Exploring the Impact of Students Demographic Attributes on Performance Prediction through Binary Classification in the KDP Model

Issah Iddrisu [a,1,*], Peter Appiahene [a,2], Obed Appiah [a,3], Inusah Fuseini [b,4]

[a] *University of Energy and Natural Resources, Post Office Box 214, Sunyani, Ghana*
[b] *University for Development Studies, Unnamed Road, Tamale, Ghana*
[1] *issah.iddrisu.stu@uenr.edu.gh*; [2] peter.appiahene@uenr.edu.gh; [3] obed.appiah@uenr.edu.gh;*
[4] *obed.appiah@uenr.edu.gh*
* corresponding author

ARTICLE INFO	ABSTRACT
	During the course of this research, binary classification and the Knowledge Discovery Process (KDP) were used. The experimental and analytical capabilities of Rapid Miner's 9.10.010 instructional environment are supported by five different classifiers. Included in the analysis were 2334 entries, 17 characteristics, and one class variable containing the students' average score for the semester. There were twenty experiments carried out. During the studies, 10-fold cross-validation and ratio split validation, together with bootstrap sampling, were used. It was determined whether or not to use the Random Forest (RF), Rule Induction (RI), Naive Bayes (NB), Logistic Regression (LR), or Deep Learning (DL) methods. RF outperformed the other four methods in all six selection measures, with an accuracy of 93.96%. According to the RF classifier model, the level of education that a child's parents have is a major factor in that child's academic performance before entering higher education.

I. Introduction

Learner assessment is central to determining students' progress in every educational establishment. Evaluating students' performance, however, 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), student information systems (SIS), and 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 (ML) 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 affecting student learning and achievement levels. Much progress has been made in 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 [1]. ML 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 [2].

The ML 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 it according to the training examples. ML 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 ML has been successfully employed, its application on educational data is a promising area in research identified as Educational Data Mining (EDM). It involves creating processes to extract patterns embedded in datasets within educational settings [3]. This concept has been implemented to improve and assess educational activities and decision-making.

Prediction, which encompasses the subcategories of classification, regression, and density estimation, is a paradigm in EDM [4]. Relation mining, association mining, correlation mining, sequential pattern mining, and causative data mining are all types of clustering [5]. In addition, prediction also incorporates data distillation to aid in human logic and model finding. EDM has proven to be the primary source of solid and dependable data analysis regarding educational decision-making at the country's educational institutions [6][7]. It carefully identifies education challenges to determine appropriate solutions that address them. The inclusion of an Expert System in managing primary education due to EDM has been enumerated in [8] and [6]. Educational Data Mining has been used to track the academic welfare of students and the general administrative procedures of educational institutions worldwide [9][10].

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 [11]. Therefore, determining the characteristics associated with students' academic accomplishment has always aroused the interest of academics who work in EDM. Many 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 influence the learner's achievement, including the algorithms that produce the best prediction result.

Students' apparent poor performance in numerous educational establishments has been influenced by various predictors [12][13]. They include personal characteristics, intellectual ability, gender and aptitude tests, academic achievement, previous college accomplishments, and demographic characteristics [14] in modeling students' academic performance based on their cognitive and non-cognitive characteristics [11]. Seven ML heterogenous lazy classifiers were employed, including DT, KNN, ANN, LR, RF, AdaBoost, and SVM. They used the 10-fold and leave-one-out cross-evaluation techniques to evaluate the selected classifiers' predictive performance. The student's absent days (SAD) were the dominant feature for predicting students' academic success. It was also concluded that the RF, LR, and ANN were viable in predicting students' performance.

Implementation of ML to determine students' academic achievement based on the student's internal assessment data constructed an ANN algorithms-based prediction model [15]. The best classification accuracy attained by the model was 95.34% through the ANN. Furthermore, the Precision, Recall, F-Score, Accuracy, and Kappa Statistics efficiency were derived as rule-based decision specifications to discover the most practical classification methods. However, the study presented inconsistent observations on which specific machine learning model is most accurate in predicting students' performance. Investigating factors affecting students' performance at the postgraduate level by using the ANN for constructing the model [16]. The study presented a model using the deep learning approach for performance prediction based on 395 postgraduate students and 30 records within the R data mining environment. A comparison of the accuracy of the LR, the RF technique, and the ANN revealed that the LR performed with 12.339% accuracy. The RF gave an accuracy of 28.101%, and the ANN had an accuracy of 97.429% on the given dataset. With this prediction accuracy, it was concluded that ANN is more reliable and demonstrates improved classification results than other traditional classifiers. The dataset used in the study was based on the attributes from institutions of higher learning. It will be interesting to apply the same model to datasets of pre-tertiary institutions to validate the model's generalization.

Investigated the prediction of students' learning outputs and explored the likelihood of recognizing the critical features in the data to be used in creating the prediction model using visualization and the clustering algorithm techniques [17]. The outcome demonstrated the capability of the clustering algorithm in classifying significant indicators within the datasets. In addition, the study showed the efficiency of SVM and Learning Discriminant Analysis (LDA) algorithms in training educational datasets while giving satisfactory classification accuracy and reliability test rates. However, the small data set cannot be generalized to prove the model's efficacy on all educational datasets.

Three different ML technique was used to forecast student performance [18]. The DT, NB, and LR classifications were employed here. The feature engineering criteria and the modification and selection of dataset characteristics were applied to enhance the predictions made by ML algorithms. The dataset used was put in two separate categories. The research findings suggest that using ML to anticipate student performance may be helpful. The most successful method from the first dataset was NB classification, with 98% accuracy; DT did better for the second batch of data, with 78% accuracy. In the study, the specific attributes and techniques capable of determining future learning outcomes could not be identified, presenting a conceptual vacuum that warrants further investigation.

Studies on the relationships between the instructional strategies employed by instructors and educators and how they impact students' academic performance have recently attracted more attention. Most research focuses on achievement due to the use of assessment techniques such as class tests, homework, class exercises, project work, and semester examinations [19]. When predicting a student's future academic success, past grades from an academic institution are seen to have the appropriate amount of weight as enumerated by [20], mainly when those grades come from continuous assessment, which shows a student's early mastery of a topic and progress of the study. Explored the efficacy of assessments using examination techniques, class tests, assignments, and mid-semester quizzes, including the influence of lecturer response on students' performance [21]. The study's outcome revealed a correlation between the assessments students took and, eventually, the student's final grades. Another investigation 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. The possibility of being able to forecast, with a degree of accuracy, how a student will perform at an end-of-course examination [22]. The effects of giving assessment feedback on time to students often result in a small quantity of enhancement in the final grades [23].

Predicting the validity of previous achievements in determining students' performance in higher education [24]. The high school Scholastic Assessment Test (SAT) score marks and 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 clearly connected these three characteristics and students' university accomplishments. Among the factors that influence students' performance are school effects, socio-economic background, and personal traits hindering students' performance [12]. 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 Tinto (1975) [24]. This is further corroborated by referring to these phenomena on students' academic success as "a one-hundred-factor problem," as many researchers focused on different aspects of students' performance in different periods and came out with diverse conclusions [25].

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 [26]. 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 a student's home environment is a contributory factor to his performance.

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 student's academic performance in Mathematics and the Serbian language [27]. The result indicated that student affluence contributed the most 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 students' performance at the university level. Integrating demographic data alongside school results is recommended because learner achievement is based almost entirely on students' past exam results, mostly without consideration for the setting where some of these performances had been accomplished [28]. Again, research on student achievement and the associations with context-specific background variables and attainment in broader terms was limited mainly [12][13]. Hence the need to delve into the correlation between students' performance 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 students' demographic characteristics to track their achievements, 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.

The paper aims to identify and apply ML algorithms 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 Ghana's Senior High School (SHS) level?

II. 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.

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. In contrast, 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 Student's 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 1.

A. Dataset Optimization and Feature Extraction

Primary and real-world data will invariably contain imbalanced data challenges [29]. 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 [30]. The Synthetic Minority Oversampling Technique (SMOTE) with default settings was used as a sampling technique to upscale the minority classes (i.e., Students' demographic variables) to manage class imbalance within the features. The upscaling synthetically increased the number of demographic variables by 79% within the local repository of Rapid Miner after the SMOTE Up-sampling application.

Table 1. Attributes extracted from the database

No.	Attribute Name	Data Value	Data Type
1	Gender	[1]: Male = M: [2] Female = F	Nominal
2	Student position in the family	[1]:1st born, [2]: last born, [3]: others, [4]: only child	Numeric
3	Parents marital status	[1]: married,[2]: Single, [3]: widowed	Nominal
4	Fathers Edu.	[1]: Primary school, [2]: Junior High School (JHS), [3] Secondary school (SHS), [4]: Tertiary, [5]: Non	Nominal
5	Mothers' Edu.	[1]: Primary school, [2] Junior High School (JHS), [3]: Secondary school (SHS), [4]: Tertiary, [5]: Non	Nominal
6	Fathers' Occ.	[1]: Retired, [2]: Government, [3]: private sector employee, [4]: self-employment, [5]	Nominal
7	Mothers' Occ	[1]: Retired, [2]: Government, [3]: private sector employee, [4]: self-employment, [5]	Nominal
8	Sponsor	[1]: Self, [2]: Parent, [3]: scholarship, [4]: others	Nominal
9	Residential Status	[1]: Boarding, [2]: Day	Nominal
10	Type of JHS attended	[1]: Private, [2]: Public	Nominal
11	BECE Aggregate	[1]: 6-8, [2]: 9-11, [3]: 12-15, [4]: 16-19, [5]: 20-24, [6]: 25-30, [7]: above 30	Numeric
12	BECE Accumulated raw score	[1]: 50-100, [2]: 101-200, [3]: 201-300, [4]: 301-400	Numeric
13	First Semester AVG score	[1]: 00-45, [2]: 46-50, [3]: 51-55, [4]: 56-60, [5]: 61-65, [6]:66-70, [7]: [8]: 71-79, [9]:80 and above	Numeric
14	Region of Residence	[1]: N/R, [2]: A/R, [3]: G/R, [4]: C/R, [5]: U/W, [6]: U/E, [7]: S/R, [8]: NE/R, [9]: W/R, [10]: E/R, [11]: V/R, [12]: O/R, [13]: Ahafo/ R, {14]: Bono East, [15]: Bono, [16]: WN/R	Nominal
15	Integrated Science	[1]: A1, [2]: B2, [3]: B3, [4]: C4, [5]: C5, [6]: C6, [7]: D7, [8]: E8,[9]: F9	Nominal
16	English Language	[1]: A1, [2]: B2, [3]: B3, [4]: C4, [5]: C5, [6]: C6, [7]: D7, [8]: E8,[9]: F9	Nominal
17	Mathematics	[1]: A1, [2]: B2, [3]: B3, [4]: C4, [5]: C5, [6]: C6, [7]: D7, [8]: E8,[9]: F9	Nominal

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, as seen in Table 2. The operator "Weight by Information Gain", was used in RapidMiner to determine the order of the attributes. Figure 1 depicts the descending order of information gained from common attributes to class attributes.

Table 2. Attributes weights by information gain

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

B. Modeling Technique and Model Building

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 analysis. The RF algorithms from DT, RI algorithms from rule-based classifiers, the NB algorithm from Bayesian Networks, the LR algorithm from Regression, and DL 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, and their popularity in recent years in education-related classification problems.
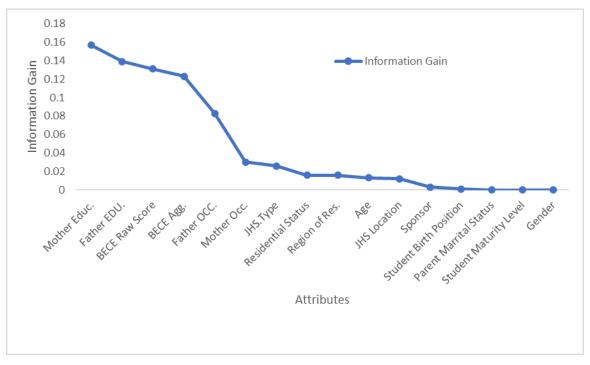
Fig. 1. A line graph of information gain of attributes

C. Description of the selected algorithms

First, a classification method is used to construct a decision tree (DT). 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 [31]. DT are made of nodes joined to constitute a rooted tree; therefore, it is a directed graph comprising nodes known as roots without incoming edges (Figure 1). The other nodes that determine the class of objects are known as the leaves or terminal nodes [32]. Every leaf is attributed to a class representing the most appropriate target value [33]. Nodes with a blend of diverse classes are to be split further. A stopping criterion determines when the decision tree algorithm should terminate. When an entire training sets in the terminal/leaf node fit within a particular class, then the stopping criterion is said to be reached [34]. Figure 2 illustrates a typical DT structure [35].
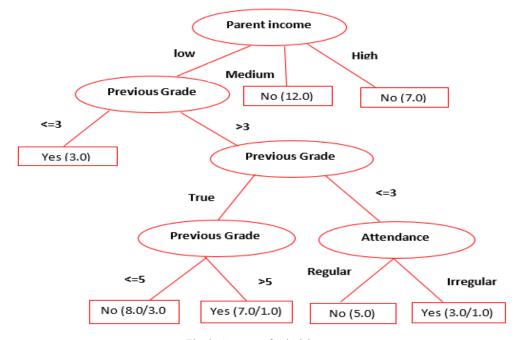

Fig. 2. Concept of a decision tree

Every node matches a characteristic, while the branches link with an array of values. All nodes are labeled with the attributes they test, and every branch has its corresponding values [36]. 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 (Figure 2).

Averaging ensemble approaches include the Random Forest (RF) algorithm. RF represent huge feature areas and are more resilient than DT. RF is a bagged classifier that connects a group of DT classifiers to form a forest of trees [37]. 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 [2]. In RF, 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 [38].

Random feature selection is used in an RF [39], where 'm' features are chosen randomly from 'M' features for every node of the DT "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 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 [40]. However, averages generally compensate for an overall model's increase in bias. Table 3 shows a description of RF-optimized parameters and data types within RapidMiner.

Table 3. Some random forest algorithm parameters with their values in RapidMiner

Parameters	Value	Description	Type
Criterion	Information Gain	Determine the criteria by which the qualities will be divided apart. The value is improved for each of these criteria about the selected parameter.	Nominal
Apply Pruning	True	Upon development, the random forest model's random trees can be trimmed. Depending on the confidence variable, some branches are supplemented by leaves if approved.	Boolean
Random Split	False	Rather than being balanced, this setting divides numerical characteristics randomly.	Boolean
Number of Trees	20	This indicates how many random trees will be produced.	Numeric
Maximal Depth	10	A tree's depth fluctuates According to the supplied example set's dimensions and properties.	Numeric
Confidence	0.1	This option specifies the confidence level used in pruning's pessimistic error computation.	Numeric
Voting strategy	Majority Vote	Outlines the prediction plan if the model prediction is in disagreement.	Nominal

The second is Bayesian Classification. The Bayesian classifiers also called the Naïve Bayes (NB), are based on statistical classifiers derived from the Bayesian theorem [41]. The accuracy and speed of the Bayesian classifiers have been proven to be of high magnitude on large databases [14]. The Bayesian classification offers a pictorial view of underlying associations on which to perform learning. A Trained Bayesian holds that networks can be helpful in classification [42]. A Bayesian classification graphical model is indicated in Figure 3.
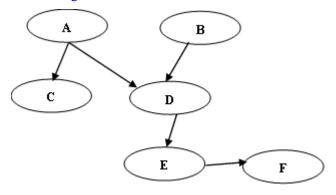

Fig. 3. Bayesian graphical model

Let X denote a data tuple labeled as the measurements on n attributes. Let H denote the hypothesis. Then, $P(H|X)$ denotes the probability of H being actual is based on X. $P(H|X)$ denotes the probability of H, conditioned on X. On the other hand, $P(H)$ denotes the prior probability of hypothesis H. Correspondingly, $P(H|X)$ denotes the posterior probability of X conditioned on H, while $P(X)$ is the prior probability of X. The Bayesian theorem offers a criterion for computing the posterior probability, $P(H|X)$ from $P(H)$, $P(X|H)$, and $P(X)$. The equation is denote in (1).

$$P(H|X) = \frac{P\left(\frac{X}{H}\right)P(H)}{P(X)} \tag{1}$$

For classification problems, X will represent an observed data tuple, assuming H as a 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 [43]. The NB algorithm makes learning simple by assuming that variables are autonomous of a specific class while offering a probabilistic interpretation of classification [11]. Though autonomy is a wrong assumption in general, the NB classifier frequently outperforms more advanced classifiers in practice. For example, while employing NB to analyze university and primary school students' performance, [44] found that the NB algorithm had superior accuracy in predicting the performance of primary school students.

Third is rule-based classifiers. A typical rule is described as follows: *IF* a condition exists, *THEN* the result [32]. The antecedent condition is on the rule's left side and consists of a variety of logical operators, comprising of >, <, =, & & OR, mainly employed on feature variables. The consequent that generates the class variable is on the rule's right side. RI is a rule presented as Qi→c, Qi being the antecedent and C as a class variable. The symbol → epitomizes a condition "*THEN*". The symbol Qi denotes a condition applied to the feature set [43]. A rule is of the form: *IF* (attribute 1; value 1) and (attribute 2; value 2) and …… (Attribute n; value n) *THEN* (decision; value).

Rule induction is experimented with in the study and is a widely applied rule-based classification technique. As stated, [33] rules are good when denoting information and aspect of information. RI 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 DT, develops a partial C4.5 decision tree, and translates the "best" leaf into a rule [41]. Typically, an if-then rule has the form: IF mother education = primary AND mother occupation = Government AND JHS location= Urban THEN Status = Low Intervention.

Fourth is Support Vector Machines (SVM). SVM is a learning algorithm 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 multilayer perceptron (MLP) classifiers [14]. The SVMs are derived from the statistical learning concept, which aims to solve related problems, except more complex ones, as a transitional step [45]. 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 fewer data samples for the training; investigations indicate that SVM is not sensitive to sample dimensions [46].

Fifth is neural networks (NN), simulating humans' NN system. It comprises an interrelated cluster of artificial neurons processing information based on a connectionist technique for calculations [3]. The NN 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) [46].

Every node in the input layer of an artificial neural network (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 input node is connected to other hidden layer nodes within the network. The inner layer nodes are linked via other inner layers or directly to an output layer. One or several response variables constitute the output layer [32].

Next to the input layer, the other nodes take in inputs, multiply the inputs by a connection weight Wxy (nodes 1 to 3 are put as $W13$), sum them, and then apply a function (known as activation or squashing function) to them, then transfer the results to the next layer. For instance, values passed from nodes 4 to 6 are put as activation functions to ($[W14$ * value of node 1] + $[W24$ * value of node 2]). Figure 4 depicts an NN structure.
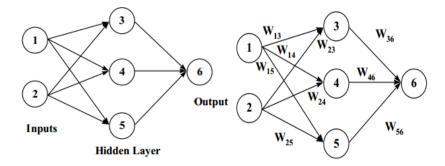

Fig. 4. A neural network with one hidden layer

The most basic deep networks are feed-forward deep networks, commonly known as multilayer perceptron (MLPs) [46]. The MLP is the most implemented NN architecture in predictive data mining. The MLP is based on the feed-forward deep network with many possibly concealed layers, with the input and output layers connected [46]. 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 [33].

The last is regression, commonly employed in predictive model building and the analytical processes in data mining. Regression predictions are primarily centered on historical data using functions and formulas [47]. It is mainly a statistical approach to data mining. Regression is implemented to derive a model between dependent and independent variables [47]. Regression is also used to build a model to analyze existing datasets to forecast trends using linear or logistic regression (LR) 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 [48].

The LR algorithm is applied to build a regression model using categorical dependent variables. LR is put into three categories (1) binary, in the case of binary response variables, (2) multinomial – for the above two non-ordered dependent variables (3) ordinal for an ordered category [33]. Researchers and data analysts generally use LR to analyze and classify proportional and binary response data [49]. The LR can effortlessly handle probability and multi-class issues in classification.

D. Research Design and Evaluation Metrics

This study is based on experimental research that employs binary classification techniques. The data comprised 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 [33]. 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 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 modeling step; and has numerous novel, clear, and specific feedback loops [33]. Figure 5 adopted indicates the six steps KDP modeling approach comprising of understanding the research problem, understanding data, preparation of data, mining the data, analyzing the knowledge base, and using the knowledge that has been discovered [50].
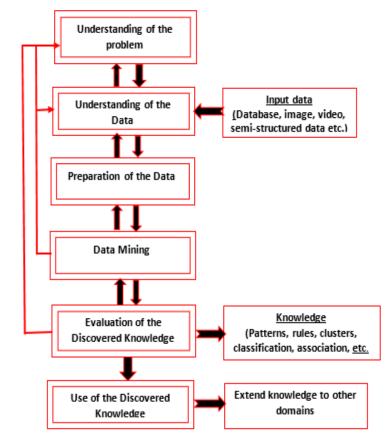
Fig. 5. The six-step KDP model

Evaluation of model performance is an essential rating for models' effectiveness, improving parameters during the iterative learning process, and choosing an acceptable model from an assortment of models [51]. The following six widely known performance metrics were used to compare and select algorithms for evaluating the classification task: accuracy, precision, sensitivity, specificity, AUC, and F-measure to construct a robust model.

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. The equation for accuracy can be seen in (2). 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 a positive class (i.e., summation of true positives and false positives) as in (3). Here, recall can be explained as a ratio of the number of true positives to the overall count of instances belonging to the positive class (i.e., summation of true positives and false negatives). i.e., instances not considered to belong to the positive class, though they belong to it). Recall carries the same value as sensitivity in model performance denote in (4).

$$Accuracy = \frac{TP+TN}{TP+TN+FN} \tag{2}$$

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

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

as in (5). It is essential to determine if a model's accuracy and recall are pretty well balanced [52]. The "true negative rate" is the name given to specificity. It provides information on the percentage of actual instances of negativity that a given model has correctly predicted as negative denote as in (6). It measures the proportion of real negatives to all negatives.

$$F - Measure = \frac{2\ X\ Precision\ X\ Recall}{Precision + Recall} \tag{5}$$

$$Specificity = \frac{True\ Negative}{True\ Negative + True\ Positive} \tag{6}$$

The area under the ROC curve (AUC) calculates the area under the ROC curve from (0,0) to (1,1) in two dimensions. The AUC defines an overall assessment of performance across all potential categorization criteria. AUC may be seen as the likelihood that a random positive instance will be ranked higher than a random negative instance in a given model.

E. Experimental Settings and Experimentation with selected algorithms

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 in 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.

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

A pictorial representation of the study method is illustrated in Figure 6.

III. Result and Discussion

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

A. 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 ML classifier capable of predicting students' academic success based on demographic characteristics. Five algorithms were explored to implement the classification modeling: RF, RI, NB, LR, and DL. The results of the experiments are presented in Table 4.
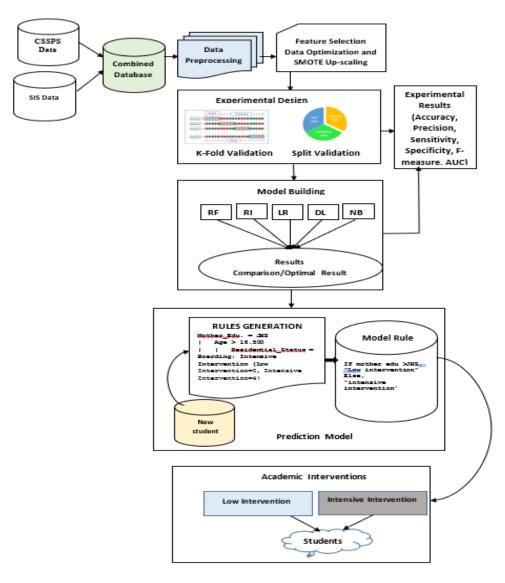
Fig. 6. A pictorial depiction of the study framework

Table 4. Summary of best-performing models from the five algorithms

Algorithm	Test mode	Accuracy	Precision	Sensitivity	Specificity	F-Measure	AUC
RF	Pruning using Bootstrap resampling with 10-Fold Cross-validation	93.96%	93.19%	94.97%	92.94%	94.04%	0.980
RI	With Ratio Split validation	83.00%	83.48%	82.29%	83.71%	82.88%	0.879
NB	Using Split validation	79.43%	78.77%	80.57%	78.29%	79.66%	0.879
LR	Using split validation	81.57%	80.44%	83.43%	79.71%	81.91%	0.892
DL	Using Bootstrap resampling with 10-Fold Cross Validation	84.45%	82.15%	88.49%	80.35%	85.11%	0.924

Comparing the six-performance metrics in Table 4, RF (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%, a precision of 93.19%, a sensitivity of 94.97%, a specificity of 92.94%, an F-measure of 94.04%, and an AUC of 0.980. As a result, the RF result with 10-fold cross-validation and bootstrap resampling was selected as the proposed model for the study.

B. 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 RF classifier. These weights were ordered in descending. The list's top two attributes were considered the most relevant in the model choice process. According to the RF classifier model simulator, the mother's and father's education levels (with the 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 7 depicts the order in which the attributes in support of the prediction are arranged according to their weights.
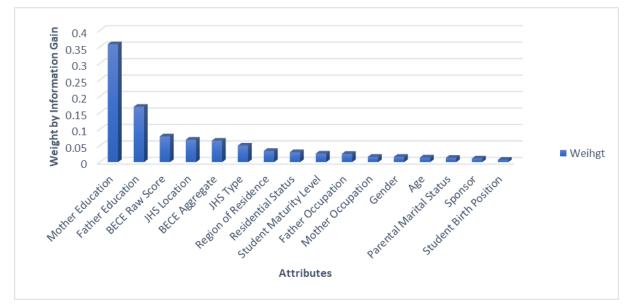

Fig. 7. Order of attributes according to weights of importance

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.

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. 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, NB, LR, and DL classifiers to construct a robust model.

Table 5 displays the confusion matrix of the chosen model, created using the RF algorithm and the Bootstrap resample approach with 10-fold cross-validation.

Table 5. Confusion matrix evaluation for random forest model

		PREDICTED CLASS		
		A	**B**	Classified As
		POSITIVE	NEGATIVE	
ACTUAL CLASS	POSITIVE	(TP)= 1108	(FN)=65	**A**=Low Intervention
	NEGATIVE	(FP)=91	(TN)= 1070	**B**= Intensive Intervention

Much may be learned by meticulously scrutinizing the errors generated by any classification model. The errors show discrepancies between the model's predictions and the tangible outcome in the actual business situation. When an appropriate model is discovered, the next step is determining why classification inaccuracies happened in the testing data. For instance, when predicting an attribute for a certain class label, 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. It indicates that 1108 2,334 incidents were accurately labeled as low intervention, whereas 1070 instances were correctly labeled 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 labeled as an intensive interventions 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.

ROC curves with averaged thresholds for all five classifiers were generated, and their Arear under Curves (AUCs) were evaluated using 10-fold cross-validation. Finally, the ROC graph is constructed and shown in Figure 8.
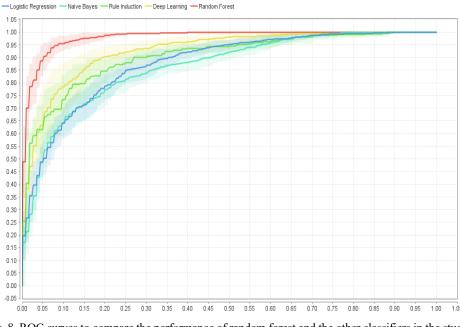

Fig. 8. ROC curves to compare the performance of random forest and the other classifiers in the study

From the ROC graph, it can be deduced that random forest achieves superior classification metrics compared to the rest of the four classifiers (i.e., RI, NB, LR, and DL). The thick red line represents the curve for the random forest with an AUC of 0.980.

C. Determining Students' Intervention Type

Eventually, Figure 8 illustrates the classifier's conclusive results after considering all features. The study's primary purpose 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 4, Out of the 2334 upscaled sample understudy, 1173 (50.26%) were labeled as needing low intervention, while 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 according to the type of intervention they needed to boost their performance. The student's classification by intervention types is illustrated in Figure 9.
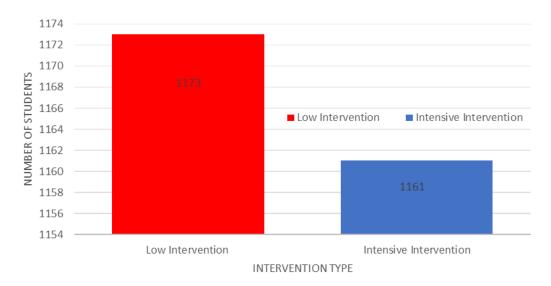
Fig. 9. Number of students classified as in need of low or intensive intervention

Overall, the RF classifier emerged as the best classification technique for the task from this study. The RF classifier correctly classified 2193(93.96%) instances, while 141(6.04%) instances were incorrectly classified. According to [53] in "Estimates of highly accurate models", the RF model is highly viable for performance determinants prediction since its accuracy extends beyond the 75% lower bound benchmark.

Again, the mother's and father's education levels (with information gains of 0.358 and 0.168, respectively) are the recognized demographic factors per this study that significantly influence pre-tertiary students' academic achievement. In their study, this finding is confirmed by [54] that well-educated parents prioritize a text-rich home environment, enhancing their academic achievement.

IV. Conclusion

The proposed demographic-based predictive model offers an innovative approach to predict learner performance accurately and recommend appropriate intervention schemes. By leveraging demographic information, educational institutions can provide targeted support to students, ultimately enhancing their educational experience and improving academic outcomes. This study has significantly reduced the gap in practical knowledge observed in the literature by introducing an intervention scheme for respective students requiring intensive or minimal academic interventions in its prediction procedure.

Declarations

Author contribution

All authors contributed equally as the main contributor of this paper. All authors read and approved the final paper.

Conflict of interest

The authors declare no known conflict of financial interest or personal relationships that could have appeared to influence the work reported in this paper.

Additional information

Reprints and permission information are available at http://journal2.um.ac.id/index.php/keds.

Publisher's Note: Department of Electrical Engineering and Informatics - Universitas Negeri Malang remains neutral with regard to jurisdictional claims and institutional affiliations.

References

[1] R. Hasan, S. Palaniappan, S. Mahmood, K. U. Sarker, and A. Abbas, "Modelling and predicting student's academic performance using classification data mining techniques," Int. J. Bus. Inf. Syst., vol. 34, no. 3, pp. 403–422, 2020.

[2] M. N. Yakubu and A. M. Abubakar, "Applying machine learning approach to predict students' performance in higher educational institutions," Kybernetes, vol. 51, no. 2, pp. 916–934, 2022.

[3] M. Arashpour et al., "Predicting individual learning performance using machine-learning hybridized with the teaching-learning-based optimization," Comput. Appl. Eng. Educ., vol. 31, no. 1, pp. 83–99, 2023.

[4] D. M. Ahmed, A. M. Abdulazeez, D. Q. Zeebaree, and F. Y. H. Ahmed, "Predicting University's Students Performance Based on Machine Learning Techniques," 2021 IEEE Int. Conf. Autom. Control Intell. Syst. I2CACIS 2021 - Proc., no. August, pp. 276–281, 2021.

[5] F. de Galiza Barbosa et al., "Genitourinary imaging," Clinical PET/MRI. pp. 289–312, 2022.

[6] F. Inusah, Y. M. Missah, N. Ussiph, and F. Twum, "Expert System in Enhancing Efficiency in Basic Educational Management using Data Mining Techniques," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 11, pp. 427–434, 2021.

[7] F. Inusah, Y. M. Missah, U. Najim, and F. Twum, "Data Mining and Visualisation of Basic Educational Resources for Quality Education," Int. J. Eng. Trends Technol., vol. 70, no. 12, pp. 296–307, Dec. 2022.

[8] F. Inusah, Y. M. Missah, U. Najim, and F. Twum, "Integrating expert system in managing basic education: A survey in Ghana," Int. J. Inf. Manag. Data Insights, vol. 3, no. 1, p. 100166, 2023.

[9] F. Inusah, Y. M. Missah, U. Najim, and F. Twum, "Agile neural expert system for managing basic education," Intell. Syst. with Appl., vol. 17, no. December 2022, p. 200178, 2023.

[10] H. Drachsler and W. Greller, "Privacy and analytics - it's a DELICATE issue a checklist for trusted learning analytics," ACM Int. Conf. Proceeding Ser., vol. 25-29-Apri, no. April, pp. 89–98, 2016.

[11] B. Owusu-Boadu, I. K. Nti, O. Nyarko-Boateng, J. Aning, and V. Boafo, "Academic Performance Modelling with Machine Learning Based on Cognitive and Non-Cognitive Features," Appl. Comput. Syst., vol. 26, no. 2, pp. 122–131, 2021.

[12] I. Issah, O. Appiah, P. Appiahene, and F. Inusah, "A systematic review of the literature on machine learning application of determining the attributes influencing academic performance," Decis. Anal. J., vol. 7, no. March, p. 100204, 2023.

[13] M. Tadese, A. Yeshaneh, and G. B. Mulu, "Determinants of good academic performance among university students in Ethiopia: a cross-sectional study," BMC Med. Educ., vol. 22, no. 1, pp. 1–9, 2022.

[14] F. Ouatik, M. Erritali, F. Ouatik, and M. Jourhmane, "Predicting Student Success Using Big Data and Machine Learning Algorithms," Int. J. Emerg. Technol. Learn., vol. 17, no. 12, pp. 236–251, 2022.

[15] S. Hussain and M. Q. Khan, "Student-Performulator: Predicting Students' Academic Performance at Secondary and Intermediate Level Using Machine Learning," Ann. Data Sci., no. Ml, 2021.

[16] V. K. Pal and V. K. K. Bhatt, "Performance prediction for post graduate students using artificial neural network," Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 7, pp. 446–454, 2019.

[17] L. M. Abu Zohair, "Prediction of Student's performance by modelling small dataset size," Int. J. Educ. Technol. High. Educ., vol. 16, no. 1, 2019.

[18] M. Pojon, "Using Machine Learning to Predict Student Performance," Univ. Tampere, no. June, pp. 1–28, 2017.

[19] B. Sekeroglu, K. Dimililer, and K. Tuncal, "Student Performance Prediction and Classification Using Machine Learning Algorithms," in Proceedings of the 2019 8th International Conference on Educational and Information Technology, Mar. 2019, pp. 7–11.

[20] M. N. Yakubu and A. M. Abubakar, "Applying machine learning approach to predict students' performance in higher educational institutions," Kybernetes, no. June, 2021.

[21] F. Aman, A. Rauf, R. Ali, F. Iqbal, and A. M. Khattak, "A Predictive Model for Predicting Students Academic Performance," in 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), Jul. 2019, pp. 1–4.

[22] J. López-Zambrano, J. A. L. Torralbo, and C. Romero, "Early prediction of student learning performance through data mining: A systematic review," Psicothema, vol. 33, no. 3, pp. 456–465, 2021.

[23] A. I. Adekitan and E. Noma-Osaghae, "Data mining approach to predicting the performance of first year student in a university using the admission requirements," Educ. Inf. Technol., vol. 24, no. 2, pp. 1527–1543, 2019.

[24] Y. Altujjar, W. Altamimi, I. Al-turaiki, and M. Al-razgan, "Predicting Critical Courses Affecting Students Performance : A Case Study," Procedia - Procedia Comput. Sci., vol. 82, no. March, pp. 65–71, 2016.

[25] A. Ahadi, R. Lister, H. Haapala, and A. Vihavainen, "Exploring machine learning methods to automatically identify students in need of assistance," ICER 2015 - Proc. 2015 ACM Conf. Int. Comput. Educ. Res., pp. 121–130, 2015.

[26] D. T. Ha, C. N. Giap, P. T. T. Loan, and T. L. H. Huong, "An Empirical Study for Student Academic Performance Prediction Using Machine Learning Techniques," Int. J. Comput. Sci. Inf. Secur., vol. 18, no. 3, pp. 21–28, 2020.

[27] J. David and G. Anastasija, "Predicting Academic Performance Based on Students ' Family Environment : Evidence for Colombia Using Classification Trees," vol. 11, no. 3, pp. 299–311, 2019.

[28] M. I. Al-Twijri and A. Y. Noaman, "A New Data Mining Model Adopted for Higher Institutions," Procedia Comput. Sci., vol. 65, no. Iccmit, pp. 836–844, 2015.

[29] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," J. Artif. Intell. Res., vol. 61, pp. 863–905, 2018.

[30] P. D. Jenssen, T. Krogstad, and K. Halvorsen, "Community wastewater infiltration at 69 o northern latitude – 25 years of experience," Soil Sci. Soc. Am. Onsite Wastewater Conf. , Albuquerque NM, 7-8 April 2014, no. April, pp. 7–8, 2014.

[31] C. Anuradha and T. Velmurugan, "Fast Boost Decision Tree Algorithm: A novel classifier for the assessment of student performance in Educational data," vol. 31, pp. 254–0223, 2016.

[32] P. Bhatia, "Introduction to Data Mining," Data Min. Data Warehous., pp. 17–27, 2019.

[33] S. Samson, "Use of Data Mining For Determining Higher Education Students 'Performance," St. Mary's University, 2019.

[34] J. Feng, "Predicting Students' Academic Performance with Decision Tree and Neural Network," University of Central Florida, 2019.

[35] K. David Kolo, S. A. Adepoju, and J. Kolo Alhassan, "A Decision Tree Approach for Predicting Students Academic Performance," Int. J. Educ. Manag. Eng., vol. 5, no. 5, pp. 12–19, 2015.

[36] Y. Liu, S. Fan, S. Xu, A. Sajjanhar, S. Yeom, and Y. Wei, "Predicting Student Performance Using Clickstream Data and Machine Learning," Educ. Sci., vol. 13, no. 1, 2023.

[37] M. M. Z. Eddin, N. A. Khodeir, and H. A. Elnemr, "A Comparative Study of Educational Data Mining Techniques for skill-based Predicting Student Performance," Int. J. Comput. Sci. Inf. Secur., vol. 16, no. 3, pp. 56–62, 2018.

[38] P. Sokkhey and T. Okazaki, "Hybrid machine learning algorithms for predicting academic performance," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 1, pp. 32–41, 2020.

[39] M. N. Yakubu, "Applying machine learning approach to predict students ' performance in higher educational institutions," no. June, 2021.

[40] Y. Denny, H. Leslie, H. Spits, and W. Budiharto, "Systematic Literature Review on Abstractive Text Summarization," no. November, 2021.

[41] K. Blackmore and T. R. J. Bossomaier, "Comparison of See5 and J48.PART Algorithms for Missing Persons Profiling," no. December, 2016.

[42] F. Ofori, E. Maina, and R. Gitonga, "Using Machine Learning Algorithms to Predict Students' Performance and Improve Learning Outcome: A Literature Based Review," J. Inf. Technol., vol. 4, no. 1, pp. 2616–3573, 2020.

[43] S. Agrawal, S. K., and A. K., "Using Data Mining Classifier for Predicting Student's Performance in UG Level," Int. J. Comput. Appl., vol. 172, no. 8, pp. 39–44, 2017.

[44] D. Gašević, V. Kovanović, and S. Joksimović, "Piecing the learning analytics puzzle: a consolidated model of a field of research and practice," Learn. Res. Pract., vol. 3, no. 1, pp. 63–78, 2017.

[45] P. G. Sameer and S. R. Barahate, "Educational Data Mining – A New Approach to the Education Systems," pp. 18–20, 2016.

[46] A. S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student Performance Prediction Model based on Supervised Machine Learning Algorithms," IOP Conf. Ser. Mater. Sci. Eng., vol. 928, no. 3, 2020.

[47] C. A. Palacios, J. A. Reyes-Suárez, L. A. Bearzotti, V. Leiva, and C. Marchant, "Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in chile," Entropy, vol. 23, no. 4, pp. 1–23, 2021.

[48] D. T. Larose and C. D. Larose, "Data Mining and Predictive Analytics (Wiley Series on Methods and Applications in Data Mining): 9781118116197: Computer Science Books @ Amazon.com," Wiley Ser., p. 794, 2015.

[49] M. Maalouf, "Logistic regression in data analysis: an overview," Int. J. Data Anal. Tech. Strateg., vol. 3, no. 3, p. 281, 2011.

[50] K. J. Cios, W. Pedrycz, R. W. Swiniarski, and L. A. Kurgan, Data mining: A knowledge discovery approach. 2007.

[51] Y. Chen et al., "Evaluation efficiency of hybrid deep learning algorithms with neural network decision tree and boosting methods for predicting groundwater potential," Geocarto Int., vol. 37, no. 19, pp. 5564–5584, 2022.

[52] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," pp. 37–63, 2020.

[53] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques. 2016.

[54] M. Idris, S. Hussain, and N. Ahmad, "Relationship between Parents' Education and their children's Academic Achievement," J. Arts Soc. Sci., vol. 7, no. 2, pp. 82–92, Dec. 2020.