Predicting and Interpreting Student Performance Using Ensemble Models and Shapley Additive Explanations

HAYAT SAHLAOUI1, EL ARBI ABDELLAOUI ALAOUI2, ANAND NAYYAR3,4, SAID AGOUJIL1, AND MUSTAFA MUSA JABER5,6

1Department of Computer Science, Faculty of Sciences and Techniques at Errachidia, University of Moulay Ismaïl, Meknes, Errachidia 52000, Morocco
2Department of Sciences, Ecole Normale Supérieure, University of Moulay Ismaïl, Meknes 52000, Morocco
3Graduate School, Duy Tan University, Da Nang 550000, Vietnam
4Faculty of Information Technology, Duy Tan University, Da Nang 550000, Vietnam
5Department of Computer Science, Al-Turath University College, Baghdad 10021, Iraq
6Department of Computer Science, Dijlah University College, Baghdad 10021, Iraq

Corresponding authors: Anand Nayyar (anandnayyar@duytan.edu.vn) and Mustafa Musa Jaber (mmjh86@gmail.com)

ABSTRACT In several areas, including education, the use of machine learning, such as artificial neural networks, has resulted in significant improvements in predicting tasks. The opacity of these models is one of the problems with their use. Prediction models that may offer valuable insights while still being simple to comprehend are preferred by decision-makers in education. Hence, this study suggests an approach that improves the previous student performance prediction by enhancing performance and explaining why a student’s performance is attaining a certain score. A prediction model was proposed and tested using machine learning models. Our models outperform previous work models developed on the same dataset. Using a combined framework of data level and algorithm approaches, the proposed model achieves an accuracy of over 98%, implying a 20.3% improvement compared with previous work models. As a balancing technique for upsampling data, we use the default strategy of synthetic minority oversampling technique (SMOTE) to oversample all classes to the number of examples in the majority class. We also use ensemble methods. For tuning the parameters, we use a simple grid search algorithm provided by scikit to estimate the optimal parameters of our model. This hyperparameter optimization along with a ten-fold cross-validation process demonstrates the dependability of the novel model. In addition, a novel visual and intuitive technique is used to help determine which factors most influence the score which helps to interpret and understand the entire model and visualizes feature attributions at the observation level for the machine learning model. Therefore, SHAP values are a powerful tool that should be incorporated within the student performance prediction framework by obtaining the prediction and explanation created through the experiment, educators can recognize students at risk early and provide suitable exhortation in an auspicious manner.

INDEX TERMS Ensemble methods, game theory, machine learning, student performance prediction, SHAP values.

I. INTRODUCTION Data mining techniques play an essential role in many application fields, such as business analytics, security analytics, financial analytics, and learning analytics. In this study, we are primarily concerned with applications of data mining in the education environment. This area of research focuses on the design and application of algorithms on educational datasets to have a good understanding of students and their educational system [1]. A primal application of educational data mining (EDM) is investigating the student learning process and predicting student performance to improve educational practices. In this context, we are attempting to approximate student performance, experience, ranking, or grade [2] by pulling out features from traditional recorded or logged data. Student performance value could be numeric in the case of a regression problem or categorical in the case of a classification problem. Currently, one of the
most promising research fields of information technology is machine learning [3]. The application of machine learning in the education domain is the primary subject of our research [4]. This technology can help predict student performance and alert teachers about students at risk to provide them with the support they need. Therefore, the question of predicting student performance has to deal with building the learning classifier by using students’ observed records that represent the training set and matching student historical data that represent features with their label that represents the actual performance [5]. The final goal is to alert teachers about students at risk of drop out and suggest means to improve their performance and increase their retention and completion rate [6].

Linear regression models are used in many existing student performance studies. Linear models are easy to implement and interpret, but they are inaccurate in terms of modeling student grades. Studies have also addressed the shortcomings of linear models using nonlinear models, such as artificial neural networks [7]. In contrast to linear models, nonlinear models have been shown to achieve better performance [8]. However, interpreting the model is difficult (e.g., which factors make student performance inefficient). This predictive model answers the “how much” question, not the reason why. Therefore, the best case is to have a robust model that can be interpreted and result in improved adoption [9].

Robust machine learning algorithms can usually make accurate predictions, but their well-known opaque nature hampers their adoption. The interpretability of the model corresponds to the tag or the label on the pill bottle. The label needs to be transparent for easy adoption. Decision-makers in education prefer prediction models that provide useful insights, as well as those that are easy to understand [10], [11]. This clear-cut view additionally provides quality assurance to the pipeline. Intuition helps researchers in determining whether any significant errors exist when inputting or executing a model. Moreover, it is a smart way to regenerate the outcome of the prediction algorithm. Most machine learning-based projects always focus on results, not interpretability. Here, we highlight the latest method for interpretability, SHAP value to illustrate and explain the accuracy of the model.

This paper describes the first attempt to use machine learning model interpretability in the context of student performance prediction. Previous studies [12] focused on updating the accuracy of student grade models. However, this paper outlines, on the one hand, the first attempt to use a combined framework of data level and algorithm approaches to improve classifier performance measures, for instance, SMOTE technique for upsampling data and bagging- (ET) and boosting-based (XGBoost) algorithms for improving the performance of single classifiers. And on the other hand, the use of SHAP values [13] and associated visualizations in the quest of interpretability and explainability of models in an education context. This work aims to improve performance measure values by using the ensemble method and then compare it with previous studies that used the Jordan dataset to investigate student performance prediction. We also apply the SHAP value to explain the inner logic of the ensemble method.

In this study, we attempt to answer the following questions: i) From previous studies related to student performance, which input attributes affect student performance, which set of algorithms is used for prediction, what are their accuracy measures, and which tools are used?; ii) can the predictive models used in student performance be improved by using ensemble methods?; iii) is there an opportunity to use these proposed models to improve the interpretability of the models that can enhance the use of these models by decision-makers in education?; iv) what are the reasons and key indicators for explaining student prediction models using game theory?; v) what are Shapley values and how do these values help improve prediction accuracy and model transparency?

The key contributions of this study are as follows:

**General objectives:**
- This research could greatly benefit school administrators in terms of their knowledge of the causes of student performance problems and in finding solutions to them.
- The results of this study could benefit officials in the ministry of education in terms of finding educational plans, solutions, programs, and activities for students at risk.
- The study could contribute to theoretical literature to extend the scope and nourish the machine learning repository and provide noteworthy insights and models to improve education across the globe.
- This study could open horizons for other researchers to conduct other studies in this field.

**Specific objectives:**
- The combination of machine learning models and SHAP value could overcome the challenges of the trade-off between student performance models’ interpretability and complexity and improve model accuracy and transparency.
- improving student performance model accuracy without compromising their interpretability via the combination of a data level and algorithm approach:
  - by solving class imbalance problems that imply strategies, such as data augmentation for the minority class, which is referred to SMOTE.
  - by leveraging two innovative model types: bagging and boosting-based algorithms, in particular, ET and XGBoost, which were examined for the first time in the context of student performance data.
- to improve the precision and to avoid model overfitting, we used hyperparameter optimization technique by choosing a set of optimal hyperparameters for the learning algorithm. Tenfold cross validation was used to train and evaluate our models and analyze the performance of the algorithms with proper metrics.
- Augmenting our ensemble models with SHAP values [13]. In real-world machine learning-based applications, model interpretability can sometimes be more
important than accuracy. Therefore, SHAP provides a more visual, intuitive, and comprehensive approach to increase the transparency of ensemble models.

- Evaluating and comparing the proposed approach with previous works: By running our set of models, we were able to validate empirically that the proposed model outperforms previous ones.

This paper is organized as follows: In Section II, related work on the prediction of the performance of university students is reviewed. In Section III, the methodology followed to show how to use ensemble methods to build a student performance model in order to improve its performance and the methods that help explain the model has been presented. This process consists of four phases: a preprocessing phase, a model building phase, a model evaluation phase and a model interpretability phase. Section IV, explores the experimental protocols, results obtained, and discussion. By providing a complete description of the variables used in predicting student grades. After that, the focus was on implementing and interpreting the model. By showing the details of the experimental implementation of the model and the accuracy in comparison with traditional models. The implementation stage is then extended through research on how to explain the model used in student performance prediction. Finally Section V concludes the paper with limitations, suggestions and future scope.

II. RELATED WORK

In this section, we provide a research background for our study. For this, we reviewed previous studies related to student performance prediction and tried to layout the input attributes that affect student performance, the set of algorithms that are used for prediction, and their accuracy metrics, as well as the methods used to complete this critical mission in web-based educational contexts.

A. DETERMINANT ATTRIBUTES/FACTORS THAT AFFECT STUDENT ACADEMIC PERFORMANCE

The improvement of the education system depends on the factors that affect student academic success [14]. Therefore, to predict student performances with efficiency and accuracy, studying the variables that influence student academic performance is of paramount importance. A comprehensive framework on the factors that influence student academic performance is presented. The framework includes five variable aspects that considerably influence student performance: personal, academic, economic, social, and institutional domains, as shown in Figure 1. Each of the five domains contributes to the performance measurement of students and is made up of attributes that work individually and conjointly for learner success. However, the degree of complexity and influence of each domain on student performance varies.

1) PERSONAL FACTORS

Personal factors include student behavior, such as feelings, thoughts, or actions, as well as students’ demographics, such as age and gender, which affect their performance. Age and gender are the most often used factors for prediction because they are considered internal factors of variability, which are simple to define and measure. The researchers in [15]–[17] looked at how psychometric factors tend to affect performance of students.

2) ACADEMIC FACTORS

These factors refer to indicators that explain the success or failure of students in the academic track in universities (e.g., the score a student obtains). According to [18], the cumulative grade point average (CGPA) is the most important attribute that has been frequently used because of its huge effect in shaping education’s future. In [19]–[21], the authors considered student GPA.

In [22], the authors disclosed that previous academic performance and parent educational background are the most important attributes in predicting the future academic performance of a student. Others looked into the effect of previous academic achievement in determining the performance of students in the future [23].

3) FINANCIAL FACTORS

Financial factors refer to the financial ability of the parents to finance their children’s education and shape their future career. Research on these factors focused on the socioeconomic status of the student [24]. Some researchers studied the correlation between academic performance and parent’s educational level and income [25].

4) FAMILY FACTORS

Family factors are related to parent’s educational background and their ability to provide educational assistance to their children and create a propitious environment for learning. In [26], results revealed that the type of school does not influence student performance, but the parent’s job plays a critical role in predicting grades. In [22], the authors found that previous academic performance and parent’s educational background are the most important attributes in predicting a student’s future academic performance. Some studies tend to analyze the influence of parent’s education background and income on academic performance [25].

FIGURE 1. Determinant factors of student academic performance.
5) INSTITUTIONAL FACTORS
The factors that correspond to this category relate to the academic program and resources that the institution allocates for the best academic performance of its students. The authors of [15]–[17] looked at how psychometric factors tend to influence the performance of students.

B. DIFFERENT ALGORITHMS IN DEVELOPING A MODEL TO PREDICT THE ACADEMIC PERFORMANCE OF STUDENTS
The above review has shown that different researchers have used different algorithms in building a model to predict the academic performance of students. Among these approaches, the most frequently used ones are decision trees, linear regression, naïve Bayes classifiers, support vector machines, and neural networks. A few studies have used a coalition of classifiers in the quest of improving predictions [27]–[29]. Many of the studies reviewed tended to choose the optimal algorithm by trying a set of machine learning approaches. The decision tree consists of a series of rules organized in a stratified structure. Most researchers used this technique because it is straightforward, that is, it can be changed into numerous classification rules. The decision tree algorithm is widely used in predictive modeling because it is suitable for large and small amounts of data [1]. It is also good for numeric and categorical variables and does not require lengthy data preparation. It can be easily understood and interpreted by users because it is based on rules that are easy for users to understand and interpret. The survey by [30] showed that decision tree models are easy to understand due to their inner logic process and can be converted directly into if then rules. The decision tree revealed that C4.5, CART, and ID3 had the best classifiers for predicting student performance [5]. In [12], some well-known decision tree algorithms, such as C4.5 and CART, were used to predict student’s final grades using their log data extracted from the Moodle system. Meanwhile, [31] used ID3 and C4.5 algorithms on student’s internal marks to identify students who are likely to fail. Experiments were conducted on 200 engineering students’ datasets, and their performance was predicted by applying a decision tree. The results revealed that the accuracy of the classifiers was 98.5%. A comparative study of classification techniques to predict academic performance was conducted by [32]. This study used the data of 350 students. The result indicated that out of the classifiers, J48 achieved an accuracy of 97% [32], [33].

Neural networks are the most established model that has been used in educational data mining. Neural networks function similarly to the human brain by linking neurons or nodes that collaborate to produce an output task [34]. Neural network algorithms are also widely used by researchers because they are well suited to large amounts of data and can also be applied in structured and unstructured data; however, they are difficult to interpret and require more execution time for algorithm training. A neural network was used in [35] to predict student performance. Results indicated that the model achieved an accuracy of 83.4%. Arsad et al. [36] used an artificial neural network algorithm to predict the performance of undergraduate students in the field of engineering. This work discovered that undergraduate subjects have a major effect on the final CGPA after graduation. In the work of [19] it was pointed out that with the SMOTE method in an imbalanced dataset, classification models for neural networks and naïve Bayes had an accuracy of 75%. When using the discretization method, the models for naïve Bayes and neural networks achieve almost identical levels of accuracy.

The authors in [37] built a model to predict the GPA of the students using the naïve Bayes technique and K-means clustering. Their models produced an accuracy of 98.8%. By contrast, [38] indicated that a naïve Bayes classifier using the cfsSubsetEval feature selection technique in a dataset consisted of 257 academic records had an accuracy of 84%. In [16], the support vector machine was found to have a decent generalization potential and is quicker than other approaches. Meanwhile, an analysis performed in [17] for the purpose of recognizing students at risk of failure obtained the best accuracy in model prediction by using a support vector algorithm. In his academic work [39] applied a support vector machine to predict which students are to pursue doctoral studies. Results revealed that the method achieved an accuracy of 96.7%. While the authors in [40] revealed that the polynomial kernel technique achieved 97.62% accuracy. The support vector machine algorithm is the least favored by the researcher because it is only suitable when the dataset is small and the algorithm is not transparent. The only disadvantage is the complexity of the model. Such complexity contradicts the general requirement that models in intelligent learning systems should be transparent. In addition, selecting proper kernel functions and other parameters is difficult, and different experiments have to be conducted empirically.

Ensemble methods were introduced by Breiman, Freund, and Schapire in 1984. The most popular algorithms, such as arcing [41], boosting [42], bagging [43], and casual forests [44], have gained a keen interest in the literature. Ensemble methods have two major families: The first is called homogeneous because it combines similar algorithms, such as multiple classification trees; multiple SVMs, bagging [43], and boosting [42] are examples of such methods. The second is called heterogeneous because it combines algorithms of different nature (e.g., classification trees, SVM, and neural networks). The most popular method is the stacking method [45].

In his paper [5] applied the ensemble method to improve the results on the best-selected classifier J48 to predict student performance. The results revealed a huge improvement. That is, the proposed model using ensemble methods could reach an accuracy of up to 98.5%. By contrast, [12] showed that the proposed algorithm (i.e., bagging and boosting) obtained an accuracy of over 80%, confirming the soundness of the algorithm. To address the student performance prediction problem, various techniques that provide accurate results are being identified. Recently, many academic papers (as shown
in Table 1) elaborated on the proper features that affect student performance, and all of them agreed, that all models can provide reasonable results when predicting student performance. However, they cannot identify a technique that is clearly superior because the accuracy of the prediction depends mainly on the context, data, parameters, and hyperparameters of the technique. Any potential alternative must consider these factors.

**C. TOOLS USED**
The most popular used tools are WEKA and SPSS Modeler most likely due to their automation and versatility features in predictive analytics [46]. In addition, R and Python package are used due to their mature and supportive community and hundreds of open-source libraries and frameworks [47].

**D. ADDITIONAL VALUE OF THE SUGGESTED APPROACH**
By scouting literature regarding student performance prediction, we found a gap specific to student performance improvement in terms of how to boost model performance without neglecting to clear out their opacity making them easy for educators to understand. And provide guidance on how the model can be used to improve school administrator’s decision-making skills and confidence in the scores calculated by the ensemble models to achieve better model adoption. For better interpretability, leads to better acceptance [9]. This study is the first to examine two novel types of models related to student performance data (i.e., XGBoost and ET) and use them to boost the predictive accuracy of student performance models. In contrast to previous machine learning models, the algorithms used are scalable, provide accurate results, and require less execution time in training the algorithm. Moreover, they do not require a huge amount of data for training the algorithm in contrast to other robust models, such as deep learning. We also solved the problem of class imbalance, which implies strategies, such as data oversampling for the less represented class, known as SMOTE technique. As a balancing technique, SMOTE was used to oversample all classes to the number of examples in the majority class. To improve the precision and avoid model overfitting, we used a hyperparameter optimization technique by choosing a set of optimal hyperparameters for the learning algorithm. Ten-fold cross validation was used to train and evaluate our models. For tuning the parameters, we used a simple grid search algorithm provided by scikit to estimate the optimal parameters of XGBoost (booster = dart, n_estimators = 1000, learning_rate = 0.1, nthread = 4, objective = binary: logistic, eval_metric = mlogloss, eta = 0.7, gamma = 4, max_depth = 9, min_child_weight = 1, max_delta_step = 0, subsample = 0.8, colsample_bytree = 0.8, silent = 1, seed = 7, base_score = 0.7). For the ET, the optimized parameters were ( n_estimators = 300, max_features = 2). Cross validation is mainly used to gauge the capabilities of our model based on new data. We also analyzed the performance of the algorithms with proper metrics, such as precision and recall, which capture some information that accuracy cannot. When considering models for predicting student performance, misclassifying students at risk as “not at risk” has more consequences than if the opposite were the case. Thus, accuracy may not be the right choice as a performance evaluation technique in this case.

This paper describes the first attempt to use the interpretability of machine learning models in the context of student performance prediction. Previous studies [12] focused on updating the accuracy of student grade models. However, this paper outlines the first attempt to use SHAP values [13] and associated visualizations in the quest for the interpretability and explainability of models in the education context.

And this by augmenting our ensembles models with SHAP values [13]. In real-world machine learning-based applications, model interpretability can sometimes be more important than accuracy. Therefore, SHAP provides a more visual, intuitive, and comprehensive approach as a great way to increase the transparency of ensemble models, which help in interpreting and understanding the entire model and in visualizing feature attributions at the observation level for any machine learning model. In contrast to surrogate models, ensemble models, regardless of the model type (model agnostic), can be improved by SHAP via outliers and missing value detection, variable selection, which can help in discovering the root cause of a decrease in model performance, and proposing changes that might redress the issues. Which augment the trustiness of the ensembles models and lead to their better adoption [9]. Finally, the suggested models were assessed and compared with previous work. By running our set of models, we could empirically validate the performance of our student performance system versus previous work models.

**III. RESEARCH METHODOLOGY**
In this section, we show how ensemble methods are used to build a student performance model, as well as other methods that help explain the model.

**A. SYSTEM MODEL**
Figure 2 depicts the four stages of this operation, including the preprocessing phase, the development phase, the appraisal phase, and the interpretability phase.

The first phase begins with a data preprocessing step in which the collected data are converted into a suitable format. The discretization process is then used to translate student performance from numerical values to nominal values that correspond to the classification problem’s class tag. In this process, we divided the dataset into three nominal intervals depending on the student’s average grade (high performer, average performer, and low performer). The values of the low performers range from 0 to 69, those of the middle performers range from 70 to 89 and those of the high performers range from 90 to 100. The post discretization dataset includes several low-performing students, average performing students, and high performing students. Afterward, we solved
the problem of class imbalance, which implies strategies, such as data oversampling for the less represented class, known as SMOTE, by balancing classes in the training data before inputting data to the machine learning algorithm. This process plays a vital role in improving accuracy measures. The original training set was 480, of which 127 accounted for the low-performing class, 211 accounted for the average performers, and 142 accounted for the high performers. As a balancing technique, SMOTE was used to oversample all classes to the number of examples in the majority class. Afterward, supervised classification algorithms were used to train models to learn the mapping function of input features to the output labels by leveraging or examining two novel model types that are related to student performance data (i.e., XGBoost and ET). These models were used to boost the predictive accuracy of student performance models. The testing set was used to test the prediction performance of the models trained by different algorithms. Ten-fold cross validation was used to train and evaluate our models. For tuning the parameters, we used a simple grid search algorithm provided by scikit to estimate the optimal parameters of XGBoost (booster = dart, n_estimators = 1000, learning_rate = 0.1, nthread = 4, objective = binary: logistic, eval_metric = mlogloss, eta = 0.7, gamma = 4, max_depth = 9, min_child_weight = 1, max_delta_step = 0, subsample = 0.8, colsample_bytrees = 0.8, silent = 1, seed = 7, base_score = 0.7). For the ET, the optimized parameters were (n_estimators = 300, max_features = 2). We recorded the accuracy, precision, and recall values for model comparison. Afterward, we focused on model interpretation. For this, we used the SHAP value to explain the inner logic of the ensemble methods. As indicated by the Shapley value, every time we provide data as input for interpreter and predictor, the predictor model supplies the accuracy value, whereas the interpreter depicts the effect of negative and positive features. As a visualization technique, the SHAP value provides a view of the prediction model functionality and increases model transparency. Specifically, the SHAP force visualization provides a clear view of the features of
TABLE 2. Attributes, algorithms, and data mining techniques frequently used to predict students academic performance.

| Publication years | Ref | Attributes/Variables used | Algorithm | Performance | Data mining technique | Tools | Preprocessing Technique |
|-------------------|-----|----------------------------|-----------|-------------|-----------------------|-------|------------------------|
| 2016              | [12] | - Demographical Features  | - Ensemble Methods | Accuracy=79.1% Recall=79.2% Precision=79.1% F-Measure=79.1% | Classification | MS Excel | -Discretization -Normalizazion - Filer based method |
|                   |     | - Academic Background     |           |             |                       |       |                        |
|                   |     | - Features               |           |             |                       |       |                        |
|                   |     | - Parents Participating  |           |             |                       |       |                        |
|                   |     | on learning process      |           |             |                       |       |                        |
|                   |     | - Behavioral Features    |           |             |                       |       |                        |
|                   |     | - Student demographic    |           |             |                       |       |                        |
|                   |     | Behavioral attitude      |           |             |                       |       |                        |
|                   |     | Family background        |           |             |                       |       |                        |
|                   |     | BA1 RESULT               |           |             |                       |       |                        |
| 2016              | [31] | - Examination marks      | ID3 C4.5  | accuracy = 98.5% | Classification | MS Excel | -                        |
|                   |     | for all the courses      |           |             |                       |       |                        |
|                   |     | taught in the four years |           |             |                       |       |                        |
|                   |     | - Students’ pre-admission |           |             |                       |       |                        |
|                   |     | marks GPA                |           |             |                       |       |                        |
|                   |     | - Discussion             |           |             |                       |       |                        |
|                   |     | - Assignments            |           |             |                       |       |                        |
|                   |     | - Attendance             |           |             |                       |       |                        |
|                   |     | - Lab work               |           |             |                       |       |                        |
|                   |     | - Last semester’s        |           |             |                       |       |                        |
|                   |     | examination              |           |             |                       |       |                        |
|                   |     | - CGPA                   |           |             |                       |       |                        |
|                   |     | - Father’s financial     |           |             |                       |       |                        |
|                   |     | income                   |           |             |                       |       |                        |
|                   |     | - Mother’s income        |           |             |                       |       |                        |
|                   |     | - Sports activity        |           |             |                       |       |                        |
|                   |     | - Living environment     |           |             |                       |       |                        |
|                   |     | - Is he/she working      |           |             |                       |       |                        |
|                   |     | as part-time student     |           |             |                       |       |                        |
|                   |     | - Resources for the      |           |             |                       |       |                        |
|                   |     | programme                |           |             |                       |       |                        |
|                   |     | - Friends circle         |           |             |                       |       |                        |
|                   |     | - Student                |           |             |                       |       |                        |
|                   |     | Demographic              |           |             |                       |       |                        |
|                   |     | - Scholarship            |           |             |                       |       |                        |
|                   |     | - Behavior features      |           |             |                       |       |                        |
|                   |     | - Gender                 |           |             |                       |       |                        |
|                   |     | - Student information    |           |             |                       |       |                        |
|                   |     | - Education information  |           |             |                       |       |                        |
| 2017              | [1]  |                           | Naïve Bayes | Accuracy=96.8%  | Clustering Classification | RapidMiner | -Cleaning |
|                   |     |                           | Decision tree |                |                       |       |                        |
|                   |     |                           | Neural Networks |              |                       |       |                        |
|                   |     |                           | Random Forest |                |                       |       |                        |
| 2017              | [37] |                           | SVM k-Means | Accuracy=96.7%  | Clustering Classification | -       | -Cleaning |
|                   |     |                           |             |                |                       |       |                        |
| 2017              | [39] |                           |             |                |                       |       |                        |
|                   |     |                           |             |                |                       |       |                        |
| 2018              | [35] |                           | Deep Neural Network | Accuracy= 84.3% | Classification | -MS Excel | -Cleaning (one-hot-encoder) |
|                   |     | - Demographic            |             |                |                       |       |                        |
|                   |     | - Scholarship            |             |                |                       |       |                        |
|                   |     | - Behavior features      |             |                |                       |       |                        |
|                   |     | - Grade                  |             |                |                       |       |                        |
|                   |     | - Student information    |             |                |                       |       |                        |
|                   |     | - Education information  |             |                |                       |       |                        |
|                   |     | - Students’ behaviors    |             |                |                       |       |                        |
|                   |     | - Family behavior        |             |                |                       |       |                        |
| 2019              | [5]  |                           | - Ensemble Methods | Accuracy=98.5% Precision=98.5% Recall=98.5% | Clustering Classification | MS Excel | -Cleaning |

Algorithm 1 SMOTE Algorithm

Input: N: amount of instances of minority groups; n number of SMOTE (in %); k: represent the number of nearest neighbors; minority data D = x_i ∈ X; with i = 1; 2; . . . . . . N.

Output: D': synthetic data from D;

for i = 1 to N do

Find the k nearest neighbors of x_i;

while N ≠ 0 do

Choose one of the k nearest neighbors of x_i

Choose a random number α ∈ [0,1]

x = x_i + α (x - x_i)

Append x to D'

end while

end for = 0

student performance that affect the score that non technical users can interpret immediately [10].

B. DATA PROCESSING

In this step, preprocessing for the purpose of balancing data was conducted using SMOTE. SMOTE is a technique that involves creating a new sample for the minority class. In this method, a new sample is built from a sample x_i of the minority class by finding the k closest neighbors to x_i from the minority class. Then, a neighbor is randomly picked, and a synthetic sample that groups x_i and the selected neighbor is finally created [48]–[50]. Algorithm 1 clearly depicts this operation. Several other SMOTE techniques have emerged from this strategy, e.g., mostly weighted SMOTE [51], SMOTE-Boost [52], and borderline SMOTE [53]; these techniques were not the subject of this study but can be the subject of future research.

C. MODEL CONSTRUCTION

In this section, we apply supervised classification algorithms to train models to learn the mapping function of the input features to the output labels. The testing set was used to test the prediction performance of the models trained by different algorithms. To improve the precision, we used ten-folds cross validation to train and evaluate our models. We experimented four classification algorithms described below:

1) K−NEIGHBOR CLASSIFIER (KNC)

The classification of a new sample is given by the common classification of the K neighbors that are closest to that sample
in our dataset. With $K$ a positive integer denoted along with the new sample.

2) DECISION TREES (DT)
Decision trees are rule-based and tree-inspired structure algorithms that represent features like a tree [54]. Each node represents a feature, each connection between nodes represents a decision rule, and the output is represented by the leaf nodes. DTs are also seen as a flow chart, where the flow starts at the root node and ends with the predictions made on the leaves. It is a decision support tool, which is also seen as a tree-like diagram to display the predictions that result from a sequence of feature-based divisions [55].

3) RANDOM FOREST (RF)
The random forest classifier [44] (also referred to as a forest of decision trees) is a classification algorithm that trim the risk of overfitting by including randomness. Such trimming is performed by combining multiple decision trees, creating samples without replacements, and dividing nodes according to a random split using a random subset of features. Let a training dataset $L = \{(x_1,y_1),\ldots,(x_n,y_n)\}$, where the sample $x_i = \{f_1,f_2,\ldots,f_n\}$ is a $D$-dimensional vector with $f_i$ as the feature and $j \in \{1,2,\ldots,D\}$. ET creates $M$ independent decision trees. For every decision tree $S_p$, indicates the subset of training dataset $L$ at child node $p$; afterward, for every node $p$, the ET algorithm selects the best split with regard to $S_p$. The ET algorithm is clearly described below:

Algorithm 2 Extra Tree

Input: Training subset $S_p = \{s_1,s_2,\ldots,s_{Q_p}\}$, where the sample $s_i = \{f_1,f_2,\ldots,f_n\}$ is a $D$-dimensional vector, the number of attributes to select randomly, $K$, and the minimum number of samples required to split a node, $n\text{min}$. if $Q_p < n\text{min}$ then Stop splitting and define the node as a leaf node.

else

Select a random subgroup of end if

for every feature $k$ in the subgroup do

identify $f_k^\text{max}$ and $f_k^\text{min}$ as the maximal and minimal values for the feature $k$ in subset $S_p$

Obtain a random cut-point $f_k^c$, uniformly in the range $[f_k^\text{min},f_k^\text{max}]$.

Set $f_k^c$, as a candidate split.

end for

Select a split $[f_a < f_a^c]$ such that $\text{score}(f_a^c) = \min_{k=1,\ldots,k} \text{score}(f_a^c)$.

Output:

best split $[f_a < f_a^c]$ at the child node $p = 0$.

4) BAGGING CLASSIFIER
Bagging classifiers, also known as bootstrap aggregation, was introduced in 1996 by Breiman in [43] to fix the problems of CART instability. It is based on majority of the votes of learner classifiers for classification or their average for regression.

5) EXTRA TREE CLASSIFIER
Extra Tree classifiers, also known as extremely random trees [56], is a classification algorithm that reduces the risk of over learning from the data by including randomness. Such reduction occurs by combining multiple decision trees, creating samples without replacements, and dividing nodes according to a random split using a random subset of features.
6) XGBoost (XGB)
The extreme gradient boosting (XGBoost) algorithm is similar to the gradient boosting algorithm; however, it is more efficient and faster because it is composed of a linear model and a tree model, in addition to its ability to perform parallel calculations on a single machine.

The trees constructed in a gradient boosting algorithm are constructed in series so that a gradient descent step can be performed to minimize a loss function. Unlike RF, the XGBoost algorithm builds the tree itself in a parallel fashion. XGBoost algorithm procedure.

Essentially, the information contained in each column can be performed to minimize a loss function. Unlike RF, the extreme gradient boosting (XGBoost) algorithm is similar to the gradient boosting algorithm; however, it is more efficient and faster because it is composed of a linear model and a tree model, in addition to its ability to perform parallel calculations on a single machine.

The trees constructed in a gradient boosting algorithm are constructed in series so that a gradient descent step can be performed to minimize a loss function. Unlike RF, the XGBoost algorithm builds the tree itself in a parallel fashion. XGBoost algorithm procedure.

In accordance with [57], at the $t^{th}$ iteration, the objective function of XGBoost can be represented as

$$
\Theta(t) = \Phi(t) + \Omega(t)
$$

$$
\Theta(t) = \sum_{i=1}^{n} \Phi(y_i, \hat{y}_i) + \sum_{k=1}^{t} \Omega(f_k)
$$

where $n$ represents the $n^{th}$ prediction, and $\hat{y}_i^{(t)}$ can be represented as

$$
\hat{y}_i^{(t)} = \sum_{j=1}^{m} f_m(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
$$

The regularization term $\Omega(f_k)$ is considered by completing the convex differentiable loss function, and this term is defined by

$$
\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \sum_{i=1}^{m} \omega_i^2
$$

The term $\Omega$ is interpreted as a combination of ridge regularization of coefficient $\lambda$ and Lasso penalization of coefficient $\gamma$.

![FIGURE 3. XGBoost algorithm procedure.](image)

D. MODEL EVALUATION
The selection of performance metrics is crucial for evaluating a model [58]. When it comes to a classification problem as in our case, the frequently used metric is the accuracy defined by Equation 4. This formula, as those of other metrics, hail from the confusion matrix given by Table 3. In our case, $TP$ stands for true positive, which implies that the set of instances for which the prediction of the model is correct; $TN$ refers to true negative, which is the set of instances correctly predicted; $FP$ is false positive, representing the number of instances wrongly predicted; and $FN$ is false negative, giving the number of instances wrongly predicted.

$$
\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}
$$

$$
\text{Precision} = \frac{TP}{TP + FP}
$$

$$
\text{Recall} = \frac{TP}{TP + FN}
$$

However, this metric is limited and does not always represent reality when data are unbalanced because it favors the majority class. Thus, in addition to accuracy, specific additional metrics allow to evaluation to adapt to unbalanced data by considering the minority and majority classes. Among these metrics we used.

| Predicted positive | Predicted negative |
|-------------------|--------------------|
| Actual positive   | $TP$               | $FN$               |
| Actual negative   | $FP$               | $TN$               |

E. MODEL INTERPRETATION
In this part, we examine one approach, particularly game theory, to interpret the ensemble methods used in predicting student performance in Jordan. In this part, we introduce the reasons and key indicators for explaining student prediction models using game theory. We also define Shapley values and how they help improve prediction accuracy and model transparency [9]. SHAP values refer to Shapley values, a game theory jargon [59]. These values consist mainly of two components: game and players, where “game” represents the outcome of the predictive model and “players” represent the features of the model. Shapley measures the contribution value each player makes to the game. By applying these components to our case, the SHAP value measures how much the feature is contributing to the output of the model. The contribution of an individual player is determined by considering all possible coalitions of players, that is, all $i$ feature combinations possible ($i$ goes from 0 to $n$, where $n$ is a total number of features available). The order of adding the feature to the model is important and influences the prediction. Lloyd Shaply proposed this approach in 1953 (hence the name “Shapley values” for phi values measured in this
manner). Provided a prediction \( p \) (this is the prediction by the complex model), the Shapley value for a specific function \( i \) (out of \( n \) total features, and \( S \) is a subset of \( n \)) is [9]:

\[
\phi_i(p) = \sum_{S \subseteq \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [p(S \cup \{i\}) - p(S)]
\]  

(7)

This equation helps calculate the importance or the influence of a feature by calculating the difference of model prediction with and without feature \( i \). The shift in model estimation is essentially the feature’s consequence. In summary, SHAP values are used whenever we have to deal with an input output model and we want to understand the decision made by this model, that is, why the model suggested that a student was more or less likely to pass. This predictive model answers the “how much” question, and the SHAP value identifies the reason why [59]. Thus, we used SHAP values to interpret the model. Improved interpretability leads to improved acceptance [9].

Robust machine learning algorithms can usually make accurate predictions, but their opaque nature hampers their adoption. The interpretability of the model corresponds to the tag or the label on the pill bottle. The label needs to be transparent for easy adoption. Shapley values and the SHAP library are a valuable toolkit for uncovering the golden nuggets that a machine learning algorithm has discovered. SHAP concerns the local interpretability of a predictive model. In particular, by considering the effects of features on individual level instead of all individuals in the entire dataset (and then summing the results), the interactions of features can also be revealed, making it possible to gain more impressive knowledge than with techniques for global feature importance [60].

As a visualization technique, SHAP values provide a view of the prediction model functionality and increases the transparency of the model. Specifically, the SHAP force visualization provides a clear view of student performance features that affect the score that non technical users can interpret immediately [10].

IV. EXPERIMENTS AND RESULT ANALYSIS

This section examines the settings and protocols of the experiments. The results obtained are then analyzed and discussed. First, we begin by laying out a complete description of the variables used in predicting the performance of students afterward, we focused on model implementation and interpretation. We define the experimental procedure used to test and validate the proposed models for predicting student performance. We evaluate the results based on the training and cross-validation score of different machine learning methods. Then, we record the accuracy, precision, and recall values for model comparison. Here are the evaluation goals:

- Performance: How performance improve by using ensemble methods in comparison with previous studies that used the Jordan dataset for student performance prediction.

- Interpretability: Whether the SHAP value can be used to explain the inner logic of the ensemble methods. The predictor model supplies the accuracy value, and the interpreter depicts the effect of negative and positive features.

A. DATA DESCRIPTION

This part describes the open-source data retrieved from the Kaggle repository 1 in connection with the study made by [12]. These data are leveraged to create a scalable student performance model. The educational dataset used in this work consists of 480 student observations with 17 features. The full description of the variables used to predict a student’s performance in the Jordanian school is given in Table 4. These variables consist of personal features, such as gender and place of origin; educational background features, such as class participation, used digital resources, and student absence days; institutional features, such as education level and grades; and social features, such as parents in charge of students, parents who responded to the survey, and parents’ review of the school. Jordanian datasets are used with the same variables chosen by the previous study to build traditional and ensemble predictive models. By running the different models, the model performance of the student performance system can be empirically assessed and compared with previous working models. All codes, records, and full documentation on how to restore the results can be found available for public in the git hub repository.

B. DESCRIPTION OF EXPERIMENTAL PROTOCOL

This section aims to explain the experimental procedure for testing and evaluating the performance of the proposed models for student performance prediction. The implementations were conducted on a Windows framework using Python 3.7 and primarily the sklearn library. The machine in use is an “HP” model with the following configuration: 16 GB of RAM, an Intel Core i7 CPU, and NVIDIA Geforce 930M graphics card.

C. RESULT ANALYSIS

In this section, we used a tenfold cross-validation approach to train our models. Considering that validation is a critical phase in building practical predictive models, we mainly used two groups of machine learning algorithms to build our student performance prediction model: conventional methods (decision tree and \( K \)-neighbor classifiers) and ensemble methods. Here, we applied two frameworks (XGBoost and extra trees) to train the decision tree, and we stored the basic performance metrics: accuracy, precision, and recall. Table 5 shows the summary of the result of the validation and test sets. We found that the ensemble methods, i.e., ET and XGBoost, always obtain the best result. Moreover, we observed that the traditional methods always yield the lowest result through validation and test. The boosting and bagging method beats

---

1. https://www.kaggle.com/aljarah/xAPI-Edu-Data
TABLE 4. Description of the variables used in predicting the performance of student.

| N  | Feature                  | Description                                                                 | Feature Type |
|----|--------------------------|-----------------------------------------------------------------------------|--------------|
| 1  | Nationality              | Place of origin                                                             | Categorical  |
| 2  | Gender                   | The student gender ('Male' or 'Female').                                    | Categorical  |
| 3  | Place of birth           | Birth place of the student                                                  | Categorical  |
| 4  | Relation                 | Parent in charge of the student ('Mom' or 'Father').                         | Categorical  |
| 5  | Stage ID                 | Educational level student belongs ('Lowerlevel', 'MiddleSchool', 'HighSchool'). | Categorical  |
| 6  | Grade ID                 | Grade levels student belongs as (from G-01 to G-12).                        | Categorical  |
| 7  | Section ID               | Classroom student belongs as (A, B, C).                                     | Categorical  |
| 8  | Semester                 | School year semester as (First or Second).                                  | Categorical  |
| 9  | Topic                    | Course topic as (Math, English, IT, Arabic, Science, Quran).                 | Categorical  |
| 10 | Student absence days     | Student absence days (< 7, > 7).                                            | Categorical  |
| 11 | Parent answering survey  | Parent is responding to the surveys or not.                                 | Categorical  |
| 12 | Parent school satisfaction| The degree of satisfaction of the parent (Good, Bad).                      | Categorical  |
| 13 | Discussion               | how many times the student participate on discussion groups.                | Numerical     |
| 14 | Announcements view       | How many times the student checks the new announcements.                    | Numerical     |
| 15 | Visited resources        | How many times the student visits a course content.                         | Numerical     |
| 16 | Raised hands             | How many times the student raises his/her hand on classroom.                | Numerical     |

other traditional stand-alone methods, that is, boosting methods improve the accuracy from 76.7% to 99.47%, indicating that the number of students accurately classified has increased from 367 to 477 out of a total of 480 students. The recall results increased from 77.097% to 99.47%, indicating that out of the student pool of unclassified and correctly classified students, 477 students were ranked correctly. The precision scores also climbed from 77.37% to 99.37%, meaning that out of the total of 480 students, 476 students were classified correctly, and four were classified incorrectly. The model performance of the boosting and bagging classifier is not significantly different. As shown Table 5, the prediction models achieved accuracy and precision over 98%. These results were obtained via the use of strategies and techniques, such as SMOTE, hyperparameter optimization, and cross-validation process, which demonstrate the dependability of the novel model. These values show that we can predict student performance and enhance the prediction by using ensemble methods.

1) CONFUSION MATRIX
The overall accuracy of the boosting and bagging method on the test and validation tests was 0.998 (accurate predictions/total or true positives/total). Nevertheless, the confusion matrix provides additional insight into accuracy by class and insight for precision and recall efficiency. An insight we can obtain from the matrix is the ensemble models were accurate at classifying low and average performers, and the accuracy was TP/total = 1.0. However, accuracy for high performers was slightly low (0.99). If for any rational motive, successful classification of a student class was particularly coveted to the case on hand, then the confusion matrix helps outline differences between the classes and identify which classifier performs better. Boosting and bagging classifiers beat other traditional individual classifiers. For instance, the accuracy by class for boosting classifier increased from 77 to 100 for the average performers, from 88 to 100 for low performers, and from 75 to 99 for high performers.

2) MODEL INTERPRETATIONS
Ensemble methods, such as ET and XGBoost, can attain good accuracy rates; however, they are difficult for people to comprehend because these models are generally viewed as a black box. However, after we introduce the SHAP values, these models can be explained in a steady manner. As a great way to increase the transparency of the ensemble models, the SHAP framework provides global and local interpretation. Global interpretation uses feature importance plot and summary plot as a way of visualization. By contrast, local interpretation uses force plot to show individual SHAP values for a single observation and try to visualize feature attributions at the observation level to provide more accurate explanations and understanding of model decisions.

The idea behind SHAP feature importance is simple, that is, feature importance values are obtained by calculating the average of the SHAP values of each feature. A feature importance plot ranks the most relevant features in degressive order.

Based on Figure 5, student absences have a higher effect than other features, indicating that the change of this feature can have a more noticeable influence than others.

If student absences are higher, then student performance would be reduced accordingly.

However, this feature has a higher negative effect specifically for low and average performers than higher performers.

If we look at the visited resource, we can find that the more resources visited, the more likely student performance improves. A similar result is found for class participation;
the more a student participates, the higher the student performance. These later features have a higher positive impact for all classes of performers. By contrast, features, such as stage ID, section ID, and semester, do not have a huge effect on student performance for all class of performers, indicating that the change of these features does not have a noticeable influence on model prediction. As far as the summary plot is concerned, the y-axis represents the feature values, the x-axis depicts the Shapley value, and the color shows the degree of the effect (red means high, whereas blue means low); the higher the SHAP value, the larger the effect. At a global level, the summary plot provides feature importance with feature effects. The features are ranked in accordance with their predictive power, and the graphs below show the features that affected the model the most and how changes in their values affect the model's prediction. How much each feature contributes, either positively or negatively, to the model output. And which variables strongly correlate with the target variable, making SHAP a tool with great benefit in variables selection.

This plot shows that student absence days, a parent in charge, raise hands Features are the most important model

| Algorithm                | Accuracy Validation score | Accuracy Test score | Precision Validation score | Precision Test score | Recall Validation score | Recall Test score |
|--------------------------|---------------------------|---------------------|---------------------------|----------------------|------------------------|------------------|
| Random Forest (RF)       | 0.9940                    | 0.9880              | 0.9930                    | 0.9879               | 0.9943                 | 0.9892           |
| Bagging Classifier (BC)  | 0.9920                    | 0.9940              | 0.9936                    | 0.9951               | 0.9940                 | 0.9942           |
| Extra Trees (ET)         | 0.9945                    | 0.9940              | 0.9938                    | 0.9929               | 0.9946                 | 0.9939           |
| XGBoost (XGB)            | 0.9935                    | 0.9940              | 0.9936                    | 0.9950               | 0.9943                 | 0.9939           |
| Decision Tree (DT)       | 0.9520                    | 0.9360              | 0.9533                    | 0.9402               | 0.9585                 | 0.9364           |
| K-Neighbors Classifier (KNC) | 0.7670                | 0.7900              | 0.7735                    | 0.7939               | 0.7709                 | 0.8003           |
| Naive Bayes Classifier (NB) | 0.7605                | 0.7360              | 0.7665                    | 0.7507               | 0.7649                 | 0.7549           |
| Neural Network (NN)      | 0.9450                    | 0.9480              | 0.9542                    | 0.9518               | 0.9445                 | 0.9528           |
features because the values of these features (high and low) are strongly correlated with high and low SHAP values. By contrast, other model features, such as stage ID, section ID, and semester, are less important as their corresponding SHAP values are closer to zero and thus have less effect on the model. By taking a look at the first rows of the summary plot, we can tell that a high level of attributes, such as student absence days, raise hands, and visited resources, have a high and positive effect on student performance. The high comes from the red color, and positive is shown on the x-axis. When the number of student absence days is under seven, the feature has a positive effect on student performance. Class participation also has a positive effect on student performance. Furthermore, the digital resources used feature is positively related to student performance. In terms of which parent is in charge, a negative effect on student performance is found when the dad is in charge (negative SHAP value). This finding is consistent with human intuition. This plot is a great tool to obtain an improved understanding of how certain features affect the model decision. To obtain a deeper understanding of our model, we use other SHAP packages, such as force plots, which provide us feature contributions for specific observation.

**FIGURE 6. Summary plots.**

In this study, SHAP value of a feature for a single record prediction is acquired by the contribution of that feature towards the prediction. The force plot in Figure 7 highlights the features responsible for predicting student performance and the features driving the model output from the baseline value to the actual output. Features that have more predictive control are shown in red, whereas features that have lower predictive control are shown in blue. The output value is the prediction with features, whereas the base value is the value that would be predicted without any features, that is, the mean prediction. This chart can inform decision-makers in education about the key features responsible for student performance at the observation level. In the plot, each feature value is a force that either increases or decreases the prediction starting from the baseline. For instance, the baseline is 0.5937. The actual prediction is 0.98. The students' absence days and visited resource features can increase the prediction. By contrast, the parent in charge and announcements view features can decrease the final output prediction.

3) DISCUSSION

This paper describes the use of two types of predictive models that show an improvement in accuracy over the usual traditional methods. The result obtained show the following: The tested and cross-validated ensemble models improved the performance prediction of student performance. These prediction models attained an accuracy and precision of over 98%, outperforming the results in [12], which used the same dataset, as illustrated in Table 6. By examining two novel types of models related to student performance data, particularly XGBoost and ET and used them to boost the predictive accuracy of student performance models. And solving the problem of class imbalance, which implies strategies such as data oversampling for the less represented class, known as SMOTE, and the optimization of hyperparameters by choosing a set of optimal hyperparameters for the learning algorithm. And a tenfold cross-validation method students to improve the classification algorithms and avoid overfitting the model. The results obtained via the use of the aforementioned strategies and techniques demonstrate the dependability of the novel model. In addition, a novel technique is used to help determine which factors influence the score (i.e., SHAP values, which increment model transparency) [9]. We apply SHAP values on student data to explain how ensemble methods predict student performance. The SHAP value and the associated visualizations provide a view into the inner operations of the prediction models used and increase the transparency of the model. Specifically, the SHAP force visualization provides a clear view of student performance features that affect the score that non technical users can interpret immediately [10]. In particular, by considering the effect of features on the prediction of the model at the macro-level: on all individuals by considering the main features that contribute to the goal. For example, student absences have a greater effect than other features. And at the micro-level: by identifying which features were most important for an individual. Force plots highlight the features responsible for predicting student performance and features contributing to the prediction for the single record for single observation for an individual (e.g., in terms of student absences, the lower the number of absences, the more positive the effect on student performance; in terms of parent in charge, a negative effect is observed when the dad is in charge; in terms of digital resources used, the more resources, the more positive the correlation with student performance). The latter represents the first attempt to use SHAP values in the educational context. This contribution fills up the gap in current prediction model interpretation, particularly in the field of education. We considered that SHAP values are a powerful tool to include in the framework for predicting student performance.
because they solve important issues about the interpretability of more opaque models. Model interpretation is a critical step that could lead to the creation of new knowledge to improve our understanding of student performance. As a result, when the prediction and explanation are established through our experiment, educators can identify students at risk early on and provide appropriate admonitions in promising ways.

V. CONCLUSION AND FUTURE WORK
This paper presents our preliminary research and depicts the importance of tree-based machine learning algorithms and their application in predicting student performance. The aim of this work is to improve performance measure values by using ensemble methods and comparing them with previous studies that used the Jordan dataset to investigate student performance prediction. We also apply the SHAP value to explain the inner logic of the ensemble methods. The result shows that our models have better accuracy than the traditional models. The prediction models achieve an accuracy of over 98%, which outperforms the result obtained in the study made by [12] working on the same dataset. These results are obtained via the use of strategies and techniques, such as SMOTE, hyperparameter optimization, and cross-validation process, which demonstrate the dependability of the novel model. Moreover, the SHAP value and the associated visualizations provide a view into the inner operations of the prediction models used and increase the transparency of the model. Specifically, the SHAP force visualization provides a clear view of the features of student performance that affects the score that non-technical users can interpret immediately. As a result, when the prediction and explanation are established through our experiment, educators can identify students at risk early on and provide appropriate admonitions in promising ways.

A. LIMITATIONS AND FUTURE WORK
This study has certain limitations that must be highlighted. The study relies on publicly available datasets and not on a dedicated student dataset. Moreover, the dataset was small, having only under a thousand records. Research with more data may offer more conclusive insights. Most researchers in EDM are currently hesitant to share their research dataset for two main reasons: The first is for privacy, ethics, and legality; the second is for dataset acquisition, which is a time-consuming, labor-intensive, and expensive task. We recommend that machine learning researchers disclose more educational datasets based on a blend of privacy shield, economic impact, and academic implications. This study used offline data although an ever-increasing number of online data remains unused so we can train the model to predict online student performance in real time. A distinctive educational dataset that can be tested by our models. Right off the bat, on the off chance that we can be given a large dataset, we can utilize the most recent big data technology to build a new model and validate the outcomes. Furthermore we can obtain additional data and attempt deep learning methods to improve model performance by using additional features, such as examining how the use of social media would influence the performance of students. Moreover, additional experimentations could be conducted by using other machine learning techniques, such as clustering. This research used decision trees, KNC and ensemble methods, such as XGBoost and ET, for classification. Other methods, such as clustering and deep neural networks, can be used for classification or regression problems to have an improved perception of the importance of method selection. Another area that can be improved is the process of feature engineering. Given limited data, the amount of feature engineering that can be made is likewise limited. For model interpretability, the Shapley value method needs to pass over “all possible combinations” of the features. When the number of features is large, the number of combinations is extremely huge, yielding in a large amount of Shapley value calculation and immense time complexity, resulting in computation infeasibility. In the future, we will consider utilizing other existing datasets with more features and will extend the scope of study to include other African countries. We are currently exploring student data obtained from the Kaggle repository related to Nigerian schools, and we look forward to using dynamic selection algorithms to improve model performance and extract knowledge that may be useful to school administrators. Our focus is to extend the scope of EDM and provide noteworthy insights and models to improve education across...
Africa. Another area of research for model interpretability is to improve the SHAP library to calculate Shapley values quicker than if a model prediction had to be calculated for each conceivable combination of features.

REFERENCES

[1] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, “Analyzing undergraduate students’ performance using educational data mining,” Comput. Educ., vol. 113, pp. 177–194, Oct. 2017.

[2] M. Goyal and R. Vohra, “Applications of data mining in higher education,” Int. J. Comput. Sci. Issues, vol. 9, no. 2, p. 113, 2012.

[3] D. Kučak, V. Juričić, and G. Dambić, “Machine learning in education—A survey of current research trends,” Ann. DAAAM Proc., vol. 29, pp. 406–410, 2018.

[4] A. Saygeh, Y. Hu, I. Zografopoulos, X. Liu, R. G. Dutta, Y. Jin, and C. Konstantinou, “Survey of machine learning methods for detecting false data injection attacks in power systems,” IET Smart Grid, vol. 3, no. 5, pp. 581–595, Oct. 2020. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/iet-stg.2020.0015

[5] A. Almasri, E. Celebi, and R. S. Alkhawaldeh, “EMT: Ensemble meta-based tree model for predicting Student performance,” Sci. Program., vol. 2019, pp. 1–13, Feb. 2019.

[6] O. Adejo and T. Connolly, “An integrated system framework for predicting students’ academic performance in higher educational institutions,” Int. J. Comput. Sci. Inf. Technol., vol. 9, no. 3, pp. 149–157, Jun. 2017.

[7] M. Yalcintas and U. A. Ozturk, “An energy benchmarking model based on artificial neural network method utilizing U.S. Commercial buildings energy consumption survey (CBECS) database,” Int. J. Energy Res., vol. 31, no. 4, pp. 412–421, 2007.

[8] Y. Wei, X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, and X. Zhao, “A review of data-driven approaches for prediction and classification of building energy consumption,” Renew. Sustain. Energy Rev., vol. 82, pp. 1027–1047, Feb. 2018.

[9] Explain Your Model With the SHAP Values. Accessed: Sep. 14, 2019. [Online]. Available: https://towardsdatascience.com/explain-any-models-with-the-shap-values-use-the-kernel-explainer-79de496487a

[10] P. Arjunan, K. Poola, and C. Miller, “EnergyStar+++: Towards more accurate and explainatory building energy benchmarking,” Appl. Energy, vol. 276, Oct. 2020, Art. no. 115413.

[11] C. Konstantinou, “Cyber-physical systems security education through hands-on lab exercises,” IEEE Des. Test. Comput., vol. 37, no. 6, pp. 47–55, Dec. 2020.

[12] E. A. Amrieh, T. Hamtini, and I. Aljarah, “Mining educational data to predict student’s academic performance using ensemble methods,” Int. J. Database Theory Appl., vol. 9, no. 8, pp. 119–136, 2016.

[13] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 4765–4774.

[14] M. Ramaswami and R. Bhaskaran, “A CHAID based performance prediction model in educational data mining,” 2010, vol. 6, no. 1, pp. 119–139, Aug. 1997.

[15] A. Tekin, “Early prediction of students’ grade point averages at graduation: A data mining approach,” Eurasion J. Educ. Res., vol. 14, no. 54, pp. 207–226, Feb. 2014.

[16] J. Bansode, “Mining educational data to predict students’ academic performance,” Int. J. Recent Innov. Trends Comput. Commun., vol. 4, no. 1, pp. 1–5, 2016.

[17] D. Kabakchieva, “Predicting student performance by using data mining methods for classification,” Cybern. Inf. Technol., vol. 13, no. 1, pp. 61–72, 2013.

[18] Z. Kovacic, “Early prediction of student success: Mining students’ enrollment data,” in Proc. Informing Sci. IT Educ. Conf. (InSITE), 2010.

[19] Z. Kovacic, “Predicting student success by mining enrolment data,” Res. Higher Educ., Tech. Rep. 15, 2012.

[20] R. B. Nandigam, A. P. N. Lam, and C. Konstantinou, “Survey of current research trends,” Ann. DAAAM Proc., vol. 9, no. 2, p. 113, 2012.

[21] A. Tekin, “Early prediction of students’ grade point averages at graduation: A data mining approach,” Eurasion J. Educ. Res., vol. 14, no. 54, pp. 207–226, Feb. 2014.

[22] J. Bansode, “Mining educational data to predict students’ academic performance,” Int. J. Recent Innov. Trends Comput. Commun., vol. 4, no. 1, pp. 1–5, 2016.

[23] D. Kabakchieva, “Predicting student performance by using data mining methods for classification,” Cybern. Inf. Technol., vol. 13, no. 1, pp. 61–72, 2013.

[24] Z. Kovacic, “Early prediction of student success: Mining students’ enrolment data,” in Proc. Informing Sci. IT Educ. Conf. (InSITE), 2010.

[25] Z. Kovacic, “Predicting student success by mining enrolment data,” Res. Higher Educ., Tech. Rep. 15, 2012.

[26] R. B. Nandigam, A. P. N. Lam, and C. Konstantinou, “Survey of current research trends,” Ann. DAAAM Proc., vol. 9, no. 2, p. 113, 2012.
EL ARBI ABDELLAOUI ALAOUI received the Ph.D. degree in computer science from the Faculty of Sciences and Technology—Errachidia, University of Moulay Ismail, Meknes, Morocco, in 2017. He is currently a Research Professor with the Ecole Normale Supérieure de Meknes, Moulay Ismail University. His main research interests include wireless networking, machine learning, DTN networks, game theory, the Internet of Things (IoT), and smart cities.

ANAND NAYAR received the Ph.D. degree in computer science from Desh Bhagat University, in 2017, with a focus on wireless sensor networks and smart intelligence. He is currently working with the Graduate School, Duy Tan University, Da Nang, Vietnam. He is also working in the areas of wireless sensor networks, the IoT, smart intelligence, cloud computing, artificial intelligence, blockchain, cyber security, network simulation, and wireless communications. He is a Certified Professional with more than 75 professional certificates from CISCO, Microsoft, Oracle, Google, Beingcert, EXIN, GAQM, and Cyberoam. He has published more than 450 research papers in various national and international conferences and international journals (Scopus/SCI/SCI/SSCI indexed) with high impact factor. He has authored/coauthored cum edited more than 30 books of computer science. He has five Australian patents to his credit in the area of wireless communications, artificial intelligence, the IoT, and image processing. He is also a member of more than 50 associations as a senior member and a life member, and acting as an ACM distinguished speaker. He is associated with more than 500 international conferences as a program committee member/the chair/an advisory board member/a review board member. He was awarded more than 30 awards for teaching and research, including the Young Scientist, the Best Scientist, the Young Researcher Award, the Outstanding Researcher Award, and the Excellence in Teaching. He is also acting as an Associate Editor for Wireless Networks (Springer), IET Quantum Communication, IET Wireless Sensor Systems, IET Networks, IJDSST, IJISP, and IJCINI. He is also acting as the Editor-in-Chief for the USA journal International Journal of Smart Vehicles and Smart Transportation (IJSVST) (IGI-Global).

SAID AGOUJIL received the M.S. and Ph.D. degrees in mathematics from the Faculty of Sciences and Technology of Marrakech (FSTM), Morocco, in 2004 and 2008, respectively. He is currently a Professor with the Department of Computer Science, Faculty of Sciences and Technology, Moulay Ismail University, Errachidia, Morocco. His current research interests include numerical analysis, wireless networks, linear algebra, and speech coding.

HAYAT SAHLAOUI received the B.S. degree in web programming from the Faculty of Science, University Moulay Ismail, Meknes, Morocco, in 2009, and the master’s degree in mathematics education and technology from the Ecole Normale Superieure Tetouan, in 2013. She is currently pursuing the Ph.D. degree in machine learning with the Faculty of Science and Technique Errachidia, University Moulay Ismail. Since 2011, she has been a distinguished educational professional in various institutions of Meknes’ district Morocco, with ten years of teaching experience with an unparalleled ability to explain complicated mathematical concepts in an easily understandable manner. She has the talent for employing unique teaching strategies to effectively engage all students and foster a fun and fascinating learning environment. Her research interests include education strategies, e-learning, web application development, learning analytics, data mining, and machine learning in education.

MUSTAFA MUSA JABER received the Ph.D. degree from the Technical University of Malaysia. He holds a Postdoctoral position at University Tun hussien Onn Malaysia. He is currently the Director of the Research Center with the Dijlah University College, Baghdad, Iraq. His research interests include telemedicine, machine learning, and human factor.

[47] (2020). Top 10 Reasons Why Python is so Popular With Developers in 2020. [Online]. Available: https://www.upgrad.com/blog/reasons-why-python-popular-with-developers/

[48] N. V. Chawla, K. K. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, no. 1, pp. 321–357, 2002.

[49] A. Roy, R. M. O. Cruz, R. Sabourin, and G. D. C. Cavalcanti, “A study on combining dynamic selection and data preprocessing for imbalance learning,” Neurocomputing, vol. 286, pp. 179–192, Apr. 2018.

[50] H. Faris, R. Abukhurma, W. Almanaseer, M. Saadeh, A. M. Mora, P. A. Castillo, and I. Aljarah, “Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: A case from the Spanish market,” Prog. Artif. Intell., vol. 9, no. 1, pp. 31–53, Mar. 2020.

[51] S. Barua, M. M. Islam, X. Yao, and K. Murase, “WMOTHE-majority weighted minority oversampling technique for imbalanced data set learning,” IEEE Trans. Knowl. Data Eng., vol. 26, no. 2, pp. 405–425, Feb. 2012.

[52] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, “SMOTEBoost: Improving prediction of the minority class in boosting,” in Proc. Eur. Conf. Princ. Data Mining Knowl. Discovery, Berlin, Germany: Springer, 2003, pp. 107–119.

[53] H. Han, W.-Y. Wang, and B.-H. Mao, “Borderline-SMOTE: A new oversampling method in imbalanced data sets learning,” in Proc. Int. Conf. Intell. Comput. Berlin, Germany: Springer, 2005, pp. 878–887.

[54] V. L. Miguéis, A. Freitas, P. J. V. Garcia, and A. Silva, “Early segmentation of students according to their academic performance: A predictive modelling approach,” Decis. Support Syst., vol. 115, pp. 36–51, Nov. 2018.

[55] J. R. Quinlan, “Induction of decision trees,” Mach. Learn., vol. 1, no. 1, pp. 81–106, 1986.

[56] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Aug. 2016, pp. 785–794.

[57] I. Lkoykentrzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, “Dropout prediction in e-learning courses through the combined use of machine learning techniques,” Comput. Educ., vol. 53, no. 3, pp. 950–965, Nov. 2009.

[58] S. Mazzanti. (Jan. 4, 2020). I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and P. A. Castillo, and I. Aljarah, “Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: A case from the Spanish market,” Prog. Artif. Intell., vol. 9, no. 1, pp. 31–53, Mar. 2020.

[59] S. Mazzanti. (Jan. 4, 2020). I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and P. A. Castillo, and I. Aljarah, “Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: A case from the Spanish market,” Prog. Artif. Intell., vol. 9, no. 1, pp. 31–53, Mar. 2020.

[60] S. Mazzanti. (Jan. 4, 2020). I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and P. A. Castillo, and I. Aljarah, “Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: A case from the Spanish market,” Prog. Artif. Intell., vol. 9, no. 1, pp. 31–53, Mar. 2020.

[61] S. Mazzanti. (Jan. 4, 2020). I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and P. A. Castillo, and I. Aljarah, “Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: A case from the Spanish market,” Prog. Artif. Intell., vol. 9, no. 1, pp. 31–53, Mar. 2020.

[62] S. Mazzanti. (Jan. 4, 2020). I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and P. A. Castillo, and I. Aljarah, “Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: A case from the Spanish market,” Prog. Artif. Intell., vol. 9, no. 1, pp. 31–53, Mar. 2020.

[63] S. Mazzanti. (Jan. 4, 2020). I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and P. A. Castillo, and I. Aljarah, “Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: A case from the Spanish market,” Prog. Artif. Intell., vol. 9, no. 1, pp. 31–53, Mar. 2020.