**AKENTEN APPIAH-MENKA**

**UNIVERSITY OF SKILLS TRAINING AND ENTREPRENEURIAL DEVELOPMENT**

**FACULTY OF APPLIED SCIENCES AND MATHEMATICS EDUCATION**

**DEPARTMENT OF INFORMATION TECHNOLOGY EDUCATION**

**EMPLOYING MACHINE LEARNING ALGORITHMS TO PREDICT STUDENTS’ PERFORMANCE**

**BY**

**FEBRUARY, 2024**

# ABSTRACT

The ultimate goal of any educational institution is to offer the best educational experience and knowledge to the students. This includes identifying the students who need extra support and taking the appropriate actions to enhance their performance. This goal could be better achieved if parents and educational administrators were able to predict the performance of their students so that remedial interventions could be applied early to underperforming students and also make informed choices in critical decisions such schools and programme selection for their wards as they progress to higher levels of their education. One application of machine learning is the prediction of students’ examination performance which is aimed, among other things, at providing an estimate of future student achievements on specific examinations or assessments. The model built here is an approach to predict the likelihood of a student passing their external exams based on their internal exams records and also obtaining the best algorithm after the training and testing is done. It uses five (5) supervised learning algorithms of machine learning to train and predict the performance of students based on their internal assessment scores and deduce the one with the highest performance accuracy in predicting the students’ performance.

***Keywords:*** Machine Learning, Artificial Intelligence, Performance Prediction, Supervised Learning

### **CHAPTER ONE**

### **INTRODUCTION**

This project entails the exploration of machine learning algorithms which will be used to model a prediction of Junior High School (JHS) students’ performance in their final Basic Education Certificate Examination using their performance in their internal assessment for the 3 years of study at the JHS. This chapter also introduces the background to the area of study, the problem statement, research or project objectives, the scope of our project, the significance of study and the questions to be addressed.

#### Background to the Project

Good performance in examinations is a critical component for the progress of students at every level of the educational ladder. Students at the Junior High Schools in Ghana are therefore assessed at the end of their study in a final summative examination known as the Basic Education Certificate Examination (BECE) by the West African Examinations Council (WEAC). Their performance in this examination is used to determine their placement into the Senior High Schools (SHS) by the Computerized Schools Selection and Placement System (CSSPS) implemented by the Ghana Education Service. The possibility of a student being placed at the SHS of their choice will depend on their performance at the BECE. This makes the examination of great concern to the pupils, school administrators and parents.

The dilemma, however, is that parents and school administrators are not able to predict the performance of their students in the BECE exams to guide them in the selection of appropriate SHS. This leads to many students and parents being frustrated because the students do not get

schools of their choice and in many instances do not get placed in any school at all since they selected senior high schools having higher grade requirements than the grades the students obtained.

#### The use of Machine Learning in predicting students’ performance

Machine Learning (ML) is “a type of artificial intelligence in which computers use huge amounts of data to learn how to do tasks rather than being programmed to do them” (Oxford Dictionary, 2020). Machine learning is a subset of Artificial Intelligence that provides systems the ability to automatically learn and improve from experience without actually being programmed. It focuses on developing computer programs that can access data and use it to learn for themselves*.* Machine learning is one of the prominent fields in artificial intelligence that came from the improvement of self-learning algorithms to get knowledge from data sets provided to predict future trends.

Machine learning algorithms, according to Brempong (2021) are classified mainly into 3 categories:

* + - 1. Supervised machine learning algorithm
      2. Unsupervised machine learning algorithm
      3. Semi-supervised machine learning algorithm.

With supervised learning, which was implemented in this study, we use labelled data (classified dataset) to infer a learning algorithm. This data is used as basis for predicting the classification of other unlabelled data through the use of machine learning algorithm.

Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision. Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning,

we have the input data but no corresponding output data. The goal of unsupervised learning is to find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format. Semi-Supervised learning is a type of Machine Learning algorithm that represents the intermediate ground between Supervised and Unsupervised learning algorithms. It uses the combination of labeled and unlabeled datasets during the training period. (Javatpoint, n.d.)

Machine learning is playing a huge role in every field recently, and it can be said to include research about various events depending on the various event related data. So in order to apply machine learning, we need to assemble all required data. The data then needs to be pre-processed so that it could be carried forward for further operations or functioning. Prediction of future occurrences based on a given set of data is one major application of machine learning. In this study, the data of students from their internal examinations and assessment was used to predict their performance in their final external BECE examination.

#### Students’ Performance Prediction Algorithms

We used Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees (DT) and Random Forest (RF) as machine learning algorithms and from that we will determine which algorithm has the best results in percentage accuracy for predicting students’ examination performance. These five algorithms are all types of supervised learning in machine learning. In supervised learning, you train machine using well labelled data. This means that some data have been labelled with the correct answer.

Machine learning algorithms have important characteristics that need careful consideration. The first is that they are only good at training, with a target task learning system based on past data. If

the data is sparse or doesn't cover task types, their learning is not enough and therefore they cannot perform well when put on the job. The second aspect is the capabilities of each machine learning algorithm, which can be different for different types of tasks. This can also be called a "generalization problem", which tells how well the trained system can predict the cases for which it was not trained, i.e., whether it can predict beyond extent of the training dataset or not.

In this study, data from a graduated year group of Junior High Schools students would be used. The data will cover the 3-year internal assessments records of the students mapped against their final grades in their external BECE examination with the objective of finding relationships in the data to predict future students’ performance.

#### Problem Statement

Parents and school administrators each year go through the difficult decision of selecting Senior High Schools for their pupils for onward processing in the Computerized School Selection & Placement System. Since they are unable to predict the performance of their students in the external exams, they sometimes end up selecting schools with higher grade requirements than the grades their students obtained leading to their inability to be placed in those schools and sometimes, not placed in any school at all. This brings untold hardship and stress on many parents and students as they wander about in search of schools because they were not automatically placed in any SHS based on the student’s performance. If school administrators and parents had a system to predict their wards performance in the external examination, it would guide them to select the right schools which will guarantee placement in SHS for their wards.

#### Significance of the Project

A students’ performance prediction model or algorithm can play an important role in providing relevant information to parents and school administrators in taking key decisions relating to the academic progress of students. Information derived from the system will guide them in:

* Providing relevant academic interventions for low performing students.
* Selecting appropriate SHS and programme of study for the students based on their predicted performance.
* Providing evidence-based data for students counselling.

#### Research Objectives

The main goal of this project is to research and develop machine learning algorithms that can be used as a model to predict the performance of JHS students in their external BECE examination. Several types of machine learning algorithms are studied, including their architecture and variations of the related learning rules to determine a higher percentage of accuracy on one of these algorithms, which will provide the best prediction for students’ performance based on the given data.

The sub-goals of this project include;

* Infer the most effective machine learning algorithm for the prediction of students’ examination performance.
* To show that a variable (internal assessment records) can be used to more effectively predict students’ performance in external examinations.
* To be able to represent this variable in order to derive accurate predictive patterns for the prediction of students’ performance in examinations.
* To compare the performance of the ML algorithms on different ratios of training and test data (that is 70:30 and 90:10 ratios of training and test data respectively).

#### Research Questions

The research questions that will be addressed in this project are:

* What is the most effective machine learning algorithm for students’ examination performance prediction?
* Which variable (internal assessment records) would most effectively predict student’s external examination performance?
* How should this variable be represented to give accurate prediction models for students’ examination performance?
* What ratio of training and test data will perform better in the prediction of students’ performance using the selected ML algorithms?

#### Scope of the Project

The project mainly focuses on using supervised machine learning algorithms trained on internal examination values to predict the performance of JHS students in their external examination (BECE). This project was born with the need to help alleviate the stress that parents and educational administrators go through in taking decisions based on the future performance of their students in external examinations. This system was therefore developed for the sole purpose of predicting the performance of students in their external examinations at the basic level.

#### Limitations of the Project

The data set used for the research was from a basic Junior High Schools in the Oforikrom Metropolis of the Ashanti region and the grading system of the BECE conducted by WAEC. Hence, the most accurate machine learning algorithm chosen after the learning processes may not accurately predict the correct results for other levels of the educational ladder (secondary, tertiary) or a different educational system or examination body.

#### Definition of Terms

* Machine Learning - It is the science of getting computers to act without being explicitly programmed
* Algorithms - A series of instructions telling a computer how to transform a set of facts about the world into useful information.
* Predictive - Relating to or having the effect of predicting an event or result.
* Accuracy - The quality or state of being correct or precise.
* Learning process - The completion of the learning cycle that includes active testing, concrete experiences, reflective observation and abstract hypothesis.

**1.9 Literature review**

This chapter surveys research findings in literature relating to the prediction of students’ performance using machine learning. It reviews the research works carried out by different researchers who used various machine learning algorithms to predict students’ performance in education.

**1.10 Methodology**

This chapter looks at the performance forecasting methods we will use in our project. Supervised machine learning algorithms applied to students’ performance forecasting will be reviewed in this chapter. The collected datasets would be pre-processed into a format suitable for training, validation and testing. Five supervised machine learning models will be selected for training and validation. The trained models will then be tested to evaluate their performance and accuracy in the forecasting of students’ performance.

1. **References**

Ahajjam, T., Haidar A., Farhaoui Y. (2021). Artificial intelligence and machine learning to predict student performance during the COVID-19. *Procedia Computer Science, Volume 184*, ISSN 1877-0509. <https://doi.org/10.1016/j.procs.2021.03.104>.

Altabrawee, H., & Ali, O., & Qaisar, S. (2019). Predicting students’ performance using machine learning techniques. *Journal Of University of Babylon for Pure and Applied Sciences*. 27. 194-205. 10.29196/jubpas.v27i1.2108.

Brempong Jnr, O. W (2021). Application of machine learning (linear regression model) to predict students enrollment among senior high schools in Ghana. *International Journal of Scientific & Technology Research Volume 10, issue 10, October 2021.* ISSN 2277-8616

Debeljak, M., & Džeroski, S. (2011). Decision trees in ecological modelling. In Modelling complex ecological dynamics (pp. 197-209). Springer, Berlin, Heidelberg.

Golino, H. , Gomes, C. & Andrade, D. (2014). Predicting academic achievement of high-school students using machine learning. *Psychology, 5,* 2046-2057.

doi: [10.4236/psych.2014.518207](http://dx.doi.org/10.4236/psych.2014.518207).

Hussain, S. & Khan, M.Q. (2021). Student-Performulator: Predicting students’ academic performance at secondary and intermediate level using machine learning.

*Annals of Data Science.* <https://doi.org/10.1007/s40745-021-00341-0>

Kheir, R. B., Greve, M. H., Abdallah, C., & Dalgaard, T. (2010). Spatial soil zinc content distribution from terrain parameters: A GIS-based decision-tree model in Lebanon. *Environmental Pollution, 158*(2), 520-528.

Kotsiantis, S., Pierrakeas, C. & Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied artificial intelligence*, 18:5, 411- 426, DOI: 10.1080/08839510490442058

Kumar, A. D., & Radhika, V. (2014). A survey on predicting student performance. *International Journal of Computer Science and Information Technologies*, *5*(5), 6147- 6149.

*ons*, *11*(3).