by students. The most significant indicators included trouble integrating with classmates, difficulty adjusting to the financial situation, age, gender, and type of teaching. This research was expanded further to determine the causes of the high dropout rates in online programs. To do this, survey data was integrated with demographics (such as age, gender, occupation, etc.). The poll asked a variety of questions about anything from why the student decided to participate in the program to what would have caused them to leave, including work obligations, a lack of teacher involvement, disappointing assignments, and much more. According to the study's findings, dropout predictors are comparable between face-to-face and online programs when there is no previous set of data. (Hlosta et al., 2018) identified at-risk students by leveraging the information from running presentations to train a

prediction model. From the actions of students who have already turned in their exams, learning patterns can be deduced.

According to (Aldowah et al., 2020), certain learning analytics techniques are most suited for particular learning issues, and using these techniques can assist in creating a student-focused approach to lowering the rate of dropping out of school. identifying at-risk learners in self-paced online courses utilizing self-regulated learning strategies. Even if only 30% of the course was completed, they anticipated that the students wouldn't finish it. (Chen et al., 2018) studied at-risk students in STEM courses across various areas. With a degree completion rate below 42%, the pace at which students drop out of STEM majors is concerning. For the purpose of quickly identifying learners who are in danger, they created a framework for survival analysis. The outcomes were encouraging and on par with more established machine learning methods like logistical regression, classification trees, and boosting. Even with fewer semester details, the methodology was effective, with characteristics like degree duration and grade-point average demonstrating good predictive ability. A time series-based prediction technique termed the Time Series Forest algorithm was put forth by (Haiyang et al., 2018). Students' actions and interactions in the learning environment made up the data they collected. They discovered that as the amount of daily data used to train the model is increased over time, its accuracy increases.

2.4 Peer Instruction

Peer instruction is a method of classroom instruction that is student-centered (Watkins & Mazur, 2013). For use in beginning Science classes, it was created in the 1990s, and mathematics and computer science eventually embraced it. It entails using questions with multiple responses in the classroom, to which students must first react individually and then again after a group discussion. Students often utilize clickers to provide their comments, including those in this study. In order to establish a shared understanding and demonstrate

instructor knowledge of the pertinent topic, instructors guide a class debate regarding the question after the two responses have been given. Numerous fields have demonstrated the effectiveness of peer instruction, including chemistry, life sciences, math, and IT (Mair et al., 2000; Nielsen, 2019; Wei et al., 2019).

This research showed that peer instruction supports learning even when none of the discussion group participants knew the right answer, as it improves students' learning and attitudes toward learning, increases student engagement and understanding, and increases learning (Khosla et al., 2010; Żytkow & Sanjeev, 1998). In the field of computing, PI has been demonstrated to improve student learning, reduce course failure rates, and boost major retention rates (Osborn, 2001). In addition, peer instruction works well across a range of computing course topics, including courses in programming, theory, and systems (Porter, 2013). Learners generate clicker data throughout every lesson by using peer instruction in conjunction with clickers. In the current work, we make use of this automatically generated data and don't need to gather or generate anymore.

2.5 Educational Data Mining

To comprehend student behavior patterns more thoroughly and learning environments, the new discipline of educational data mining uses statistical and machine learning techniques to evaluate a massive library of academic information. Finding out what influences the results of learners, failure, participation, and interaction on online learning platforms requires some steps. Several educational data mining research has been conducted. Some of these research studies also use demographic info track student learning (Waheed et al., 2018), but the majority of these studies focus on analyzing variables that are generated from students' online activities (Baneres et al., 2019; Shahiri et al., 2015). Studying frequency, length, subject matter, and social interaction-related parameters were formerly the main variables to be taken into consideration for analysis. Variables including click stream, examination results, task

results, and web forums involvement were introduced to the analytical process as the online learning platform became more dependable and interactive (Zacharis, 2015). Researchers find it difficult to identify relevant variables because of the variety of Learning Management Systems, massive open online courses, courses available, and course activity types. Literature focus on gathering data and forecasting how well learners perform at the end of the semester from failure and dropping out. On the contrary, from the very beginning of the course, online learning platforms produce a vast amount of data related to student involvement and courses, etc.

By examining information about variables from the beginning of the course, a thorough predictive model may be created that will let teachers intervene effectively and at the proper time, preventing failures and dropouts. Four machine learning methods were used in a study by (Mair et al., 2000) to identify pupils who are likely to fail early on. The Support Vector Machine, which had an accuracy rate of 79%, was found to be the most efficient method for identifying pupils in the past. The study also showed that improving the performance of machine learning algorithms requires significant data initial processing. Predictive model building is feasible earlier in the course, according to the findings of prior studies, but there are many obstacles that restrict their applicability to a particular learning platform. The various course structures, instructional styles, and online platforms are major factors preventing predictive models from being flexible, universal, and transferrable (Cicchinelli et al., 2018). In the past several years, research projects in both both formal and informal contexts for education have used statistical and predictive models to identify insights in a vast store of data (Waheed et al., 2018). For instance, various studies looked into the impact of demographic factors on student retention or successful learning performance. More than 200 distinct datasets pertaining to the demographics and online interaction patterns of undergraduate students studying economics and business were examined and analyzed in a

study by (Tempelaar et al., 2015) The study looked at the effects of many factors on students' success, including educational history, clickstream information, and exam results, entrance examination results, and studying character data.

While the majority of research focuses on determining how important factors affect the achievement of learners, some studies advocate for early intervention, knowledgeable support, and prompt feedback to help students who are at risk. Several studies conducted at Open University in the UK attempted to identify students who were at risk by using a variety of predictor variables (Cui et al., 2020). The studies used students' study habits to map out poor performance and Whether or not learners complete the course successfully or unsuccessfully. The experiments also showed that adding demographic information to behavior among learners' data enhanced the effectiveness and precision of model predictions. (Abumalloh et al., 2021) conducted a study in which they sought to pinpoint pertinent factors that influence students' decisions to drop the class. Three categories were created to group the factors influencing student dropout rates. First factors relating to the students' demographics, such as their gender, background, relevant experience, abilities, psychological characteristics, and prior schooling. Second, variables relating to the requirements and structure of the course, such as the number of assessments, institutional assistance, interaction, level of difficulty, and length of time. Third environmental/contextual aspects, such as the type of technology utilized, the area, the level of outside noise, the setting at work or home, etc. A time-series clustering method was used to identify at-risk students who signed up for online courses as soon as it was practical to do so (Huang & Fang, 2013). The time-series clustering method produced prediction models with greater accuracy compared to conventional aggregation procedures. (Buschetto Macarini et al., 2019) examined the effects of engagement characteristics derived from online course performance achievements using a variety of learning analytics methodologies. The findings showed that students with higher

academic standings have a higher engagement rate than students with lower academic standings.

By analyzing students' electronic book downloading patterns, (Harlim et al., 2021) created an early warning system that might identify learners who could struggle academically. 14 machine learning methods were utilized to train the model utilizing data from various weeks of the semester to create the best predicting model. According to the accuracy and Kappa measure, the best predictive model was chosen, and it suggested the ideal moment for teachers to step in. According to the study, all prediction models performed better as they were trained using progressively more weekly data. Starting in the third week, the early warning system predictive models had an accuracy of 85% in predicting low- and high-performance learners. While J48 outperformed all other algorithms when given the converted data, Random Forest outperformed other algorithms when given the full 16 weeks of data. Additionally, the Nave Bayes performed better with categorical data. Due to the variety in course structures and Massive Open Online Course designs, predicting students' performance early in the course can be a difficult challenge. Even if a learning management system is becoming more and more popular, there is still a need for an automated intervention system that can give learners immediate responses. Researchers have built a number of machine-learning techniques that can help teachers provide knowledgeable assistance to students during the learning process in order to combine an automated intervention system with a learning management system. In order to identify students' learning patterns, ML algorithms like Decision Trees, k-nearest-neighbors, Support Vector Machines, and random forest algorithm among others, are frequently trained using daily, weekly, or monthly student log data.

As a result of their ability to process raw data directly, deep learning algorithms are now also used to develop predictive models. (Minnich et al., 2016) used the Recurrent Neural Network algorithm trained on raw log student records to forecast how well learners performed at the end of the course. The use of genetic programming was made since it complements multi-view learning well. Without additional adjusting, the learned and evolved prediction model is directly explicable. Additionally, applying the genetic computing approach causes classification rules to naturally evolve as new data becomes available. In other words, categorization rules changed as new data became available. Underperforming and underrepresented learners are the focus of the early warning system, which was created using understandable Genetic Programming classification principles. This study's main flaw was the author's failure to specifically highlight the various semester phases at which performance markers, namely accuracy, sensitivity, specificity, and Kappa, were computed using a multi-view genetic programming algorithm in conjunction with other machine learning techniques.

The program uses the students' prior grades as input to simulate learning outcomes. In terms of precision, recall, specificity, and accuracy, the suggested dropout detection method outperformed feed-forward neural networks, Support Vector Machines, educational data mining systems, and Probabilistic Ensemble Simple Fuzzy Adaptive Resonance Theory Mapping methods. The dropout rate could be reduced by 20% by implementing a tutoring action plan based on logistic regression. (Lara et al., 2014) recommended employing knowledge discovery to pull information from databases that could help teachers comprehend how students interact with online learning environments. The method being discussed produces historical student reference models that can be utilized to determine whether or not learners are failures. The System for Educational Data Mining suggested system compares learners from the course who aced the exams and are eligible to take the final exam to dropout students who are unable to sit for the final test. For the two groups, The System for

Educational Data Mining was able to produce study patterns, which can be very beneficial for instructors in terms of enhancing students' study performance.

2.6 Empirical review

At-risk learners were found using a variety of machine learning methods, such as regularized logistic regression (LR), support vector algorithms, random forest, decision tree, and naive Bayes models (Buschetto Macarini et al., 2019). Some characteristics, such as the frequency of the learners who visited the home page and the duration of each session, were recorded from behavior log data. The outcomes showed that regularized logistic regression models had the best performance. An example of a deep neural network is the ConRec Network model, which was proposed by (Xing et al., 2015). Convolutional and recurrent neural networks were integrated in this study to predict whether students are likely to drop within the next seven days of virtual classroom. The records of learners were organized based on a chronology of timestamps and comprised information such as the date and time of the learners' enrolment as well as other qualities. The lower and top sections make up the two parts of the hybrid neural network model. Convolutional neural networks' hidden layer was used in the lower section to automatically extract features. Recurrent Neural Networks were utilized in the upper section to produce a forecast by merging and averaging the retrieved information each time. The model was contrasted with different industry standards. The findings showed that all models performed similarly. The authors claimed that despite similar performance, the ConRec Network model is more effective than baseline approaches because it can automatically extract features from student data without the requirement for the use of feature engineering.

Researchers have taken into account a number of factors to determine the degree of student achievement in the online environment, including the length of time students spend interacting with digital resources during assessment submission and the total number of

attempts made, educational level, geographic location, and gender. Genetic algorithms were used to enhance the feature set (Abumalloh et al., 2021). The results demonstrated that behavioral traits, not demographic traits, are associated with highly ranked features. Four classifiers DT, ANN, naive bayes, and k-nearest neighbor were used to predict student performance. The use of the feature set improved using genetic algorithms raised accuracy by 19%, according to simulation results. When using the genetic algorithms-optimized feature set instead of the decision tree with the full feature set, accuracy increased to 97% (Minaei-Bidgoli et al., 2003). The impact of latent variables in combination with observable variables on student results in online classrooms was evaluated using hidden Markov models. (Boughoula et al., 2017) suggested a two-layer hidden Markov model to deduce latent student behavior patterns. The ability of the two-layer hidden Markov model to more thoroughly identify the micro-behavioral patterns of students and spot the change from one latent state to another set it apart from standard HMM. For instance, students often join in forum discussions when they take quizzes. The quiz assessment date and the submission date are two distinct transitions that the model can learn. According to the study's findings, Less latent behavioral states are present in high-achieving students because they already have the necessary knowledge and do not require additional assistance.

The development of models for predicting students' success has received a lot of attention in Educational Data Mining research for a variety of reasons (Taruna & Pandey, 2014). Many of these studies formulated the problem of grade identification, such as class grade semester Grade, or graduation (i.e. final) grade, as a way to forecast the students' performance (Hamsa et al., 2016). Some works used students' grades or scores as criteria to create their difficulties in score prediction or performance classification. The most frequently used characteristics to predict students' performance in a typical classroom are their academic records. The academic performance history that was accessible prior to the start of the specific class was

employed in the majority of the studies in this field (Yang et al., 2018). These characteristics include GPA, cumulative GPA, and grades from prerequisite courses. Academic information gathered over the course of the semester, such as grades on assignments, participation in class activities, quizzes, and attendance, has also been used in numerous research (Hamsa et al., 2016). Although these characteristics can be quite useful in predicting course grades, they might not always be present. Additionally, it is important to note how these in-class evaluations vary from semester to semester.

Some studies have used information on the student's prior educational outcomes, including their enrollment results and middle school grades, as well as information about their high school, including its name, school type, and locality, since the existing academic curriculum lacks performance statistics, particularly for first-year students (Oladokun et al., 2008). Numerous works also made use of non-academic factors, such as demographic information about students, or information about students' interests and extracurricular activities (Almayan & Mayyan, 2016). To create the prediction model, several machine-learning approaches have been used DT, k-Nearest Neighbor, ANN, SVM, naive bayes, and logistic regression are examples of commonly used methods. (Kumar et al., 2020) created a Neuro-Fuzzy model for forecasting learning outcomes by combining the benefits of a neural network and a fuzzy system. The student's GPA and CGPA make up the features employed for the forecast in this study. Several researchers have revealed that combined models have promising potential.

The use of traditional statistical approaches in identifying at-risk students of quitting their field of study and leaving higher education has been studied recently in research activities. Based on enrollment data. (Golding & Donaldson, 2006) determined the connection between students' overall academic achievement and the likelihood of matriculation in the first year. One of the first researchers to use machine learning methods to investigate the student

retention issue (Żytkow & Sanjeev, 1998). Factor analysis and logistic regression were applied to a set of student attributes collected from data extracted by (Xing et al., 2015). In order to predict student retention, (Borko et al., 2008) evaluated numerous methods of data mining (decision trees, logistic regression, Bayesian classifiers, and neural networks) and found that logistic regression performed the best. Additionally, several contexts have employed logistic regression to predict early dropout rates. (Kovacic, 2012) used a variety of tree-based techniques and logistic regression to investigate the impact of demographic factors and the study environment on the outcome of student enrolment. Retention data was examined by (Nandeshwar et al., 2011), who came to the conclusion that giving high-risk student groups more attention and resources increases the likelihood that they will graduate from college. (Mohd & Yahya, 2018) used attributes collected from attendance reports, class test results, and submitted assignments to classify data using a decision tree. In an online course setting, (Albreiki et al., 2021) employed the k-nearest neighbor approach to predict student failure. None of these techniques can fully utilize the extra semesters in the data from the previous student cohort for early dropout prediction, even if they can all predict student dropout.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

In a course at the intermediate levels of education, the project aims to apply machine learning techniques for the early identification of students at risk. Weka software, which has built-in functions for the three approaches chosen for this research (linear regression, artificial neural network, and decision tree) is used to provide the research design and machine learning technique for the study in this Chapter.

3.2 Research Design

This study uses machine learning to create a model that forecasts at-risk students. Building a binary classification model with two classes for samples to fall into the successful class and the at-risk class is the major goal of this work. A single sample's variables are input attribution x = (X_1, X_2,..., X_n) and category attribution y; building a classification model entails developing a mapping function called y = f(x) that can be used to establish a sample's category attribution y based on its input attribution x. Figure 1 depicts the general layout of the study.

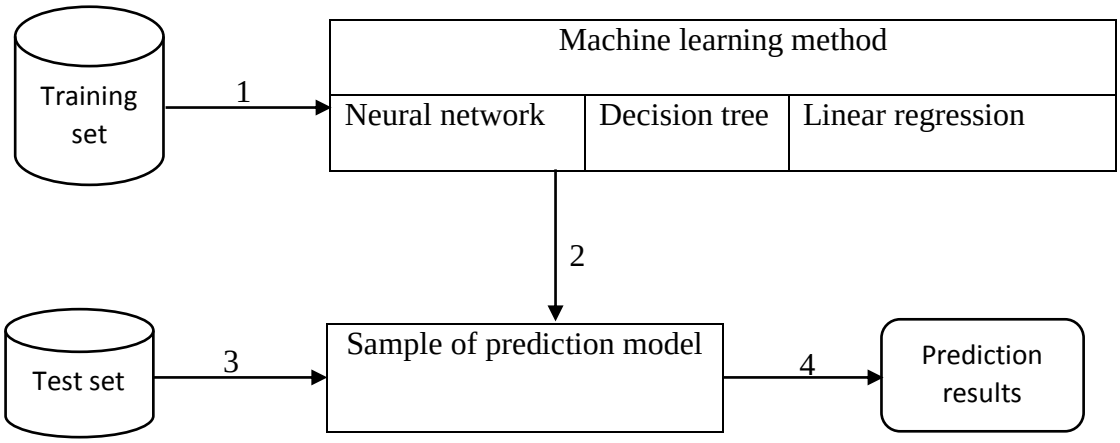

Figure 1: Study Framework

The study can be broken down into the four sections below:

1. Create the training data set by extracting student at-risk attribution data from the school data, and then load the data into the at-risk prediction model.
2. Create samples for the prediction model by using the data to train machine learning algorithms like linear regression, artificial neural networks, and decision trees.
3. In order to create a test data set, take another part of the data from the information systems and feed it into the actual samples of the previously created prediction models.
4. Use the samples from the prediction model to generate predictions on the test data set and evaluate how those predictions turned out.

3.3 Data Collection

One dataset is utilized in this research. This set is obtained from Bia Secondary School with a student population of 785. The dataset was collected from the entire student population enrolled in the schools. The demographic and academic dataset was collected from the school administration. Despite potentially being essential for comprehending student at-risk rates, data on the satisfaction of students, socioeconomic status, motivational factors, an individual fit of the institutional structure, attention to detail when choosing the program of study, and academic or societal student integration is not available. They are not accessible to every student at the intermediary level of education, and even if they were, data privacy and protection rules would prevent their use. Therefore, information regarding student demographics and academic achievement is used for the prediction.

Table 1: Variable descriptions of student's demographics

Serial Number	Name of variables	Meaning of variables	Type of variable
1	gdr	Gender	binary
2	age	Age	numerical
3	dbg	Day/Boarding	binary
6	fss	Financial support source	numerical
7	stl	study level	binary
8	hsta	health status	binary
9	esfs	support from school	binary
10	bece	BECE grade	numerical

Table 2: Variable descriptions of student's academic performance

Serial number	Name of variables	Meaning of variables	Type of variable
1	awtr	Average test written	numerical
2	hwtr	High test results	numerical
3	lwtr	Low test results	numerical
4	tsc	Number of courses studied	numerical
5	tpc	Number of courses passed	numerical
6	tfc	Number of courses failed	numerical
7	pte	Rate of pass	numerical

3.4 Data Analysis

3.4.1 Linear Regression Model

Regression analysis is a method for determining the statistical significance of a connection between a dependent variable and multiple independent variables. The fundamental multivariate linear regression model in this situation is

$$y_{it} = \beta_0 + \beta_1 x_i + \beta_2 z_{it} + \varepsilon_{it}$$

with student and terms represented by i and t, accordingly. The dependent variable, y_{it}, has two possible outcomes: successful (0) and at-risk (1). While data about performance, z_{it}, is time-variant, data on demographics, x_i, is time-constant. The ease of interpretation and improved comprehension of the significance and depth of the explanatory factors on the likelihood of at-risk are two benefits of the linear regression approach. The linear regression model's ability to predict values of the dependent variable that are less than zero and more than one is a drawback when it comes to estimating probability.

3.4.2 Artificial Neural Network (ANN)

Artificial intelligence (AI) research has a particular interest in behavioral simulation techniques. Brain research served as an inspiration for AI, which has been working since the 1955s to simulate the structure and operation of the human brain. Training algorithms were shown to be capable of calculating linearly separable functions. (Berens et al., 2018) employed the error-back propagation method for neural networks reviving artificial intelligence, which had been in a state of standstill. In addition to pattern recognition, image analysis, and detection of speech, higher-dimensional neural networks are being utilized to improve procedures, control systems, evaluate and forecast different results. In broad terms, the parameters, borders, and values of the chosen activation function are changed to learn the multilayer perceptrons. Artificial neural networks are computer models that mimic the neural

networks seen in animal brains. It is composed of links between the input and hidden layers as well as input and output layer components. In contrast to the variables of category attributions, which are represented by units in the output layer, each variable of the input attributions is represented by a unit in the input layer. Training is a process that modifies the weighting of inter-layer connections depending on the training utilizing categories of previously known information to create a more precise categorization of data with unknown categories.

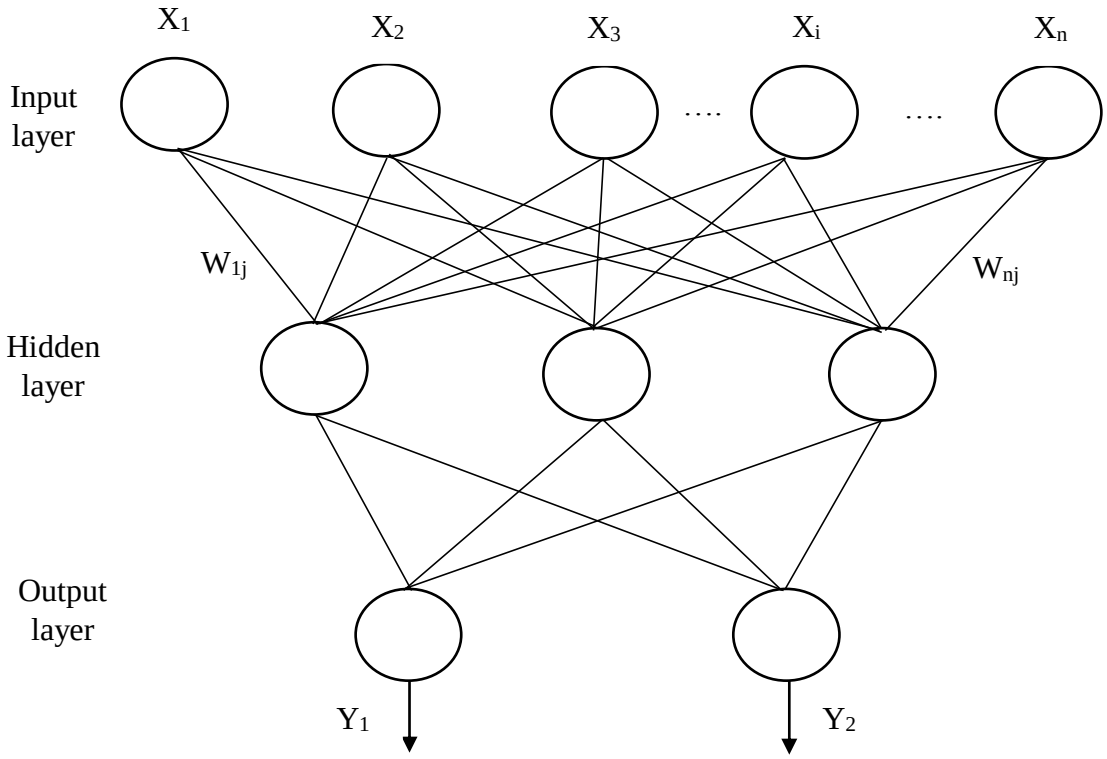

Figure 2: Algorithm model for artificial neural network

3.4.3 Decision tree

Using rules drawn from an existing data collection, a decision tree allocates entities (students) to one or more specified classes (successful and at-risk) of the goal variable. A decision tree builds itself by choosing observational attributes as nodes and building branches from each potential value for those attributes, iteratively (Zacharis, 2015). The process of choosing attributes is guided by the ideas of entropy and information gain. The decision tree algorithm

employs a top-down selection process, starting with the observed attribute that delivers the greatest amount of information gain. The root node is the property that has the best ability to predict the result. The observations are divided into progressively smaller data groupings by the root and subsequent nodes in order of predictive strength until all observations in one group have the same result or a pure group. Entropy assesses the homogeneity of results in a subset of the data, with 0 entropy corresponding to a subset that is entirely uniform and one with equal shares of all outcomes (Mair et al., 2000). Various decision tree algorithms produce predictions for the outcome variable across observation. You can find a summary of the most popular algorithms in (Pilotti et al., 2022). The researcher extends the well-known ID3 algorithm in this study by using the decision tree algorithm C4.5. It eliminates the ID3 constraint. The C4.5 builds trees utilizing knowledge gained recursively. This approach also makes use of improved choice of attributes and splitting. Decision trees have a tendency to overfit the data since they are a non-parametric machine-learning technique with a lot of adaptability.

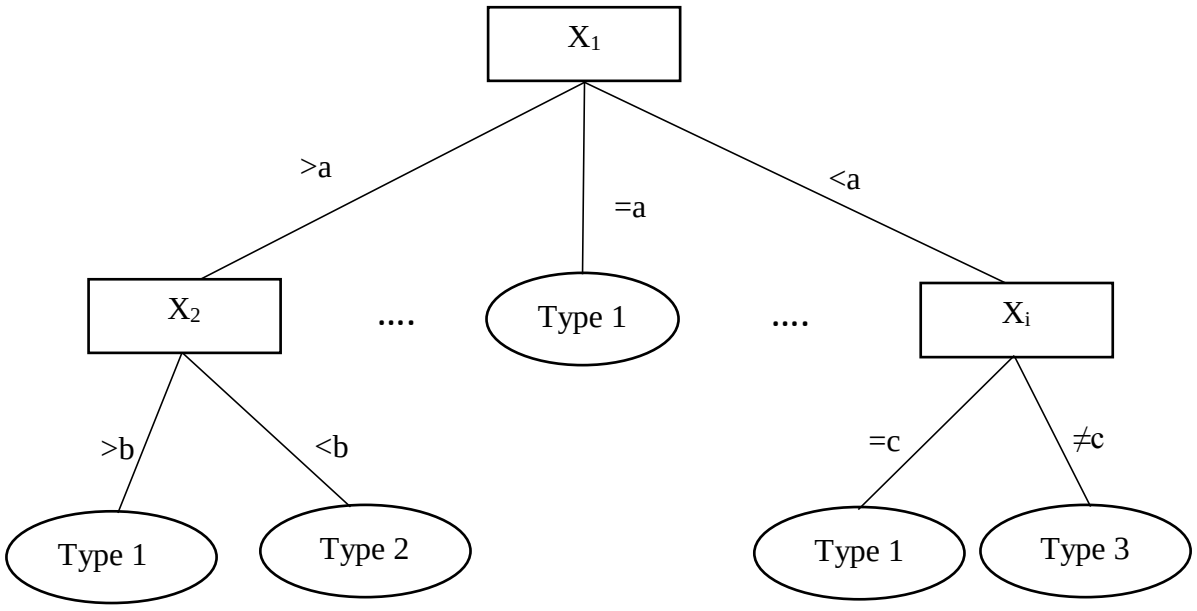

Figure 3: Decision Tree Algorithm Model

3.5 WAKA Software

WEKA software will be employed for this research project. Machine-learning algorithms for classification and grouping tasks in data analysis are widely available in WEKA software. Both direct dataset application and Java code calls for the WEKA algorithms are available. Data pre-processing, regression analysis, clustering, classification, rule-based association, and display are among the capabilities included in WEKA. WEKA is suitable for creating new machine-learning techniques as well. Data from ".arff" files are read into the program as input. WEKA can read ".csv" file formats in addition to its native ARFF data file format. This is beneficial since most spreadsheets and database software programs may export or store data as flat files in the.csv format (Hall et al., 2009). There is extra software available for combining many interconnected database tuples into a single table that can be handled using Weka, despite the fact that it cannot perform mega-relational analysis of data. The methods provided in the Weka distribution do not now cover sequence modeling, which is crucial.

3.6 Predictive Model's Performance

WEKA uses a confusion matrix to show how well the classification methods performed. Counting the number of false positives, false negatives, true positives, and true negatives is done using a predictive analytics table called a confusion matrix, which has two rows and columns. Any predictive model could make false negatives (under-prediction) or false positives (over-prediction) errors. An error matrix, or confusion matrix, is used to represent the percentages that constitute each of these errors (Fielding & Bell, 1997). The total number of instances in which the data meet the criteria for a positive result is (p). Condition negative (n) refers to the total number of genuine negative cases in the data. True positives (tp) are test results that reliably demonstrate the existence of a condition or trait. A "true negative" (tn) exactly demonstrates the absence of a characteristic in a test result. False positive (fp) is a test

result that incorrectly suggests the existence of a certain condition or trait. A false negative (fn) is a test outcome that untruthfully shows the absence of a certain condition or characteristic. The four elements of the confusion matrix are displayed in the table below.

Table 3*:* Confusion matrix

	At-risk prediction	**Successful prediction**
At-risk Students	True Positive (TP)	False Negative (FN)
Successful Students	False Positive (FP)	True Negative (TN)

A student who is appropriately rejected as an at-risk student, or truly negative, is a student whose success is correctly projected. As a result, a student who was accurately forecasted as being at danger was correctly identified as such, making it a true positive from the confusion matrix, derived. A confusion matrix is the name given to this matrix, which displays potential prediction outcomes (Lizzio & Wilson, 2013). These data can be used to determine various evaluation criteria. The first is accuracy, as stated by Powers (2011):

$$\text{Accuracy} = \frac{TP+TN}{\text{FP} + \text{FN} + \text{TP} + \text{TN}} \quad \text{..... Equation (1)}$$

Basically, accuracy is the proportion of accurate forecasts. Accuracy, however, has certain limitations when assessing the outcome of predictions. Particularly when the distribution of classes is unbalanced, accuracy does not reveal how the cases of the minority class are classified. One possible data set is one with 200 students in it, 180 of whom have passed the test. 90% of the time, the majority rule, a basic forecast that makes no use of machine learning but instead assumes that every student will pass the test, is correct. In comparison to simply assuming that each example belongs to the majority class, the model ought to perform better. Three more criteria are used in this research Precision and recall are two of them, and they are described as follows (Berens et al., 2018):

$$\text{Precision} = \frac{TP}{\text{TP} + \text{FP}} \quad \ldots\ldots\ldots\ldots\ldots\ldots \text{ Equation (2)}$$

$$\text{Recall} = \frac{TP}{\text{TP} + \text{FN}} \quad \ldots\ldots\ldots\ldots\ldots\ldots \text{ Equation (3)}$$

Recall and precision are used to create a superior assessment. The essential point is that making precise predictions about a favorable result is insufficient. Successful positive and unsuccessful negative predictions must be present in a solid predictive model. The F-measure, which is the third criterion this research uses, is defined as follows by Fawcett (2005):

$$\text{F} - \text{measure} = \frac{2 * Recall * Precision}{\text{Precision} + \text{Recall}} \quad \ldots\ldots\ldots\ldots\ldots\ldots \text{ Equation (4)}$$

F-measure is a means to have a single value that accounts for both recall and precision. The last criterion for comparing in this research is the F-measure.

CHAPTER FOUR

RESULTS, FINDINGS, AND DISCUSSION

4.0 Introduction

The research's outcomes are reported in this chapter using the effectiveness of machine learning algorithmic techniques. Accuracy, recall, and precision are the three evaluation criteria that have been used to assess how well each technique performs in identifying successful or at-risk students.

4.1 Predicted Models

Predictive modeling is the technique of forecasting future circumstances and actions that utilize a model, find in previously unknown data developed from similar prior data (Schneider et al., 2018). It has several applications in a range of fields, including finance, skill development, medicine, and the legal system (Nandeshwar et al., 2011; Willging & Johnson, 2009). The application procedure is the same across all of these disciplines. Using previously obtained data, a machine learning technique establishes the correlations between different data properties. Based on characteristics, the generated model is capable of foretelling one of a new dataset's features (Mackenzie, 2015).

Building a prediction model using known data is known as training, and the data used in this process is known as the test dataset or training dataset. After being built, the model needs to be evaluated on another data set to determine how well it performs. To make sure the model is adaptable enough to be used other than the one it was constructed with, two other sets of data are employed. Otherwise, the issue of over-fitting could develop, which happens when a model performs well with its original data set but badly on additional data sets due to its excessive complexity (Quinn & Gray, 2020). Training and test sets are often generated from the input data set to lower prediction errors. The model is used to predict the dependent

variable for the test set in order to evaluate the model using test data. The dependent variable's anticipated and actual values are then compared. Comparing assessment to the overall number of accurate forecasts is more difficult.

4.2 Correctness of Classifiers

Figure 4 displays the forecasting accuracy for the Decision Tree, Artificial Neural Network, and linear regression before describing the results for the various classifiers. Each approach calculates a student's at-risk likelihood, which ranges from 0 (successful) to 1 (at-risk). Probabilities that are close to 0 or 1 when predicting at-risk are correct. The accuracy of forecasts at the identification threshold is unclear. The precision is demonstrated in Figure 4. The proportion of accurate predictions among all predictions is, as expected, lowest when the threshold is approached. This is true for all classifiers, but the Decision Tree performs better, although not throughout the entire range of observations when compared to Linear Regression and Artificial Neural Network. Particularly for students with risks just beyond the threshold, the DT's accuracy is greater.

The models had to be built after the test and training sets had been made. The linear regression approach was used to develop the initial model. The major steps in creating a model with WEKA include defining the input data, dependent variables, and independent variables. The test data collection is used to apply the model after it has been constructed. A confusion matrix is the result of this process, which is relevant to the research. It includes information on both actual and expected values. The confusion matrix for the first prediction model is displayed in Table 4. 93% is the accuracy determined using the confusion matrix. Compared to the baseline accuracy of 74%, this is an improvement. Findings are assessed once again.

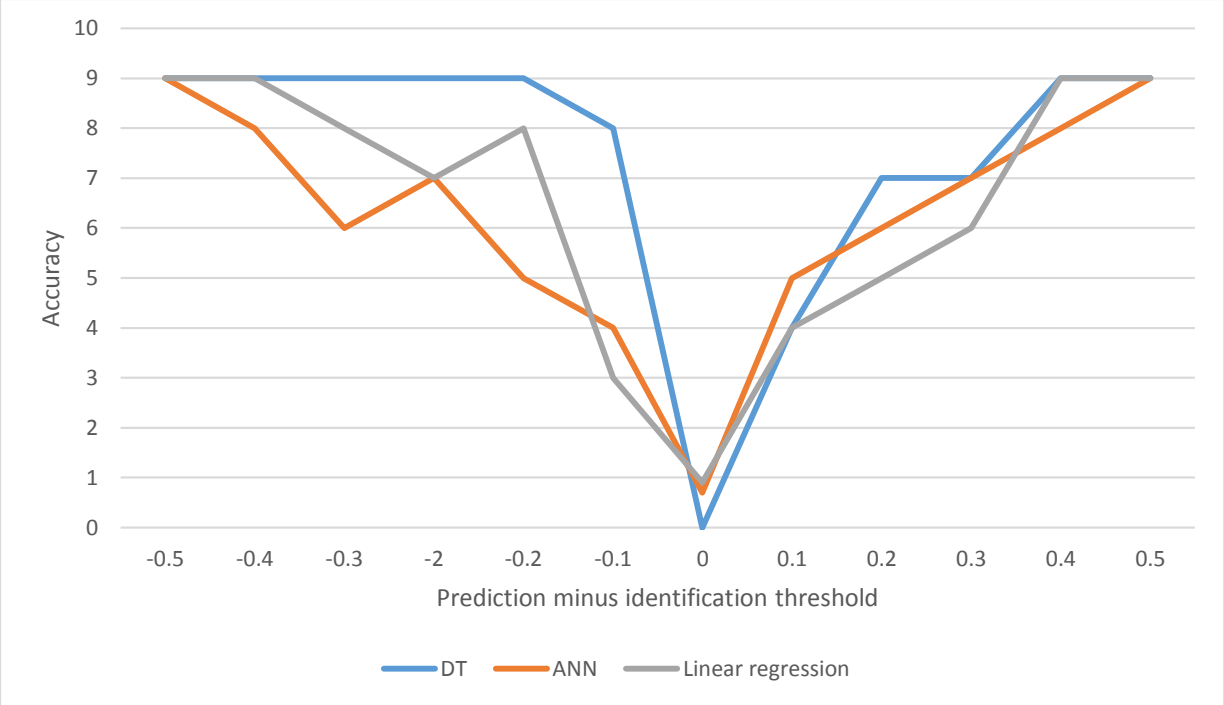

Figure 4: Prediction Accuracy

4.3 Results of Data Set from Bia Secondary School

Applying machine learning techniques to the raw data, including student demographics and academic performance, was the initial step. The dependent variables were changed in this instance to become binary, which constitutes the only data processing that took place. The majority rule is 74% accurate because 584 out of 785 students performed adequately or satisfactorily. The accuracy of prediction models constructed using this data set was compared to the standard accuracy for this data set in order to determine whether the models were capable of producing relevant predictions. The data had to be split into training and test sets after the standard accuracy had been determined.

The prediction model was constructed using a training set that represented 75% of the data, and a test set that represented 25% of the data. It is crucial to keep in mind when creating training and test sets that each must have about equal proportions of students from both classes. A built-in feature of the WEKA software makes sure that the cases of various classes are distributed fairly among the training and test sets.

Table 4: For the raw data set, a confusion matrix for linear regression was employed.

	False prediction	True Prediction
False (Actual)	26	6
True (Actual)	2	86

The decision tree approach was used to generate the second model. This model's WEKA function implements the CART classification algorithms. The process is the same as with the prior model, except for the function used. A model was developed using the training set, then applied to the test set after creating training and test sets. The confusion matrix for this model is displayed in Table 5. This model's accuracy, which is 93%, is identical to that of the preceding model.

Table 5: CART confusion matrix applied to the raw data set

	False Prediction	True prediction
False (Actual)	24	8
True (Actual)	0	88

The Artificial Neural Network (ANN) classification technique is used to create the final model for this data set. The model's confusion matrix is displayed in Table 6. The accuracy of this confusion matrix is 95%.

Table 6: For the raw data set, an ANN confusion matrix was applied.

	False Prediction	True Prediction
False (Actual)	28	4
True (Actual)	1	87

After the complete data set was split into a training set and a test set in a ratio of 7:3, the test set had 521 samples. Utilizing examples from three different machine learning approaches, the test data set was categorized. The predicted results are shown in Table 7.

Table 7: Results of the confusion matrix's prediction

	Linear Regression		ANN		DT	
	Successful	At-risk	Successful	At-risk	Successful	At-risk
Successful	372	19	451	27	471	20
At-risk	38	302	29	43	46	174
Total	**410**	**321**	**480**	**70**	**517**	**194**

The outcomes of the prediction were assessed using the abovementioned evaluation techniques in Table 8. The three prediction models performed similarly in relation to the precision rate and recall rate of the successful students: Artificial Neural Network (ANN) had the best precision rate (98.88%), followed by Decision Tree (DT) (98.37%), and Linear regression had the lowest precision rate (97.30%). The recall rate is ranked by Artificial Neural Network ANN (98.88%), Linear regression (98.37%), and Decision Tree (DT) 97.30% in ascending order. There are some disparities between the three prediction models in terms of the precision rate and recall rate for the at-risk students: The Decision Tree (DT) had the best accuracy (63.89%), followed by the linear regression (63.39%), while the Artificial Neural Network (ANN) had the least precision rate (53.54%). The prediction model's total efficacy is indicated by the overall accuracy rate. The overall accuracy rate is ranked as follows: Decision Tree (94.63%), Artificial Neural Network (93.97%), and linear regression (93.92%), with all three models having relatively good overall accuracy rates that were above 93%.

Table 8: Evaluation of prediction outcomes

Analysis Index	ANN	DT	Linear Regression
Precision Rate of Successful Students	98.88%	98.37%	97.30%
Precision Rate of At-risk Students	53.54%	63.89%	63.39%
Recall Rate of Successful Students	94.63%	95.76%	95.67%
Recall Rate of At-risk Students	84.85%	82.22%	76.15%
Overall Effectiveness	93.97%	94.63%	93.92%
F-measure of at-risk Students	65.65%	71.91%	69.19%

Finding potential at-risk students was the goal of this research. The at-risk students' specific F-measure score represents the general efficacy of the prediction models in predicting the at-risk students. The models are ranked Decision Tree (71.91%), Linear regression (69.19%), and Artificial Neural Network (65.65%) in decreasing order. The DT was, by contrast, the most effective and was more accurate in predicting the at-risk students. In summary, the three prediction models were all relatively good at screening prospective risky students.

The accuracy of all three prediction models was higher for the successful students than the at-risk students due to the incomplete data attributions. A similar study suggests that a range of variables, including personal qualities, may have an impact on students who are at risk of failing. Because of the comprehensiveness of the data collected from the academic administration systems of the educational institutions, the study's restrictions prevented the researcher from using more than only students' personal attributes and academic performance as input variables in the predictive model. This affected the predictive model's accuracy. The precision of the forecast may also be improved through advancements in machine learning algorithm techniques. In contrast to the use of an integrated multi-model technique, which

will help to enhance the precision to some level, we only used one model for prediction in this work.

CHAPTER FIVE

SUMMARY, CONCLUSION, AND RECOMMENDATION

5.0 Summary

Enhancing student retention rates and improving educational institution efficacy are all made possible by the students' performance prediction. Intervention initiatives in schools assist students who are at risk of dropping out of school. Accurately identifying and prioritizing the students who need help is key to the success of such initiatives. As discussed in chapter one, this study investigates the issues relating to students who are at risk at the intermediary level of education, using machine learning methods for early identification through the use of student demographics and academic information.

The reviewed literature covered Educational Data Mining, students at-risk of low performance, and machine learning. As a result, machine learning has the ability to advance education more quickly, and it is evident that education is becoming much more efficient. Teaching and learning will be radically altered by the right and effective application of machine learning techniques in the educational sphere. In order to aid difficult students early and take action to increase success and retention, educators that use machine learning will acquire a better picture of how their students are developing with their learning.

This study has shown that early detection based on student performance is important in identifying the necessary corrective actions. On the other hand, corrective measures are conducted while utilizing technological innovations and course characteristics. The research also showed that Decision Tree was the typical early detection algorithm and the corrective algorithm because most earlier studies had used Decision Tree for the purpose and had shown meaningful outcomes. Finally, student demographic and academic data were used to apply

Decision Tree, ANN, and Linear regression methods for performance and during at-risk predictions.

5.1 Conclusion

By summarizing and analyzing pertinent literature on the factors affecting successful students and at-risk students in addition to the student's attribution data stored in the information system of the educational institution, this study used the demographics and academic performance of students as the prediction model's input attributions. Three machine learning techniques Artificial Neural Network, Decision Tree, and linear regression were used to forecast whether a student will succeed or encounter risk factors. The Decision Tree provided the most accurate predictions, but the results showed that all three prediction models could predict student behavior to some extent.

The researcher conducted successful and at-risk predictions on current students at the Bia Secondary School using the methodology suggested in this study. For the school to take targeted retention measures to keep the potential students before dropout behavior happens, the divisions associated with students' learning support services were given the list of students who are anticipated to be at risk of dropping out.

5.2 Recommendations

Based on the findings, the following suggestions are made:

1. Due to their direct impact, group projects have a favorable link with a student's final grade. Therefore, it is advised that teachers encourage students to participate more actively in group projects.
2. Given that financial support and academic performance are positively correlated, it is advised that educational institutions award bursaries to deserving candidates.
3. Booster or support classes for students who are at risk should be made available.

5.3 Suggestion for Further Research

The improvement of prediction accuracy is the main objective of research on student at-risk identification. Future studies could be done to strengthen attributions and enhance algorithms given this purpose. It is first possible to collect information on learning behavior from the learning management system to enhance the input attributions for the predictive model, which will improve prediction accuracy. Second, machine learning techniques can be enhanced to boost forecast accuracy. One example is the usage of the combined model. The method of feature creation is the third area that may need improvement. Due to a lack of data, there is a limit on the number of feature adjustments that may be done. The research's data source was a single table, and variables from that database were used to create custom variables. A larger data set with more tables will make it easier to create new customized variables, but bear in mind that the more customized a variable is, the harder it is to comprehend how it interacts with the dependent variable.

References

Abumalloh, R. A., Asadi, S., Nilashi, M., Minaei-Bidgoli, B., Nayer, F. K., Samad, S., Mohd, S., & Ibrahim, O. (2021). The impact of coronavirus pandemic (COVID-19) on education: The role of virtual and remote laboratories in education. *Technology in Society*, *67*, 101728. https://doi.org/10.1016/j.techsoc.2021.101728

Acharya, A., & Sinha, D. (2014). Early Prediction of Students Performance using Machine Learning Techniques. *International Journal of Computer Applications*, *107*(1), 37–43. https://doi.org/10.5120/18717-9939

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques. *Education Sciences*, *11*(9), Article 9. https://doi.org/10.3390/educsci11090552

Aldowah, H., Al-Samarraie, H., Alzahrani, A. I., & Alalwan, N. (2020). Factors affecting student dropout in MOOCs: A cause and effect decision- making model. *Journal of Computing in Higher Education*, *32*(2), 429–454. https://doi.org/10.1007/s12528-019 09241-y

Almayan, H., & Mayyan, W. A. (2016). Improving accuracy of students' final grade prediction model using PSO. *2016 6th International Conference on Information Communication and Management (ICICM)*, 35–39. https://doi.org/10.1109/INFOCOMAN.2016.7784211

Al-Mobayed, A. A., Al-Madhoun, Y. M., Al-Shuwaikh, M. N., & Abu-Naser, S. S. (2020). Artificial Neural Network for Predicting Car Performance Using JNN. *International Journal of Engineering and Information Systems (IJEAIS)*, *4*(9), 139–145.

Al-Shabandar, R., Hussain, A. J., Liatsis, P., & Keight, R. (2019). Detecting At-Risk Students With Early Interventions Using Machine Learning Techniques. *IEEE Access*, *7*, 149464–149478. https://doi.org/10.1109/ACCESS.2019.2943351

Baek, C., & Doleck, T. (2021). Educational Data Mining versus Learning Analytics: A Review of Publications From 2015 to 2019. *Interactive Learning Environments*, *0*(0), 1–23. https://doi.org/10.1080/10494820.2021.1943689

Bagaa, M., Taleb, T., Bernabe, J. B., & Skarmeta, A. (2020). A Machine Learning Security Framework for Iot Systems. *IEEE Access*, *8*, 114066–114077. https://doi.org/10.1109/ACCESS.2020.2996214

Baker, R. S., & Siemens, G. (n.d.). *Learning Analytics and Educational Data Mining*. 29.

Baneres, D., Rodríguez-Gonzalez, M. E., & Serra, M. (2019). An Early Feedback Prediction System for Learners At-Risk Within a First-Year Higher Education Course. *IEEE Transactions on Learning Technologies*, *12*(2), 249–263. https://doi.org/10.1109/TLT.2019.2912167

Berens, J., Schneider, K., Görtz, S., Oster, S., & Burghoff, J. (2018). Early Detection of Students at Risk – Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3275433

Borko, H., Jacobs, J., Eiteljorg, E., & Pittman, M. E. (2008). Video as a tool for fostering productive discussions in mathematics professional development. *Teaching and Teacher Education*, *24*(2), 417–436. https://doi.org/10.1016/j.tate.2006.11.012

Boughoula, A., Geigle, C., & Zhai, C. (2017). A Probabilistic Approach for Discovering Difficult Course Topics Using Clickstream Data. *Proceedings of the Fourth (2017)*

ACM Conference on Learning @ Scale, 303–306. https://doi.org/10.1145/3051457.3054010

Buschetto Macarini, L. A., Cechinel, C., Batista Machado, M. F., Faria Culmant Ramos, V., & Munoz, R. (2019). Predicting Students Success in Blended Learning—Evaluating Different Interactions Inside Learning Management Systems. *Applied Sciences*, *9*(24), Article 24. https://doi.org/10.3390/app9245523

Chen, Y., Johri, A., & Rangwala, H. (2018). Running out of STEM: A comparative study across STEM majors of college students at-risk of dropping out early. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 270–279. https://doi.org/10.1145/3170358.3170410

Cicchinelli, A., Veas, E., Pardo, A., Pammer-Schindler, V., Fessl, A., Barreiros, C., & Lindstädt, S. (2018). Finding traces of self-regulated learning in activity streams. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 191–200. https://doi.org/10.1145/3170358.3170381

Cui, Y., Chen, F., & Shiri, A. (2020). Scale up predictive models for early detection of at-risk students: A feasibility study. *Information and Learning Sciences*, *121*(3/4), 97–116. https://doi.org/10.1108/ILS-05-2019-0041

Daimiel, L., Martínez-González, M. A., Corella, D., Salas-Salvadó, J., Schröder, H., Vioque, J., Romaguera, D., Martínez, J. A., Wärnberg, J., Lopez-Miranda, J., Estruch, R., Cano-Ibáñez, N., Alonso-Gómez, A., Tur, J. A., Tinahones, F. J., Serra-Majem, L., Micó-Pérez, R. M., Lapetra, J., Galdón, A., … Ordovás, J. M. (2020). Physical fitness and physical activity association with cognitive function and quality of life: Baseline cross-sectional analysis of the PREDIMED-Plus trial. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-59458-6

Er, E. (2012). Identifying At-Risk Students Using Machine Learning Techniques: A Case Study with IS 100. *International Journal of Machine Learning and Computing*, 476 480. https://doi.org/10.7763/IJMLC.2012.V2.171

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, *24*(1), 38–49. https://doi.org/10.1017/S0376892997000088

Gajane, P., & Pechenizkiy, M. (2018). *On Formalizing Fairness in Prediction with Machine Learning* (arXiv:1710.03184). arXiv. https://doi.org/10.48550/arXiv.1710.03184

Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, *28*, 68–84. https://doi.org/10.1016/j.iheduc.2015.10.002

Golding, P., & Donaldson, O. (2006). Predicting Academic Performance. *Proceedings. Frontiers in Education. 36th Annual Conference*, 21–26. https://doi.org/10.1109/FIE.2006.322661

Goto, T., Camargo, C. A., Jr, Faridi, M. K., Freishtat, R. J., & Hasegawa, K. (2019). Machine Learning–Based Prediction of Clinical Outcomes for Children During Emergency Department Triage. *JAMA Network Open*, *2*(1), e186937. https://doi.org/10.1001/jamanetworkopen.2018.6937

Haiyang, L., Wang, Z., Benachour, P., & Tubman, P. (2018). A Time Series Classification Method for Behaviour-Based Dropout Prediction. *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, 191–195. https://doi.org/10.1109/ICALT.2018.00052

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18. https://doi.org/10.1145/1656274.1656278

Hamsa, H., Indiradevi, S., & Kizhakkethottam, J. J. (2016). Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm. *Procedia Technology*, *25*, 326–332. https://doi.org/10.1016/j.protcy.2016.08.114

Harlim, J., Jiang, S. W., Liang, S., & Yang, H. (2021). Machine learning for prediction with missing dynamics. *Journal of Computational Physics*, *428*, 109922. https://doi.org/10.1016/j.jcp.2020.109922

Heuer, H., & Breiter, A. (2018). *Student Success Prediction and the Trade-Off between Big Data and Data Minimization*. Gesellschaft für Informatik e.V. http://dl.gi.de/handle/20.500.12116/21041

Hlosta, M., Herrmannova, D., Vachova, L., Kuzilek, J., Zdrahal, Z., & Wolff, A. (2018). *Modelling student online behaviour in a virtual learning environment* (arXiv:1811.06369). arXiv. https://doi.org/10.48550/arXiv.1811.06369

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education*, *61*, 133–145. https://doi.org/10.1016/j.compedu.2012.08.015

Khor, E. T. (2022). A data mining approach using machine learning algorithms for early detection of low-performing students. *The International Journal of Information and Learning Technology*, *39*(2), 122–132. https://doi.org/10.1108/IJILT-09-2021-0144

Khosla, A., Cao, Y., Lin, C. C.-Y., Chiu, H.-K., Hu, J., & Lee, H. (2010). An integrated machine learning approach to stroke prediction. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 183 192. https://doi.org/10.1145/1835804.1835830

Kovacic, Z. (2012). *Predicting student success by mining enrolment data.* https://repository.openpolytechnic.ac.nz/handle/11072/1486

Kumar, S. N., Saxena, P., Patel, R., Sharma, A., Pradhan, D., Singh, H., Deval, R., Bhardwaj, S. K., Borgohain, D., Akhtar, N., Raisuddin, S., & Jain, A. K. (2020). Predicting risk of low birth weight offspring from maternal features and blood polycyclic aromatic hydrocarbon concentration. *Reproductive Toxicology*, *94*, 92–100. https://doi.org/10.1016/j.reprotox.2020.03.009

Lara, J. A., Lizcano, D., Martínez, M. A., Pazos, J., & Riera, T. (2014). A system for knowledge discovery in e-learning environments within the European Higher Education Area – Application to student data from Open University of Madrid, UDIMA. *Computers & Education*, *72*, 23–36. https://doi.org/10.1016/j.compedu.2013.10.009

Lizzio, A., & Wilson, K. (2013). Early intervention to support the academic recovery of first year students at risk of non-continuation. *Innovations in Education and Teaching International*, *50*(2), 109–120. https://doi.org/10.1080/14703297.2012.760867

Lotter, W., Kreiman, G., & Cox, D. (2020). A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nature Machine Intelligence*, *2*(4), Article 4. https://doi.org/10.1038/s42256-020-0170-9

Mackenzie, A. (2015). The production of prediction: What does machine learning want? *European Journal of Cultural Studies*, *18*(4–5), 429–445. https://doi.org/10.1177/1367549415577384

Mair, C., Kadoda, G., Lefley, M., Phalp, K., Schofield, C., Shepperd, M., & Webster, S. (2000). An investigation of machine learning based prediction systems. *Journal of Systems and Software*, *53*(1), 23–29. https://doi.org/10.1016/S0164-1212(00)00005-4

Makombe, F., & Lall, M. (2020). A Predictive Model for the Determination of Academic Performance in Private Higher Education Institutions. *International Journal of Advanced Computer Science and Applications*, *11*(9). https://doi.org/10.14569/IJACSA.2020.0110949

Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., & Punch, W. F. (2003). Predicting student performance: An application of data mining methods with an educational Web-based system. *33rd Annual Frontiers in Education, 2003. FIE 2003.*, *1*, T2A-13. https://doi.org/10.1109/FIE.2003.1263284

Minnich, A., Abu-El-Rub, N., Gokhale, M., Minnich, R., & Mueen, A. (2016). ClearView: Data cleaning for online review mining. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 555–558.

Mohd, N., & Yahya, Y. (2018). *A Data Mining Approach for Prediction of Students' Depression Using Logistic Regression And Artificial Neural Network*. 1–5. https://doi.org/10.1145/3164541.3164604

Musumeci, F., Rottondi, C., Nag, A., Macaluso, I., Zibar, D., Ruffini, M., & Tornatore, M. (2019). An Overview on Application of Machine Learning Techniques in Optical Networks. *IEEE Communications Surveys & Tutorials*, *21*(2), 1383–1408. https://doi.org/10.1109/COMST.2018.2880039

Nandeshwar, A., Menzies, T., & Nelson, A. (2011). Learning patterns of university student retention. *Expert Systems with Applications, 38*(12), 14984–14996. https://doi.org/10.1016/j.eswa.2011.05.048

Nielsen, A. (2019). *Practical Time Series Analysis: Prediction with Statistics and Machine Learning*. O'Reilly Media, Inc.

Okubo, F., Yamashita, T., Shimada, A., & Ogata, H. (2017). A neural network approach for students' performance prediction. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 598–599. https://doi.org/10.1145/3027385.3029479

Oladokun, V. O., Adebanjo, A. T., & Charles-Owaba, O. E. (2008). *Predicting students academic performance using artificial neural network: A case study of an engineering course*. http://ir.library.ui.edu.ng/handle/123456789/1796

Osborn, V. (2001). Identifying at-risk students in videoconferencing and web-based distance education. *American Journal of Distance Education*, *15*(1), 41–54. https://doi.org/10.1080/08923640109527073

Papamitsiou, Z., & Economides, A. A. (2014). Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. *Journal of Educational Technology & Society*, *17*(4), 49–64.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, *41*(4, Part 1), 1432–1462. https://doi.org/10.1016/j.eswa.2013.08.042

Pilotti, M. A. E., Nazeeruddin, E., Nazeeruddin, M., Daqqa, I., Abdelsalam, H., & Abdullah, M. (2022). Is Initial Performance in a Course Informative? Machine Learning

Algorithms as Aids for the Early Detection of At-Risk Students. *Electronics*, *11*(13), Article 13. https://doi.org/10.3390/electronics11132057

Pojon, M. (n.d.). *Using Machine Learning to Predict Student Performance*. 39.

Porter, S. R. (2013). Self-Reported Learning Gains: A Theory and Test of College Student Survey Response. *Research in Higher Education*, *54*(2), 201–226. https://doi.org/10.1007/s11162-012-9277-0

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), Article 1. https://doi.org/10.22554/ijtel.v5i1.57

Rawat, K. S., & Malhan, I. V. (2019). A Hybrid Classification Method Based on Machine Learning Classifiers to Predict Performance in Educational Data Mining. In C. R. Krishna, M. Dutta, & R. Kumar (Eds.), *Proceedings of 2nd International Conference on Communication, Computing and Networking* (pp. 677–684). Springer. https://doi.org/10.1007/978-981-13-1217-5_67

Rose, S. (2018). Machine Learning for Prediction in Electronic Health Data. *JAMA Network Open*, *1*(4), e181404. https://doi.org/10.1001/jamanetworkopen.2018.1404

Schneider, K., Berens, J., Oster, S., & Burghoff, J. (2018). *Early Detection of Students at Risk—Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods*. *D20-V1*. http://hdl.handle.net/10419/181544

Seko, A., Hayashi, H., Nakayama, K., Takahashi, A., & Tanaka, I. (2017). Representation of compounds for machine-learning prediction of physical properties. *Physical Review B*, *95*(14), 144110. https://doi.org/10.1103/PhysRevB.95.144110

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, *72*, 414 422. https://doi.org/10.1016/j.procs.2015.12.157

Taruna, S., & Pandey, M. (2014). An empirical analysis of classification techniques for predicting academic performance. *2014 IEEE International Advance Computing Conference (IACC)*, 523–528. https://doi.org/10.1109/IAdCC.2014.6779379

Tempelaar, D. T., Rienties, B., & Giesbers, B. (2015). In search for the most informative data for feedback generation: Learning analytics in a data-rich context. *Computers in Human Behavior*, *47*, 157–167. https://doi.org/10.1016/j.chb.2014.05.038

Waheed, H., Hassan, S.-U., Aljohani, N. R., & Wasif, M. (2018). A bibliometric perspective of learning analytics research landscape. *Behaviour & Information Technology*, *37*(10–11), 941–957. https://doi.org/10.1080/0144929X.2018.1467967

Watkins, J., & Mazur, E. (2013). Retaining Students in Science, Technology, Engineering, and Mathematics (STEM) Majors. *Journal of College Science Teaching*, *42*(5), 3641.

Wei, J., Chu, X., Sun, X.-Y., Xu, K., Deng, H.-X., Chen, J., Wei, Z., & Lei, M. (2019). Machine learning in materials science. *InfoMat*, *1*(3), 338–358. https://doi.org/10.1002/inf2.12028

Weiss, S. M., & Indurkhya, N. (1995). Rule-based Machine Learning Methods for Functional Prediction. *Journal of Artificial Intelligence Research*, *3*, 383–403. https://doi.org/10.1613/jair.199

Willging, P. A., & Johnson, S. D. (2009). Factors that Influence Students' Decision to Dropout of Online Courses. *Journal of Asynchronous Learning Networks*, *13*(3), 115 127.

Xing, W., Guo, R., Petakovic, E., & Goggins, S. (2015). Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. *Computers in Human Behavior*, *47*, 168–181. https://doi.org/10.1016/j.chb.2014.09.034

Yang, S. J. H., Lu, O. H. T., Huang, A. Y. Q., Huang, J. C. H., Ogata, H., & Lin, A. J. Q. (2018). Predicting Students' Academic Performance Using Multiple Linear Regression and Principal Component Analysis. *Journal of Information Processing*, *26*, 170–176. https://doi.org/10.2197/ipsjjip.26.170

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zacharis, N. Z. (2015). A multivariate approach to predicting student outcomes in web enabled blended learning courses. *The Internet and Higher Education*, *27*, 44–53. https://doi.org/10.1016/j.iheduc.2015.05.002

Zoabi, Y., Deri-Rozov, S., & Shomron, N. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *Npj Digital Medicine*, *4*(1), Article 1. https://doi.org/10.1038/s41746-020-00372-6

Żytkow, J. M., & Sanjeev, A. P. (1998). Business Process Understanding: Mining Many Datasets. In L. Polkowski & A. Skowron (Eds.), *Rough Sets and Current Trends in Computing* (pp. 239–246). Springer. https://doi.org/10.1007/3-540-69115-4_33