Determining Student Demographic Attributes Influencing Performance Using Binary Classification in KDP Model

Iddrisu Issah (issah.iddrisu.stu@uenr.edu.gh)  
University of Energy and Natural Resources  https://orcid.org/0000-0002-1407-5746

Peter Appiahene  
University of Energy and Natural Resources  https://orcid.org/0000-0002-6098-4537

Obed Appiah  
University of Energy and Natural Resources  https://orcid.org/0000-0002-4680-3015

Fuseini Inusah  
University for Development Studies  https://orcid.org/0000-0001-9785-4464

Research Article

Keywords: Academic, Intervention, Performance, Demographic, Prediction

Posted Date: June 1st, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3005805/v1

License: ©  This work is licensed under a Creative Commons Attribution 4.0 International License.  Read Full License
Abstract

Machine learning (ML) is one way that can help decipher the intricate relationship between students’ data and their performance. When implemented correctly in learning environments, machine learning will improve knowledge of fundamental processes by simplifying the identification, extraction, and evaluation of underlying factors that affect student learning and levels of achievement. This study employed the experimental research approach using binary classification techniques based on the six-step Knowledge Discovery Process (KDP) model. Five classifiers were used within the Rapid Miner’s 9.10.010 educational environment as both experimental and analytical tool. The dataset comprised of 2334 records, 17 attributes with one class variable (students’ semester average score) inclusive. Twenty different tests were conducted. The experiments’ results were evaluated using 10-fold cross-validation and ratio split validation with bootstrap sampling. The Random Forest algorithm (RF), Rule Induction methods (RI), Naive Bayes (NB), Logistic Regression (LR) and Deep Learning (DL) algorithms were used in the experiment. The experimental results demonstrated that the RF method outperforms the other four techniques in all six-evaluation metrics that were employed for the selection process with the accuracy being 93.96%. According to the RF classifier model, the mother’s and father’s education levels of students are two recognized demographic factors per this study that significantly influence pre-tertiary students’ academic achievement. This study has significantly reduced the gap in practical knowledge observed in the literature by introducing an intervention scheme for respective student’s requiring intensive or minimal academic interventions in its prediction procedure.

1.0 INTRODUCTION

Student assessment takes centre stage in educational activities when it comes to the evaluation of the educational attainment of students in pre-tertiary institutions in the country. Evaluating students’ performance however of late has become a daunting task as more factors are now involved when it comes to the determinants of student achievements due to the paradigm shift now taking place in the educational sector; the use of Learning Management Systems (LMS), Students Information Systems (SIS) as well as Educational Management Information Systems (EMIS).

The data produced by these systems tend to overwhelm educational decision-makers due to the diversity and the massive volume of data housed by these data sources. However, recent research improvements have made powerful computational prediction methods and techniques, such as machine learning, a realistic alternative for various applications, including Educational Decision Support Systems (EDSS).

Machine learning is one way that can help decipher the intricate relationship between these students’ data and their performance. When implemented correctly in learning environments, machine learning will improve our knowledge of fundamental processes by simplifying the identification, extraction, and evaluation of underlying factors that affect student learning and levels of achievement.

Much progress has been made in the area of machine learning about its use in other fields such as medicine, commerce, the transport industry, bioinformatics, road traffic detection and control and in diverse fields where decision-making is crucial (Hasan, Palaniappan, Mahmood, Sarker, & Abbas, 2020).

Machine learning involves searching through many possible hypotheses to ascertain the most appropriate and relevant data and then comparing it with existing data generated by the learner. The idea of machine learning is
derived from various disciplines, such as probability and statistics, computational complexity, information theory, neurology, evolutionary theories, and models (Yakubu & Abubakar, 2022).

The machine learning design approach leans itself against several criteria that embody identifying the natural experiences acquired from training, the exact function to learn, a demonstration for the said function, and the optimal algorithm for learning the said function according to the training examples. Machine learning algorithms commonly used include; Decision trees (DT), Support Vector Machines (SVM), Artificial Neural Networks (ANN), Logistic Regression (LR), Naïve Bayes (NB), and rule inductions (RI) algorithms.

Similar to the other fields where machine learning has been successfully employed, its application on educational data is a promising area in research identified as Educational Data Mining (EDM). It involves coming out with processes to extract patterns embedded in datasets within educational settings (Arashpour et al., 2023). This concept has been implemented to improve and assess educational activities and decision-making.

On the part of (Ahmed, Abdulazeez, Zeебaree, & Ahmed, 2021), Prediction, which encompasses the subcategories of classification, regression, and density estimation, is a paradigm in educational data mining. Per the view of Barbosa et al. (2022) relation mining, association mining, correlation mining, sequential pattern mining, and causative data mining, are all types of clustering. In addition, prediction also incorporates data distillation to aid in human logic and model finding. From the views of (Inusah, Missah, Ussiph, & Twum, 2021) and (Inusah, Missah, Najim, & Twum, 2022) EDM has proven to be the primary source of solid and dependable data analysis when it comes to educational decision-making at the country's educational institutions. It carefully identifies the challenges of education to determine appropriate solutions which address the challenges. The use of Expert System with DM to manage basic education in the country as a result of challenges can be seen in (Inusah, Missah, Najim, & Twum, 2023b) and (Inusah et al., 2021). Educational Data Mining has been used to track the academic welfare of students and the general administrative procedures of educational institutions worldwide (Inusah, Missah, Najim, & Twum, 2023a) and (Drachsler & Greller, 2016).

The paper aims at identifying and applying machine learning (algorithm) to uncover the key demographic factors that influence newly admitted students' academic achievement as well as identify students to receive appropriate academic intervention so that overall school performance can be scaled up in the West Africa Senior Secondary Certificate Examination (WASSCE).

The research aims to examine and address the following set of questions:

1. Which machine learning classification algorithms are more viable in predicting students' academic attainment based on their demographic attributes?

2. What primary demographic attributes influence students' academic performance at the Senior High School (SHS) level in Ghana?

The rest of the paper is arranged as follows: The second section is a review of relevant literature. Section three discusses the study's materials and methodologies. Section four then presents the study's findings and comments. Section five finally discusses the conclusion and recommendations for possible future studies.

2.0 RELATED LITERATURE

This section discusses related studies on student characteristics and their influence on academic performance.
In recent decades, machine learning has become a famous avenue for data analysis and predictions. Due to the advent of computerization of institutional data management, many studies investigating building new machine learning models and determining their efficacies for a wide range of fields have been done. In addition, machine learning has been proposed to advance learner experience and determine the factors significantly influencing the capacity to predict students' performance. The issue at hand in the research was to ascertain the efficiency of machine learning algorithms in deciding if a student needs a high or low intervention using some features referred to here as "predictors", mostly students' characteristics.

2.1 The Role of Predictive Models in Educational Data Mining

Predictive modelling aims to determine class of a given attribute by using the values of other given attributes. The attributes being predicted are referred to as the target or dependent variables, whereas the independent/explanatory variables are the attributes used in the prediction process (Hasan et al., 2020). The resultant model is typically based on carefully examining the training datasets. Classification and prediction techniques are two techniques that may be employed during the development of a model to determine future trends and patterns to gain a thorough knowledge of a given data set (Hasan et al., 2020). The classification techniques commonly used predictive models include K-Nearest Neighbor, Decision Trees, Neural networks, support vector machines, and Rule-Based Classifiers (Rule Induction Naïve Bayes).

2.2 Predictive Modelling Using Classification Algorithms

Prediction models or functions distinguishes the data clusters and concepts by applying the model to predict the class of variables with the unknown class (Hasan et al., 2020). The resultant model is typically based on the careful examination of the pattern of the training dataset. Derived models are presented in several forms, including the classification: neural networks, (IF-THEN) rules, mathematical formulae, or decision trees (Hashim, Awadh, & Hamoud, 2020).

According to (Hashim et al., 2020) Naïve Bayes, decision trees and rule inductions are among the data mining techniques to classification problems. The following classification methods are briefly discussed: decision trees, support vectors, Bayesian classification, logistic regression, neural networks, and rule induction.

2.2.1 Decision Trees

A classification method is used to construct a decision tree. The classification processes are described in this instance via a hierarchical array of decisions on feature variables that manifest in the shape of a tree (Anuradha & Velmurugan, 2016). Decision trees are made of nodes that are joined to constitute a rooted tree; it is, therefore, a directed graph comprising nodes known as roots without incoming edges (figure 2.1). The other nodes that determine the class of objects are known as the leaves or terminal nodes (Bhatia, 2019). Every leaf is attributed to a class that represents the most appropriate target value (Samson, 2019). Nodes with a blend of diverse classes are to be split further. A stopping criterion determines when the decision tree algorithm should terminate. When all training examples in the terminal/leaf node fit within the same class, then the stopping criterion is said to be reached (Junshuai, 2019).
Every node matches a characteristic, while the branches link with an array of values. All nodes are labelled with the attributes they test, and every branch is labelled with its corresponding values (Liu et al., 2023). The range of values is mutually exclusive and complete. The properties of a tree being disjoint or complete are vital as they ensure every instance maps to one case (see Figure 2.1) as described in (David Kolo et al., 2015).

**Random Forest Algorithm**

Averaging ensemble approaches include the Random Forest algorithm. Random forests represent huge feature areas and are more resilient than decision trees. *Random Forests* is a bagged classifier that connects a group of decision tree classifiers to form a forest of trees (Eddin, Khodeir, & Elnemr, 2018). A diverse collection of classifiers is formed by integrating randomization into the classifier-building process. The ensemble prediction is presented as the average prediction of the discrete classifiers (Yakubu & Abubakar, 2022). In random forests, every tree in the ensemble is created using a unique bootstrap sample, which includes a random selection of instances with replacements from the entire training dataset (Sokkhey & Okazaki, 2020).

According to (Yakubu, 2021), Random feature selection is used in a random forest, where 'm' features are chosen randomly from 'M' features for every node of the decision tree t, and the optimal value is taken from m. Therefore, the split determined when splitting a node throughout tree formation is no longer the best split among all features.

Alternatively, the chosen split is the best among a randomly picked collection of characteristics. As a result, the forest bias often grows concerning the bias of a single non-random tree (Denny, Leslie, Spits, & Budiharto, 2021). However, averages generally compensate for an overall model's increase in bias.

### 2.2.2 Rule-Based Classifiers

A typical rule is described as follows: **IF** a condition exists, **THEN** the result (Bhatia, 2019). The antecedent condition is on the rule's left side and consists of a variety of logical operators, comprising of >, <, =, & and OR, that mainly are employed on feature variables. The consequent that generates the class variable is located on the rule's right side. Ri is a rule presented as Qi→c, Qi being the antecedent and C as the class variable. The symbol → epitomizes the condition 'THEN'. Induction rules are mostly constructed by using the datasets gathered during the training phase. The symbol Qi denotes a condition applied to the feature set (Agrawal, K., & K., 2017). A rule is expressed in the format: IF (attribute 1; value 1) and (attribute 2; value 2) and …… (Attribute n; value n) THEN (decision; value).

**Rule induction Algorithm**

Rule induction is experimented in the study and is widely applied rule-based classification technique. As stated, (Samson, 2019), rules are good when denoting information and aspect of information. Rule induction generates rules by dividing and conquering the training set, bringing out all instances bound by the rule. Rule induction uses the divide-and-conquer and separate-and-conquer rule learning approaches. The rule algorithms generate a decision list, an ordered set of rules. Through J48, rule induction discovers rules based on partial decision trees, develops a partial C4.5 decision tree, and translates the "best" leaf into a rule (Prof, 2018).
Typically, an if-then rule has the form: IF mother education = primary AND mother occupation = Government AND JHS location = Urban THEN Status = Low Intervention.

2.2.3 Bayesian Classification

Observations made by (Prof, 2018) indicates that, the Bayesian classifiers, also called the Naïve Bayes, are based on statistical classifiers which are built on the Bayes theorem. The accuracy and speed of the Bayesian classifiers have been proved to be of high magnitude on large databases (Ouatik, Erritali, Ouatik, & Jourhmane, 2022). The Bayesian classification offers a pictorial view of underlying associations on which to perform learning. A Trained Bayesian holds the principle that networks can be useful in classification (Ofori, Maina, & Gitonga, 2020), A Bayesian classification graphical model is indicated in figure 2.2.

Let X denote a data tuple, labelled as the measurements on a set of n attributes. Let H denote the hypothesis. Then, p (H|X) denotes the probability that the hypothesis H holds per the data tuple X. P (H|X) denotes the posterior or posteriori probability H, conditioned on X. On the other hand, P (H) denotes the prior probability of hypothesis H. Correspondingly, P(X|H) denotes the posterior probability of X conditioned on H while P(X) is the prior probability of X. Bayes theorem offers a criterion for calculating the posterior probability, P (H|X) from P(H), P (X|H), and P(X). the Bayes' Theorem is put as follows:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$  \[\text{Eqn. 2.1}\]

For problems in classifications, X will represent an observed data tuple assuming H as hypothesis binding on X with class C. these are used to establish the probability of P(H|X) that binds on tuple X in class C, according to the attribute depiction of X (Agrawal et al., 2017)

Naïve Bayes Algorithm

Naive Bayesian classifier is based on the Bayes rule for its probability computations (Balamurugan, 2018). Each target attribute value is assigned an instance, and all input attributes are conditionally independent of assigned target attribute (i.e., class conditional independence) (Balamurugan, 2018). Naïve Bayes dwells much on the simplistic but unrealistic hypothesis.

(Agrawal et al., 2017) suggested a stepwise implementation of the naïve Bayesian classifier as follows:

1. D is the training set of tuples and their related class labels. Each tuple is denoted as an n-dimensional attribute vector, X = (x1, x2, ......., xn), representing n measurements on the tuple based on n attributes, A1, A2, ...., An respectively.

2. Supposing m classes are involved, C1, C2, ...., Cm, and Given tuple X, the classifier can predict X to belong to a class with the highest posterior probability, based on X.

Here, the naïve Bayesian classifier will predict the tuple X as belonging to the class Ci if and only if P (Ci/X) > P(Cj/X) for 1 ≤ j ≤ m; j ≠ i. P (Cij X) is maximized here. The class Ci on which P (Cij X) is maximized is known as
the maximum posterior hypothesis.

\[
P(C_i/X) = \frac{P(X_j/C_i)P(C_i)}{P(X)} \quad \text{[Eqn. 2.2]}
\]

\(P(X)\) is recurrent for all classes, but only \(P(X/C_i)\) and \(P(C_i)\) need maximization. If prior class probabilities are missing, then it is assumed that \(P(C1) = P(C2) = \ldots = P(Cm)\), and therefore \(P(X/C_i)\) is maximized, else \(P(X/C_i) P(C_i)\) is maximized.

The class prior probabilities may be projected using \(P(C_i) = \frac{|C_i, D|}{|D|}\), where \(|C_i, D|\) denotes the training tuples of class \(C_i\) in \(D\).

3. With a data set of several attributes, it will be computationally difficult to calculate \(P(X/C_i)\). To minimize the calculation in evaluating \(P(X/C_i)\), the naive hypothesis of conditional class autonomy is made. This is on the assumption that the values of all attributes are conditionally discrete based on the class label of tuples (i.e., all attributes are discrete). Thus

\[
P(X|X_i) = \prod_{k=0}^{n} P(X_k | C_i) \quad \text{[Eqn. 2.3]}
\]

The probabilities can be estimated using the training tuples. \(x_k\) denotes attribute values of \(A_k\) for tuple \(X\) for every attribute. We determine whether its value is categorical or continuous-valued.

4. To determine the class label of \(X\), \(P(X|C_i) P(C_i)\) is valued for every class \(C_i\). The classifier then predicts the class label of tuple \(X\) in class \(C_i\) if and only if \(1 \leq j \leq m, j \neq i\).

The class \(C_i\) with the highest \(P(X/C_i) P(C_i)\) value is the one with the most accurate label prediction.

### 2.2.4 Support Vector Machines

Support vector machine, or SVM, is a learning algorithm used to study and understand classification and regression rules. Support Vector Machines (SVMs), for example, can be used to train radial base functions (RBFs), polynomials, and multi-layer perceptron (MLP) classifiers (Ouatik et al., 2022).

The SVMs are derived from the statistical learning concept, objective is to solve related problems, excepting more complex ones, as a transitional step (Sameer & Barahate, 2016)

The SVM belongs to the supervised learning algorithm family capable of generating learning rules based on the given training dataset. The SVM has a comprehensive theoretical basis and entails comparatively a lesser number of data samples to be used for the training; investigations indicate that SVM is not sensitive to sample dimensions (Hashim et al., 2020)

### 2.2.5 Neural Networks
The artificial neural network (ANN), also referred to as the neural network (NN), simulates the neural network system of humans. It is made up of an interrelated cluster of artificial neurons that process information by using a connectionist technique for computation (Arashpour et al., 2023).

The Neural network framework is made up of interconnected nodes through a directional link. Every node presents itself as a processing unit and each link depicts a causal association among the nodes. The nodes are adaptive (the outputs of the nodes are based on the modifiability of the parameters concerning these nodes) (Hashim et al., 2020).

Every node in the input layer of an ANN matches a predictor when the ANN is first constructed. After that, the input nodes are connected to various other nodes contained within the hidden layer. Every single input node is connected to other hidden layer nodes within the network. The hidden layer nodes may be linked to other corresponding hidden layers or directly to an output layer. One or several response variables constitute the output layer (Bhatia, 2019).

Next to the input layer, the other nodes take in inputs, multiply the inputs by a connection weight \( W_{xy} \) (nodes 1 to 3 are put as \( W_{13} \)), sum them, and then apply a function (known as activation or squashing function) to them, then transfer the output to the next layer. For instance, values passed from nodes 4 to 6 are put as activation function to \([W_{14} \times \text{value of node 1}] + [W_{24} \times \text{value of node 2}]\)

**Feed Forward deep Networks/Multilayer Perceptron**

The most basic deep networks are feedforward deep networks, commonly known as multilayer perceptron (MLPs) (Hashim et al., 2020). The Multilayer perceptron is the most implemented NN architecture in predictive data mining. The multilayer perceptron is based on the feed-forward deep network with many possibly concealed layers, with the input and output layers connected, respectively (Hashim et al., 2020). The feed-forward neural network has no interconnections between nodes within a given layer; instead, outputs from one layer are used as input information to nodes in subsequent layers. This ensures modularity within the network, i.e., nodes are coherent in functionality or provide an equivalent level of abstraction on input vectors (Samson, 2019).

**2.2.6 Regression**

Regression is commonly employed in predictive model building as well as the analytical processes in data mining. Regression predictions are primarily centred on historical data using functions and formulas (Palacios, Reyes-Suárez, Bearzotti, Leiva, & Marchant, 2021). It is mainly a statistical approach to data mining. Regression is implemented to derive a model between dependent variables and independent variables (Palacios et al., 2021). Regression is also used to build a model to analyse existing datasets to forecast trends by using linear or logistic regression techniques derived from statistical methods where functions are driven from an existing dataset. The derived data is subsequently mapped to the functions to assist in predicting (Larose & Larose, 2015).

**Logistic Regression Algorithm**

The Logistic regression (LR) algorithm is applied to build a regression model using categorical dependent variables. Logistic regression is put into three categories (1) binary, in the case of binary response variables (2) multinomial – for above two non-ordered dependent variables (3) ordinal for an ordered category (Samson,
2019). Logistic Regression is generally used by researchers and data analysts in analyzing and the classification of proportional and binary response data (Journal et al., 2016). The LR can effortlessly handle probability and multi-class issues in classification.

2.3 DETERMINANT FACTORS OF STUDENTS' ACADEMIC PERFORMANCE

It is essential to be aware of the factors (also known as the predictor variables) that influence students' academic performance to comprehend and enhance the current state of the educational system (Owusu-Boadu, Nti, Nyarko-Boateng, Aning, & Boafo, 2021). Therefore, the determination of the characteristics associated with students' academic accomplishment has always aroused the interest of academics who work in the field of educational data mining. A significant number of earlier studies dissected this phenomenon by isolating one variable at a time. They attempted to investigate the relationship between a single element and its impact on academic accomplishment by collecting data, the majority of which was obtained using instruments of the survey type. Previous research works have been published in the academic world to determine the primary elements or characteristics that contribute to influencing the performance of students as well as the algorithms that produce the best prediction result.

According to (Tadese, Yeshaneh, & Mulu, 2022), students' apparent poor performance in numerous educational establishments has been influenced by a variety of predictors. They include personal characteristics, intellectual ability, gender and aptitude tests, academic achievement, previous college accomplishments, and demographic characteristics (Ouatik et al., 2022).

2.3.1 Grades and Assessment Reports as Determinants of Academic Performance

Recently, there has been a rise in interest in studies on the correlations between the teaching techniques used by teachers and educators and how they affect learners' academic achievements. Most of the studies dwell on improvement as a result of the implementation of assessment methods in the form of mid-semester tests, assignments, class exercises, practical assessments, and end-of-semester examinations (Sekeroglu, Dimililer, & Tuncal, 2019). Past grades in an academic institution are anticipated to portray an ideal weight when predicting the future academic achievements of the learner (Khudheir, n.d.), especially in a situation where grades are derived from the continuous assessment that portrays the early understanding of a subject.

According to (Yakubu & Abubakar, 2021), grades are common indicators of academic outcomes in any educational institution.

(Aman, Rauf, Ali, Iqbal, & Khattak, 2019), explored the efficacy of assessments using the techniques of examination, class test, assignment, and mid-semester quizzes, including the influence of lecturer response on students' performance. The study's outcome revealed that a correlation exists between the assessments students took and eventually the students' final grades.

Another investigation by (Aman et al., 2019) in exploring the relevance of formative assessment to improve the prediction of learner grades in examinations suggested the possibility of identifying students who may perform
poorly in their final examinations. (López-Zambrano, Torralbo, & Romero, 2021) also suggest the possibility of being able to forecast, with a degree of accuracy, how a student will perform at an end of course examination. (Adekitan & Noma-Osaghae, 2019) concluded that the effects of giving assessment feedback on time to students often result in a little enhancement in the final grades.

(Altujjar, Altamimi, Al-turaiki, & Al-razgan, 2016) did a case study on predicting the validity of previous achievements in determining students' performance in higher education. The high school Scholastic Assessment Test (SAT) score marks, as well as the early years' university grades, were considered possible predictors of future performance. The impact of subjects on students' advanced placements was also investigated. Their finding indicated a clear connection between these three characteristics and students' university accomplishments.

### 2.3.2 Student Demographic Features as Determinants of Academic Performance

The (Issah, Appiah, Appiahene, & Inusah, 2023) in a literature identified various factors affecting inequity and diversified performance among school students. Among the background factors include school effects, socioeconomic background, and personal traits hindering students' performance.

Tinto (1975), as cited in (Altujjar et al., 2016), stated that student background characteristics such as education levels, the profession of parents/guardians, and place of residence all play an essential part in defining students' success. This is further corroborated by (Ahadi, Lister, Haapala, & Vihavainen, 2015) referring to these phenomena on students' academic success as "a one-hundred-factor problem" as many researchers focused on different aspects of student's performance in different periods and came out with diverse conclusions.

In a study by (Ha, Giap, Loan, & Huong, 2020), "examining the impact of socio-economic influence on the upbringing of students and the final results of their education" realized that students from privileged backgrounds attained higher grades or had necessary skills that proved valuable within the academic setting. This suggests that the level of poverty and even the area students come from can affect a student's academic output. Furthermore, this suggests that the home environment of a student is a contributory factor to his performance.

In research conducted by (David & Anastasija, 2019) in Serbia, some demographic features, including gender, ethnicity, and the students' school background, were investigated to determine which among them had more influence on the students' academic performance in Mathematics and the Serbian language. The result indicated that student affluence had the highest contributory factor to poor mathematics performance whereas the Serbian language grades were less affected. Gender had a relatively minimal effect on the grades suggesting that gender had less effect on the performance of students at the university level. in their study, (Al-Twijri & Noaman, 2015) suggest that integrating demographic data alongside school results is recommended because learner achievement is based almost entirely on students' past exam results, mostly without due consideration for the setting in which some of these performances had been accomplished.

Again, research on student achievement and the associations with both context-specific background variables and attainment in broader terms (Issah et al., 2023) was largely seen to be limited (Tadese et al., 2022). Hence the need to delve into the correlation between the performance of students and their demographic variables.
More so, literature in this regard has failed to provide further remedies or intervention strategies based on identifiable traits early in a student’s programs of study. As a result, the goal of this research is to execute ML on dynamic data of students to track their performance, as well as design a classification model capable of mapping student features and performance in order to effectively implement the Ministry of Education’s (MOE) flagship early intervention scheme to improve underperforming students’ academic achievements in schools.

3.0 MATERIALS AND METHODS

This study employed the experimental research approach using binary classification techniques based on the six-step KDP model. The classification technique was used to sort the students into either in need of intensive intervention or low intervention.

3.1 Datasets

We employed secondary data from two sources. Based on the placement forms of students from the Computerized School Selection and Placement System (CSSPS), the demographic, Basic Education Certificate Examination (BECE) average score and previous school data were extracted, while the semester average score and the Grades for English Language, Mathematics and Integrated Science for their Senior High School (SHS) performance were also extracted from the Students Information System (SIS). Also, with the suggestion of the domain expert (ICT coordinator of Tamale Islamic Science Senior High School (TISSEC)), the following student attributes were considered helpful for the task at hand: mother’s education level, father’s education level, Sponsor for the student’s education, the birth position of the student in the family, and parental status of students. This study used 1854 records and 17 common attributes (including the class attribute) for training and evaluating the various models. The description of students features used in the study is summarized in Table 3.1.
Table 3.1
Attributes extracted from database

<table>
<thead>
<tr>
<th>No</th>
<th>Attribute Name</th>
<th>Data Value</th>
<th>Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gender</td>
<td>[1]: Male = M; [2] Female = F</td>
<td>Nominal</td>
</tr>
<tr>
<td>2</td>
<td>Student position in the family</td>
<td>[1]: 1st born, [2]: last born, [3]: others, [4]: only child</td>
<td>Numeric</td>
</tr>
<tr>
<td>3</td>
<td>Parents marital status</td>
<td>[1]: married, [2]: Single, [3]: widowed</td>
<td>Nominal</td>
</tr>
<tr>
<td>4</td>
<td>Fathers Edu.</td>
<td>[1]: Primary school, [2]: Junior High School (JHS) [3] Secondary school (SHS) [4]: Tertiary, [5]: Non</td>
<td>Nominal</td>
</tr>
<tr>
<td>5</td>
<td>Mothers’ Edu.</td>
<td>[1]: Primary school, [2]: Junior High School (JHS) [3]: Secondary school (SHS) [4]: Tertiary, [5]: Non</td>
<td>Nominal</td>
</tr>
<tr>
<td>8</td>
<td>Sponsor</td>
<td>[1]: Self, [2]: Parent, [3]: scholarship, [4]: others</td>
<td>Nominal</td>
</tr>
<tr>
<td>9</td>
<td>Residential Status</td>
<td>[1]: Boarding, [2]: Day</td>
<td>Nominal</td>
</tr>
<tr>
<td>10</td>
<td>Type of JHS attended</td>
<td>[1]: Private, [2]: Public</td>
<td>Nominal</td>
</tr>
<tr>
<td>12</td>
<td>BECE Accumulated raw score</td>
<td>[1]: 50–100, [2]: 101–200, [3]: 201–300, [4]: 301–400</td>
<td>Numeric</td>
</tr>
<tr>
<td>16</td>
<td>English Language</td>
<td>[1]: A1, [2]: B2, [3]: B3, [4]: C4, [5]: C5, [6]: C6, [7]: D7, [8]: E8, [9]: F9</td>
<td>Nominal</td>
</tr>
<tr>
<td>17</td>
<td>Mathematics</td>
<td>[1]: A1, [2]: B2, [3]: B3, [4]: C4, [5]: C5, [6]: C6, [7]: D7, [8]: E8, [9]: F9</td>
<td>Nominal</td>
</tr>
</tbody>
</table>

3.2 Dataset optimization and sampling technique
Primary and real-world data will invariably contain imbalanced data challenges (Fernández, García, Herrera, & Chawla, 2018). For example, whenever the number of instances from one class (the minority class) is significantly lower than the number of instances from the other classes (the majority class), the minority class may be the most effective, leading to the highest error cost in terms of learning (Jenssen, Krogstad, & Halvorsen, 2014). To address the issue of class imbalance across features, the Synthetic Minority Over Sampling Technique (SMOTE) with default settings was used as a sampling technique to upscale the minority classes (i.e. students demographic variables). The upscaling synthetically increased the number of demographic variables by 79% within local repository of Rapid Miner after the SMOTE Up-sampling application.

### 3.3 Feature Extraction and Feature Hierarchy

Since not all attributes have equal significance in prediction within a defined dataset, feature extraction and order are critical. Given this, the attributes were sorted on information gain by weight. The operator "Weight by Information Gain" was used in RapidMiner to determine the order of the attributes. Figure 3.1 depicts the descending order of information gained from common attributes to class attributes.

<table>
<thead>
<tr>
<th>No.</th>
<th>Attribute</th>
<th>Information Gain</th>
<th>No.</th>
<th>Attribute</th>
<th>Information Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mother Education</td>
<td>0.157</td>
<td>9</td>
<td>Region of Residence</td>
<td>0.016</td>
</tr>
<tr>
<td>2</td>
<td>Father Education</td>
<td>0.139</td>
<td>10</td>
<td>Age</td>
<td>0.013</td>
</tr>
<tr>
<td>3</td>
<td>BECE Raw Score</td>
<td>0.131</td>
<td>11</td>
<td>JHS Location</td>
<td>0.012</td>
</tr>
<tr>
<td>4</td>
<td>BECE Aggregate</td>
<td>0.123</td>
<td>12</td>
<td>Sponsor</td>
<td>0.003</td>
</tr>
<tr>
<td>5</td>
<td>Father Occupation</td>
<td>0.083</td>
<td>13</td>
<td>Student Birth Position</td>
<td>0.001</td>
</tr>
<tr>
<td>6</td>
<td>Mother Occupation</td>
<td>0.030</td>
<td>14</td>
<td>Parent Marital Status</td>
<td>0.000</td>
</tr>
<tr>
<td>7</td>
<td>JHS Type</td>
<td>0.026</td>
<td>15</td>
<td>Student Maturity</td>
<td>0.000</td>
</tr>
<tr>
<td>8</td>
<td>Residential Status</td>
<td>0.016</td>
<td>16</td>
<td>Gender</td>
<td>0.000</td>
</tr>
</tbody>
</table>

### 3.4 Model Building

Modelling comprises one of the significant roles played by hybrid data mining processes. Therefore, various modelling tasks in the data mining process were conducted at this phase of the study. The tasks involve selecting a modelling approach, setting up an experiment, developing a model, and assessing the model's efficacy.

### 3.5 Selecting a Modeling Technique

Experiments were conducted in this study to build models by incorporating specified classifiers for predicting the performance of pre-tertiary students based on demographic information. Five classification approaches were used for model construction to meet the study's aims. RapidMiner Studio was used to conduct the experiments and the analysis. The RF algorithms from decision trees, RI algorithms from rule-based classifiers, the NB algorithm from Bayesian Networks, the LR algorithm from Regression, and Deep Learning algorithms from NN were chosen for the experiments among the various classification algorithms available in RapidMiner. The grounds for selecting the algorithms are:
Their capacity to handle polynomial attributes effectively,
the ease of understanding and interpretation of the model's outcomes for the investigations.

3.7 Research design

This study is based on experimental research that employs binary classification techniques. The data used comprised of both numerical (e.g., age, test scores, etc.) and nominal (textual data), e.g., gender, residential status, and former school. The experimental study concepts are chosen because they are the basic approach to studying cause and effects (cause/effect) connections and studying the relations between two variables (Samson, 2019). Also, Experimental research is used by researchers to make comparisons between two or more groups on one or more metrics.

The research again employed a hybrid data mining model development approach based on the KDP model to carry out the study. This approach gives the researcher a deeper understanding of the problem when used than deploying only one approach. This design methodology was employed to obtain a much more broad-minded, research-oriented explanation of the phases; it symbolizes a data mining process rather than just a modelling step; and has numerous novel, clear, and specific feedback loops (Samson, 2019). Figure 3.2 indicates the six steps KDP modeling approach comprising of:

1. understanding the research problem,
2. understanding of data,
3. preparation of data,
4. mining the data,
5. analyzing knowledge base, and
6. using the knowledge that has been discovered.

3.8 Evaluation Metrics

Evaluation of model performance is an essential rating for models’ effectiveness, improving parameters during iterative learning process, and choosing an acceptable model from an assortment of models (Chen et al., 2022). To construct a robust model, the following six widely known performance metrics were used for comparing and selection of algorithms for evaluating the classification task: accuracy, precision, sensitivity, specificity, AUC and F-measure.

Accuracy

The most prevalent metric for measuring the feasibility of a model is its accuracy. A data mining classifier's correct accuracy is measured by how well its predictions match the actual true or false values.

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FN} \quad \text{[Eqn.3.1]}
\]

Precision

Precision for a class is equivalent to counts of true positives (i.e., the counts of instances rightly considered as positive) divided by the total count of instances considered as positive class (i.e., summation of true positives and false positives).
Precision = \frac{TP}{TP+FP} \quad [\text{Eqn.3.2}]

Recall

Here, recall can be explained as a ratio of the number of true positives to overall count of instances belonging to positive class (i.e., summation of true positives and false negatives). i.e., instances not considered as belonging to positive class, though they belong to it).

Recall = \frac{TP}{TP+FN} \quad [\text{Eqn.3.3}]

Similarly, negative class's precision and recall are defused. Use precision is determined by the proportion of instances categorized as negative that is negative. In contrast, the ratio of true negatives to the total number of instances of negative class will provide a recall for users.

F-Measure

The F-measure is a metric for evaluating the performance of classifiers using confusion matrices. F-Measure is the opposite correlation between accuracy and recall, defined as harmonic mean of precision and recall. It is important to determine if a model's accuracy and recall are pretty well balanced (Powers, 2020).

\[
\text{F-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad [\text{Eqn. 3.4}]
\]

3.9 Experimental Settings

The models were developed and simulated in the design view of the Rapid Miner’s modeling environment using a Fujitsu laptop computer with Windows 10 pro (version 21H2) 64-bit operating system, an X64-based processor (Intel(R) Core (TM) i7-4702MQ CPU @ 2.20GHz 2.20 GHz) and 8 Gigabytes of Random Access Memory (RAM).

The K-fold cross-validation and the split validation were employed to each experiment as metrics evaluation techniques. The default parameter relative ratios of 0.7 for training and 0.3 for testing were adopted in split validation. In 10-fold cross-validation, data is arbitrarily subdivided into ten mutually exclusive equal subgroups of one to ten. Training and testing are repeated ten times. The initial subgroup is reserved as a test set.

3.10 Experimentation with selected algorithms

The exploration method was used during the experimentation process to identify the most suitable algorithm.

Four different experiments were conducted for each of the five algorithms used in the study: (random forest, rule induction, naïve bayes, regression, and deep learning) as follows:

Experiment 1: Experimenting algorithm with split (ratio split) validation test mode.

Experiment 2: Experimenting algorithm by employing Bootstrap resampling with a split (ratio split) validation test mode.

Experiment 3: Experimenting algorithm with 10-Fold Cross validation test mode.

Experiment 4: Experimenting by employing Bootstrap resampling with 10-Fold Cross validation test mode.
4.0 RESULTS AND DISCUSSIONS

This section presents the results of the random forest model on the dataset to discover the student demographic variables influencing their performance.

4.1 Determination and Evaluation of the Best Classification Model for predicting students’ achievements

RQ1: Which machine learning classification algorithms are more viable in predicting students' academic attainment based on their demographic attributes?

One of the primary goals of this study is to identify a suitable machine learning classifier capable of predicting students’ academic success based on demographic characteristics. Five algorithms were explored to implement the classification modelling: Random Forest, Rule Induction, Naive Bayes, Logistic Regression, and Deep Learning. The results of the experiments are presented as Table 4.1.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Test mode</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>F-Measure</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>Random Forest (pruned) using Bootstrap resampling with 10-Fold Cross validation</td>
<td>93.96%</td>
<td>93.19%</td>
<td>94.97%</td>
<td>92.94%</td>
<td>94.04%</td>
<td>0.980</td>
</tr>
<tr>
<td>Rule Induction</td>
<td>Rule Induction with Ratio Split validation</td>
<td>83.00%</td>
<td>83.48%</td>
<td>82.29%</td>
<td>83.71%</td>
<td>82.88%</td>
<td>0.879</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>Naïve Bayes using Split validation</td>
<td>79.43%</td>
<td>78.77%</td>
<td>80.57%</td>
<td>78.29%</td>
<td>79.66%</td>
<td>0.879</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>Logistic Regression using split validation</td>
<td>81.57%</td>
<td>80.44%</td>
<td>83.43%</td>
<td>79.71%</td>
<td>81.91%</td>
<td>0.892</td>
</tr>
<tr>
<td>Deep Learning</td>
<td>Deep Learning using Bootstrap resampling with 10-Fold Cross Validation</td>
<td>84.45%</td>
<td>82.15%</td>
<td>88.49%</td>
<td>80.35%</td>
<td>85.11%</td>
<td>0.924</td>
</tr>
</tbody>
</table>

Comparing the six-performance metrics in Table 3, Random Forest (pruned) implementing Bootstrap resampling with 10-Fold Cross validation had the most outstanding performance metrics among the five classifiers for predicting students’ characteristics influencing their academic performance. The RF had an accuracy of 93.96%, precision of 93.19%, the sensitivity of 94.97%, specificity of 92.94%, an F-measure of 94.04%, and an AUC of 0.980. As a result, the Random Forest result with 10-fold cross-validation and bootstrap resampling was selected as the propose model for the study.

4.2 Analysis of Attributes of Importance in the Random Forest Classifier Model

RQ2: What primary demographic attributes influence students’ academic performance at the SHS level in Ghana?
The weights of the respective attributes by information gain were determined using the model simulator operator to determine the attributes that had a significant impact on the decision made by the random forest classifier. These weights were ordered in descending. The top two attributes in the list were then considered to be the most relevant in the model choice process. According to the Random Forest classifier model simulator, the mother's and father's education levels (with highest weights of 0.358 and 0.168 respectively) are the two discovered demographic factors that significantly support the classification model per this study. Figure 4.1 depicts the order in which the attributes in support of the prediction are arranged according to their weights.

The two most contributing demographic attributes based on the weight of contributions to the decision made by the model are the mother's and father's education levels. The BECE attributes happened to belong to academic features; hence they were excluded.

4.2.1 Evaluation of the Model

This section explains the evaluation technique for the model developed to evaluate the demographic factors impacting student performance in pre-tertiary institutions. The study included twenty specific tests with various classifiers. To construct a robust model, the following evaluation parameters were used: the confusion matrix, the number of trees in the forest, and comparing the ROC of a random forest with the ROCs of rule induction, Naïve Bayes, Logistic Regression, and Deep Learning classifiers.

Model Evaluation using Confusion Matrix

Table 4.2 displays the confusion matrix of the chosen model, which was created using the Random Forest algorithm and the Bootstrap resample approach with 10-fold cross-validation.

<table>
<thead>
<tr>
<th>CONFUSION MATRIX</th>
<th>PREDICTED CLASS</th>
<th>ACTUAL CLASS</th>
<th>A</th>
<th>B</th>
<th>Classified As</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREDICTED CLASS</td>
<td>POSITIVE</td>
<td>NEGATIVE</td>
<td>(TP) = 1108</td>
<td>(FN) = 65</td>
<td>A = Low Intervention</td>
</tr>
<tr>
<td>ACTUAL CLASS</td>
<td>POSITIVE</td>
<td>NEGATIVE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NEGATIVE</td>
<td>POSITIVE</td>
<td>91</td>
<td>1070</td>
<td>B = Intensive Intervention</td>
</tr>
<tr>
<td></td>
<td>POSITIVE</td>
<td>NEGATIVE</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Much may be learnt by meticulously scrutinizing the errors generated by any classification model. The errors show discrepancies among the model's predictions and the factual outcome in the actual business situation. When an appropriate model is discovered, the next step is to figure out why classification inaccuracies happened in the testing data. When predicting an attribute for a certain class label, for instance, the predicted and actual results may differ.

However, because comparable features reside within the same class limit, the classifier predicts the data into a particular class.
Table 4 displays the confusion matrix of the final model for the study. In summary, it indicates that 1108 of the 2,334 incidents were accurately labelled as low intervention, whereas 1070 instances were correctly labelled as intensive intervention. This classifier identified 91 instances as a low intervention when they should have been classified as intensive intervention. Again, 65 cases were wrongly labelled as an intensive intervention when they should have been classified as low intervention. The misclassification of the two groups might be because if low intervention status happens, there is also a potential for intensive intervention status to occur, and vice versa.

**Evaluation of Model using the Receiver Operation Curve (ROC)**

ROC curves with averaged thresholds for all five classifiers were generated, and their Area under Curves (AUCs) were evaluated using 10-fold cross-validation. Finally, the ROC graph is constructed and shown in Fig. 4.2.

From the ROC graph, it can be deduced that random forest achieves a superior classification metrics compared to the rest of the four classifiers (i.e., Rule Induction, Naïve Bayes, Logistic Regression and Deep Learning). The thick red line represents the curve for the random forest with an AUC of 0.980.

### 4.3 Determining students’ intervention type

Eventually, Fig. 4 illustrates the classifier’s conclusive results after considering all features. The major purpose of the study was to discover the demographic determinants of learners’ academic success.

These determinants help educational administrators define learners as needing intensive or low intervention. Per the confusion matrix class prediction of the random forest model in table……., Out of the 2334 upscaled sample understudy, 1173 (50.26%) were labelled as needing low intervention, whilst 1161 (49.74%) of the second-year students whose data was used were classed as needing intensive academic intervention to enhance their performance. As a result, it is possible to examine and conclude that the model effectively categorized the 2334 students in accordance to the type of intervention they need to boost their performance.

### 5.0 CONCLUSION

Traditional educational establishments’ academic intervention techniques do not give the necessary information for resolving students’ academic challenges. This eventually leads to the wrong selection of academic intervention methods to prevent low learner accomplishment. EDM is a recently acknowledged branch of research that mines massive volumes of data to answer education-related research questions concerning the learning process and its associated challenges. EDM has placed a strong emphasis on creating mechanisms for analyzing distinctive and increasingly large-scale data within academic contexts and using these systems to assess learners and the learning environments more effectively. The main goal of this study is to find distinctive and fascinating trends in students’ academic records that might help them with their academic work.

The fundamental purpose of this research is to identify relevant and intriguing patterns in individuals' academic histories that may assist them in performing better in Ghanaian second-cycle educational institutions.

The Hybrid Data Mining Process Model was employed in this study, and it comprises six phases that the researcher methodically goes through and iterates as needed. Five classifiers for classification techniques were used within the Rapid Miner's studio educational 9.10.010 environment. The RF model was constructed using 2334 datasets within the Rapid Miner repository. In all, 17 attributes with one class variable (students’ semester
average score) inclusive, were used in the model construction procedure. Twenty different tests were conducted. The experiments’ results were analyzed using 10-fold cross-validation and ratio split validation with bootstrap sampling. The Random Forest algorithm, rule induction methods, Naive Bayes, Logistic Regression, and deep learning algorithms were used in each experiment. The experimental results demonstrated that the random forest method outperforms the other four techniques in all six-evaluation metrics were employed for the selection process with the accuracy being 93.96%.

The study's primary goal was to identify students' demographic characteristics that influence their academic achievement. According to the random forest classifier model, the mother's and father's education levels are two recognized demographic factors per this study that significantly influence pre-tertiary students’ academic achievement. This study has significantly reduced the gap in practical knowledge observed in the literature by introducing an intervention scheme for respective student's requiring intensive or minimal academic interventions in its prediction procedure.

RECOMMENDATIONS

The researchers are confident that relevant academic institutions could subsequently explore the study's findings to fine-tune the implementation of their academic interventions. Therefore, the researchers offer the following recommendations based on the study's findings:

This study employs a collection of attributes regarded as more relevant by domain experts. However, examining the situation reveals that several attributes are omitted from the dataset. Capturing crucial characteristics such as class attendance, family size, disabilities, teachers’ competence, social media network usage, psychological factors, and study methods may aid in appropriately classifying the student. As a result, pre-tertiary education institutions must scale up the size of their database systems in order to capture more student biodata.

Education administrators and other stakeholders should use the recommended possible set of characteristics of students to implement appropriate strategies to address student academic weaknesses. This can be done by deploying the suggested model to identify potential traits that will hinder a student's performance.

Statements and Declarations

Funding.
This research did not receive any specific grant from funding agencies in the public, commercial or non-for-profit sectors.

Competing interest.
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author contributions
All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Iddrisu Issah, Obed Appiah, Peter Appiahene and Fuseini Inusah. The first draft of the
manuscript was written by Iddrisu Issah and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References


**Figures**
Figure 1

Fig2.1: concept of a Decision tree ((David Kolo, A. Adepoju, & Kolo Alhassan, 2015))

Figure 2

Fig2.2: Bayesian graphical model
Figure 3

Fig2.3: A neural network with one hidden layer.

Figure 4

Fig3.1: A line graph of Information Gain of Attributes
Figure 5

Fig3.2: The six-step KDP model (adapted from (Cios, Pedrycz, Swiniarski, & Kurgan, 2007))
Figure 6

Fig4.1: Order of attributes according to weights of importance

Figure 7

Fig4.2: ROC curves to compare the performance of random forest and the other classifiers in the study.
Figure 8

Fig 4.3: Number of students classified as in need of low or intensive intervention