IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 23, 2020, accepted April 4, 2020, date of publication April 9, 2020, date of current version April 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2986809

Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques

RAMIN GHORBANI AND ROUZBEH GHOUSI
School of Industrial Engineering, Iran University of Science and Technology, Tehran 16846-13114, Iran
Corresponding author: Rouzbeh Ghousi (ghousi@iust.ac.ir)

ABSTRACT In today's world, due to the advancement of technology, predicting the students' performance is among the most beneficial and essential research topics. Data Mining is extremely helpful in the field of education, especially for analyzing students' performance. It is a fact that predicting the students' performance has become a severe challenge because of the imbalanced datasets in this field, and there is not any comparison among different resampling methods. This paper attempts to compare various resampling techniques such as Borderline SMOTE, Random Over Sampler, SMOTE, SMOTE-ENN, SVM-SMOTE, and SMOTE-Tomek to handle the imbalanced data problem while predicting students' performance using two different datasets. Moreover, the difference between multiclass and binary classification, and structures of the features are examined. To be able to check the performance of the resampling methods better in solving the imbalanced problem, this paper uses various machine learning classifiers including Random Forest, K-Nearest-Neighbor, Artificial Neural Network, XG-boost, Support Vector Machine (Radial Basis Function), Decision Tree, Logistic Regression, and Naïve Bayes. Furthermore, the Random hold-out and Shuffle 5-fold cross-validation methods are used as model validation techniques. The achieved results using different evaluation metrics indicate that fewer numbers of classes and nominal features will lead models to better performance. Also, classifiers do not perform well with imbalanced data, so solving this problem is necessary. The performance of classifiers is improved using balanced datasets. Additionally, the results of the Friedman test, which is a statistical significance test, confirm that the SVM-SMOTE is more efficient than the other resampling methods. Moreover, The Random Forest classifier has achieved the best result among all other models while using SVM-SMOTE as a resampling method.

INDEX TERMS Classification, data mining, educational data mining, imbalanced data problem, machine learning, resampling methods, statistical analysis.

I. INTRODUCTION

Recent advancement in several fields has led to a large amount of collected data [1]. Since analyzing the considerable amount of data to reach useful information is a tedious task for humankind, data mining techniques can be used to discover valuable and significant knowledge from the data [2]. It is well-known that universities are operating in a very complex and highly competitive environment [3], [4]. The main challenge for universities is to examine their performance profoundly, identify their uniqueness, and build tactics for further development and future achievements [5]. The educational system understands the potential of using data mining to improve its performance dramatically.

Educational Data Mining (EDM) is the implementation of data mining methods for analyzing available data at educational institutions [6]. Although data mining leads to knowledge discovery, machine learning algorithms provide the needed tools for this purpose. The high accuracy prediction in students' performances is useful as it helps to identify the students with low academic achievements at the early stage of academics [7], [8]. Educational data mining helps educational organizations to extend their understanding of the learning process by analyzing the related educational data [9], [10]. In fact, the prediction of student academic performance is indispensable for student academic progression, and it is also

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Asaduzzaman.

challenging due to the influence of different factors affecting students' performance [11], [12]. In recent years, researchers have introduced new strategies for educational data mining.

There have been numerous researches in the education field. In 2008, [13] introduced an Artificial Neural Network (ANN) model using a sample of 1,407 students' profiles to predict their performance. The proposed algorithm was trained and tested by applying the hold-out method, which is one of the most popular cross-validation techniques. It should be noted out that there are other researches that have implemented the Artificial Neural Network algorithm as a predictive model. In 2015, [14] developed two different models of the Artificial Neural Network algorithm. The results of this research indicated that the Artificial Neural Network model could predict 95% of students' performance accurately, which shows the effectiveness of this model in prediction. Furthermore, [15] tested the Artificial Neural Network model with the overall accuracy result of 84.6%, which proves the potential of this model in predicting students' performance. It is apparent that other machine learning models have also been developed. [16] formed a Naive Bayes model using the 700 students' data to predict their performance. Also, [17] used the Decision Tree models with an overall correct classification percentage of 60.5%. In addition, this research indicated the essential features using feature importance method.

It is important to note that there are some research works that have introduced and compared different machine learning and data mining models with other models. In 2014, [18] applied various machine learning models to predict students' performance. The results show that the Decision Tree has obtained the best performance among other models. Also, [19] assessed the performance of different classifiers such as Logistic Regression, Support Vector Machine, Decision Tree, Artificial Neural Network, Naive Bayes, and K-Nearest Neighbor. Moreover, the feature selection method is used to increase the models' accuracy. Furthermore, [20] compared the performance of the Support Vector Machine, Logistic Regression, Naive Bayes, Random Forest, and XG-Boost data mining methods. Similarly, [21] studied the differences among the performance of Artificial Neural Network, XG-Boost, Random Forest, and Logistic Regression. The results of this research show that XG-Boost has demonstrated excellent predictive accuracy. It is significant to consider that all of these research works have used the hold-out method, which is the most straightforward cross-validation technique.

The k-fold cross-validation is a reliable cross-validation method which is not used as much as the hold-out method in the field of education data mining. In 2013, [22] implemented and compared the Decision Tree, Naive Bayes, and K-Nearest Neighbor models while using 10-fold cross-validation. Furthermore, some other researches, such as [23], [24] used the k-fold cross-validation to compare different data mining models in the purpose of predicting students' performance.

It is a well-established fact that it can be challenging to improve a model's predictive accuracy. Different factors have an impact on enhancing prediction accuracy. Using feature selection and handling imbalanced class distribution problem are among the essential factors. The class imbalance distribution is a common problem for educational data, which can extremely affect models' performance. Therefore, [25] developed the Decision Tree and Logistic Regression models to predict students' performance while handling the imbalanced class problem. [26] concentrated on developing different algorithms while using the feature importance method and SMOTE oversampling method as a way to solve the imbalanced data problem. Moreover, [27] compared random oversampling and SMOTE balancing methods along with four popular data mining models to assess the students' performance. Choosing a way to solve the imbalanced data problem can be challenging, and many resampling methods are available to handle the imbalanced data problem. However, there is not any research comparing these methods with each other.

A summarized list of research works on educational data mining and predicting students' performance is presented in Table 1.

In summary, due to the importance of imbalanced data problem, a lack of comprehensive comparison among the popular resampling methods as a way to handle this problem is evident. This paper tries to study the impact of the imbalanced data problem on the machine learning models' performance. It uses different resampling methods to solve the imbalanced data problem and compares these methods while using various machine learning classifiers to fill the gaps in the literature. The novel innovations and vital processes of this research as compared to similar research works include:

- Applying feature scaling to normalize the variety of independent data features.
- Implementing and comparing different resampling methods, namely Borderline SMOTE, Random Over Sampler, SMOTE, SMOTE-ENN, SVM-SMOTE, and SMOTE-Tomek.
- Applying different model validation methods, namely Random Hold-Out and Shuffle K-fold cross-validation methods, to perform the validation step.
- Comparing the performance of resampling methods using various machine learning models such as Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Naive Bayes, Artificial Neural Network, and Decision Tree, and XG-Boost.
- Measuring the performance of the implemented models using different evaluation measure methods such as Accuracy, Recall, Precision, and F1-Score.
- Showing the effect of the resampling methods on the classifiers' performance.
- Analyzing and examining the differences between resampling methods and indicating the best method among others using the Friedman test as a statistical significance test.

TABLE 1. Review of research works in the field of educational data mining and predicting students' performance.

Article	Machine Learning Model							Imbalance Data Problem							validation		Statistical Evaluation	Model Comparison
	Logistic Regression	K-Nearest-Neighbor	Naive Bayes	Support Vector Machine	Artificial Neural Network	Decision Tree	Random Forest	XG-Boost	Borderline SMOTE	Random Over Sampler	SMOTE	SVM-SMOTE	SMOTE-ENN	SMOTE-Tomek	Hold-Out	K-Fold Cross-validation		
[13]					✓										✓			✓
[23]			✓	✓	✓	✓	✓									✓		✓
[17]						✓									✓			
[22]		✓	✓			✓									✓	✓		✓
[27]	✓			✓	✓	✓				✓	✓					✓		✓
[25]	✓					✓				✓						✓		
[18]		✓	✓		✓	✓									✓			✓
[14]					✓										✓			✓
[15]					✓										✓			✓
[26]			✓		✓	✓					✓					✓		✓
[16]			✓												✓			✓
[19]	✓	✓	✓	✓	✓	✓									✓			✓
[20]	✓		✓	✓			✓	✓							✓			✓
[21]	✓				✓		✓	✓							✓			✓
[24]	✓		✓	✓	✓	✓										✓		✓
Present Work	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

• Investigating the difference between multiclass and binary classification and the importance of the features' structure.

This paper is prepared as follows: The next section explains the methodology of this paper and the information about the datasets and all the preprocessing operations, such as different solving methods of the imbalanced data problem. In section 3, the implemented predictive models are introduced. Section 4 describes the validation methods used to evaluate the generalization of statistical analysis results. In section 5, the employed evaluation measure methods are described. Section 6 presents the results and complete analysis to demonstrate the performance of the different resampling methods while using various machine learning classifiers. Finally, Section 7 exposes the conclusion and recommends some directions for future research.

II. MATERIAL & METHODS

This paper attempts to compare the different resampling methods of handling the imbalanced data problem to find the best approach and classifier while predicting the students' performance. Also, examining the difference between multiclass and binary classification and the importance of the features' structure are among the goals of this research. The steps of the applied methodology to achieve the goals of this paper are as follows:

1. Data Collection
2. Data Preprocessing
3. Handling Imbalanced dataset
4. Implementing Predictive Models
5. Analyzing the Results

A. DATASET INFORMATION

This research has used two different educational datasets from educational institutions of Iran and Portugal [23]. In the Iran dataset, all available information about postgraduate students collected and registered manually from Iran University of Science and Technology between 1992-93 and 2014-15 academic years. This dataset consists of a set of factors that can affect the students' performance. This dataset includes information on the 650 students with 19 different attributes. Also, in the Portugal dataset, all the information is related to student achievement in the secondary education of two Portuguese schools. This dataset includes information on the 394 students with 19 different attributes. The output variable in this study is the Final GPA. The information about the output attribute for both datasets is divided into four categories based on the grade point average of the students. These four categories are Poor, Medium, Good, and Excellent students, so this paper faces a multi-classification problem. Table 2 presents the main features of these datasets. Using these two datasets helps to better express the imbalanced data problem in all levels of educational fields, to have a better comparison among different resampling methods, and to gain more trustable results. Moreover, different structure of these datasets helps to have a more comprehensive analysis in the effect of the formation of a dataset.

B. DATA PREPROCESSING

One of the most significant steps in machine learning is data preprocessing. This step transforms the raw data into a proper and understandable format. In the real world, datasets contain many errors; therefore, this step can solve the errors, and the

TABLE 2. Main features of students' dataset of Iran University of Science and Technology.

Iran Dataset		Portugal Dataset	
Feature Name	**Type**	**Feature Name**	**Type**
Sex	Nominal	Sex	Nominal
Age	Numeric	Age	Numeric
Health Status	Ordinal	Travel Time to School	Numeric
Age on Entrance to the University	Numeric	Family Quality Life	Ordinal
Entrance Semester	Nominal	Going Out with Friends	Ordinal
Students' Scholarship Status	Nominal	Health Status	Ordinal
Remedial Courses	Nominal	Absences in Classes	Numeric
Failed Academic Terms	Nominal	Size of the Family	Numeric
Marital Status	Nominal	Free Time after School	Numeric
Children	Nominal	Mother's Job	Nominal
Residence Area (Capital or Other Cities)	Nominal	Father's Job	Nominal
Master's degree Period	Nominal	Extra Educational Support (Financial)	Nominal
Thesis Project (Mandatory or Not)	Nominal	The Goal of Pursuing Education	Nominal
Failed Course	Nominal	Internet Access	Nominal
Major of Study	Nominal	Romantic Relationship	Nominal
Bachelor's Degree GPA	Nominal	Mother's Education	Nominal
Rank of Bachelor's University	Nominal	Father's Education	Nominal
Gap Years Between Bachelor and Master	Nominal	Parents Status	Nominal
Master's Degree GPA	Nominal	Final GPA	Nominal

datasets become easy to handle [28]. Fortunately, handling the missing data as a step of data preprocessing is not needed because the datasets used in this research have no missing data.

1) IMBALANCED DATA PROBLEM

Imbalanced data problem occurs in many real-world datasets where the class distributions of data are highly imbalanced. It is important to note that most machine learning models work best when the number of instances of each class is approximately equal [29]. The imbalanced data problem causes the majority class to dominate the minority class; hence, the classifiers are more inclined to the majority class, and their performance cannot be reliable [30].

Analyzing the introduced datasets reveals that they are highly imbalanced, and the four categories of students based on their grade point average are not equal. In fact, the Iran dataset includes more samples from Medium (40% of samples) and Good classes (40% of samples), while the other two classes have fewer samples (the Poor class with 11% of samples and the Excellent class with only 9% of samples). The Portugal dataset involves 15% of samples related to Poor class, 44% of samples to Medium class, 35% of samples to Good class, and only 6% of samples to Excellent class. Accordingly, it is necessary to solve the imbalanced data problem because this problem may lead to unpredictable outcomes. Figure 1 shows the distribution of the students' performance based on the different classes of both datasets.

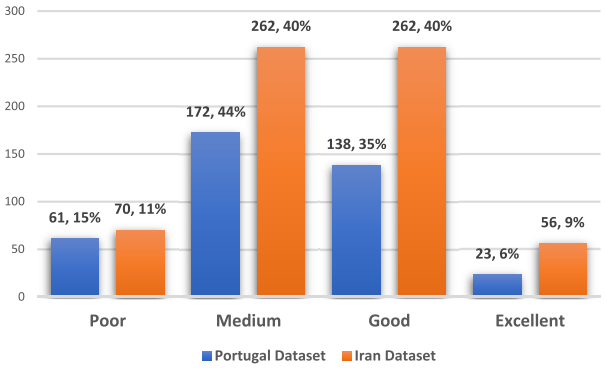

FIGURE 1. The distribution of the students' performance of both datasets.

Many strategies have been generated that can handle the imbalanced data problem. The sampling-based approach is one of the most effective methods that can solve the imbalanced data problem. The sampling-based approach can be classified into three categories, namely: Over-Sampling [31], Under-Sampling [32], and Hybrid-Sampling [33].

a: OVER-SAMPLING METHOD

Over-sampling raises the weight of the minority class by replicating or creating new minority class samples. There are

different over-sampling methods; moreover, it is worth noting that the over-sampling approach is generally applied more frequently than other approaches.

- Random Over Sampler

This method increases the size of the dataset by the repetition of the original samples. The point is that the random over sampler does not create new samples, and the variety of samples does not change [34].

- SMOTE

This method is a statistical technique that increases the number of minority samples in the dataset by generating new instances. This algorithm takes samples of the feature space for each target class and its nearest neighbors, and then creates new samples that combine features of the target case with features of its neighbors. The new instances are not copies of existing minority samples [35].

- Borderline SMOTE

In this method, samples and the neighboring ones are more likely to be misclassified than the ones far from the borderline. This method uses the number of majority neighbors of each minority sample to divide the minority samples into the three groups, namely Safe, Danger, and Noise. It should be noted that the Danger group is used to generate new instances [36].

- SVM-SMOTE

This method generates the new minority class samples along with directions from existing minority class instances towards their nearest neighbors. The SVM-SMOTE focuses on creating new minority class samples near borderlines using the SVM model to help set boundaries between classes [37].

b: UNDER-SAMPLING METHOD

Under-sampling is one of the most straightforward strategies to handle the imbalanced data problem. This method under-samples the majority class to balance the class with the minority class. The under-sampling method is applied when the amount of collected data is sufficient. There are different under-sampling models, such as Edited Nearest Neighbors (ENN) [38] and Tomek links [39], which are the most popular ones.

c: HYBRID METHODS

Over-sampling and under-sampling have different advantages and disadvantages. Combining these two methods can help to get benefits and drawbacks of both approaches.

- SMOTE-ENN

This method is one of the well-known methods that combines the SMOTE as over-sampling model and ENN as an under-sampling model to improve the results [40].

- SMOTE-Tomek

This method is another common hybrid method that connects the SMOTE as an over-sampling model to Tomek links as an under-sampling model to enhance the results [40].

TABLE 3. Resampling methods with their parameters' settings.

Methods	Parameters
SMOTE	K_Neighbors = 5
Borderline SMOTE	K_Neighbors = 5, M_ Neighbors = 10
Random Over Sampler	No Parameters
SMOTE-ENN	$K_{(SMOTE)} = 5$, $K_{(ENN)}=3$
SVM-SMOTE	K_Neighbors = 5, M_ Neighbors = 10
SMOTE-Tomek	$K_{(SMOTE)} = 5$

All of the used resampling methods in this paper are listed in Table 3, together with their most important parameters' settings. Best results are achieved using these settings.

2) FEATURE SCALING

Feature scaling or data normalization is a technique that helps to normalize the range of independent variables or features of the dataset. Most of the machine learning models use the Euclidean distance between two data points, so they may not work well without Feature Scaling [41]. There are four popular ways to implement Feature Scaling, namely Standardization, Mean Normalization, Min-Max Scaling, and Unit Vector. The range of students' performance dataset values used in this paper is widely varied. This paper uses the Standardization method to rescale the features. As a result, all the features have the standard normal distribution characteristics with $\mu = 0$ and $\sigma = 1$ where μ is the average, and σ is the standard deviation from the average. The formula used to scale the values is as follows [42]:

$$\mathcal{Z} = \frac{x - \mu}{\sigma} \tag{1}$$

III. MACHINE LEARNING MODELS

There are various classifications machine learning models. This paper carries out different classifiers, including Random Forest [43], [44], K-nearest-neighbor [45], [46], Artificial Neural Network [47], [48], XG-boost [49], [50], Support Vector Machine (Radial Basis Function kernel) [51], [52], Decision Tree [53], [54], Logistic Regression [55], [56], and Naïve Bayes [57]. It is a well-established fact that most machine learning classifiers support multiclass classification inherently, such as Artificial Neural Network (ANN), K-Nearest Neighbor (KNN), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), and Naïve Bayes (NB). Since Support Vector Machine (SVM) and XG-Boost do not support multiclass classification inherently, one vs. one method is used for applying the Support Vector Machine model, and one vs. all method is used for implementing XG-Boost model.

All of the used machine learning models in this paper are listed in Table 4, together with their specific parameters' settings.

TABLE 4. Machine Learning models with their specific parameters' settings.

Methods	Parameters
Artificial Neural Network	1 hidden layer, Activation Function = rectified linear unit, Maximum iterations = 200
K-Nearest Neighbor	N_Neighbors = 2, weight function = distance, leaf_size = 30
Random Forest	N_estimators = 113, min_samples_leaf = 2
Logistic Regression	C = 1.2, penalty = l2, solver = liblinear
Decision Tree	Criterion = entropy, Splitter = best
Naïve Bayes	No Parameters
Support Vector Machine	C=1, Kernel = rbf, Gamma= scale
XG-Boost	N_estimators = 70, loss function = deviance, learning_rate = 1.0

IV. MODEL VALIDATION

Cross-validation is a model validation technique applied to evaluate how the statistical analysis results are generalized into an independent dataset. This paper uses two popular different cross-validation approaches, which are random hold-out (randomly divides the 80% of the data into the training set and 20% into the test set) and shuffle 5-fold cross-validation. It should be noted that the resampling method can only be used on the training set, and the test set classes should not be balanced at all. Therefore, all the resampling methods are applied to the training set while using different model validation.

V. EVALUATION METHODS

Evaluating the performance of classifiers is an essential part of comparing and finding the best model. There are many ways to measure and check the performance of machine learning algorithms. This paper uses various evaluation methods such as prediction Accuracy, Sensitivity, Precision, and F1-score; moreover, the statistical evaluation strategy is used for a more trustable and powerful analyzing and comparing.

Analyzing and comparing the classifiers' performance is a significant procedure. Although it is simple to use evaluation measures, the obtained results may be misleading. Therefore, finding the best model or method based on their capabilities is a critical challenge. Statistical significance tests are planned to solve this problem [58]. The repeated-measures ANOVA is the typical statistical test method which is used to determine the differences between more than two related sample means. The null-hypothesis being examined in the ANOVA test is that all resampling methods perform the same, and the detected differences are only arbitrary [59]. It should be noted that the ANOVA test considers three assumptions. These assumptions are as follows:

1- The samples should be normally distributed.

2- the sample cases should be autonomous from each other.

3- the variance between the groups (methods which are being compared) should be approximately equal.

This paper uses the Anderson–Darling normality test [59] to evaluate the normality of data. This test is a modification of the Kolmogorov-Smirnov test [60]. The null hypothesis of this normality test is that the data have a normal distribution; accordingly, if the p-value of this normality test is less than α ($\alpha = 0.05$), the null hypothesis will be rejected, and the data do not have a normal distribution.

It is a well-established fact that the ANOVA assumptions can be violated. Therefore, the Friedman test, which is a non-parametric option of the ANOVA test, can be applied to examine the differences between models and methods [61]. The null-hypothesis of the Friedman test is that all resampling methods perform the same; also, the rejection of this null hypothesis implies that one or more of the resampling methods have a different performance. This paper uses the accuracy data gathered by shuffle 5-fold cross-validation for each resampling method.

The Freidman test ranks the data of each classifier for each resampling method, then analyzes the values of ranks [62]. Accordingly, the Friedman test gives a sum of ranks for each resampling method that assists in defining the most effective resampling method among all others.

VI. RESULTS & DISCUSSION

This paper tries to show the effect of imbalanced data problem and handle this problem using various resampling methods; additionally, determining the best resampling method and the best classifier compare to all other models and examining the difference between multiclass and binary classification and the importance of the features' structure are among the aims of this paper. All presented models and methods have coded in Python, which is an interpreted, general-purpose, high-level programming language. Moreover, all practical operations are performed with a 2 GHz Intel Core i7 MacBook Pro with 4GB of RAM. It should be pointed out that all the classifiers are first executed on the imbalanced data to show the effect of the imbalanced data problem on the models' performance. Next, all the classifiers are implemented on balanced data generated by resampling methods to notify a better perception of the effectiveness of the resampling methods as ways to solve the imbalanced problem.

A. RANDOM HOLD-OUT METHOD RESULTS

Table 5 shows the performance of the different classifiers on the imbalanced datasets using the random hold-out approach. Various evaluation measure methods such as Accuracy, Recall, Precision, and F1-score are used to provide a better understanding of the performance of the models.

TABLE 5. Performance of the classifiers based on the hold-out strategy on imbalanced data.

Model	Dataset	Test Set Accuracy	Recall	Precision	F1 Score			
					Poor	Medium	Good	Excellent
Artificial Neural Network	Iran	**58.46%**	**58.46%**	**52.98%**	0%	67%	56%	53%
	Portugal	67.97%	67.97%	67.93%	55%	71%	71%	0%
XG-Boost	Iran	57.69%	56.15%	51.19%	0%	64%	57%	59%
	Portugal	69.24%	69.24%	64.85%	41%	71%	66%	0%
Support Vector Machine (RBF Kernel)	Iran	56.15%	56.15%	44.80%	0%	64%	61%	0%
	Portugal	69.11%	69.11%	61.07%	0%	81%	62%	0%
K-Nearest-Neighbor	Iran	54.61%	54.61%	43.52%	0%	65%	57 %	0%
	Portugal	75.56%	75.56%	66.32%	0%	85%	77%	0%
Random Forest	Iran	54.61%	54.61%	43.57%	0%	62%	60%	0%
	Portugal	**76.83%**	**76.83%**	**78.44%**	72%	84%	73%	0%
Logistic Regression	Iran	53.07%	53.07%	48.21%	0%	61%	52%	46%
	Portugal	69.24%	69.24%	61.35%	0%	71%	79%	0%
Decision Tree	Iran	48.46%	48.46%	49.94%	33%	55%	52%	27%
	Portugal	56.58%	56.58%	58.23%	63%	59%	56%	0%
Naïve Bayes	Iran	46.15%	46.15%	27.76%	0%	67%	0%	56%
	Portugal	47.72%	47.72%	61.71%	51%	0%	65%	0%

One of the most popular evaluation techniques to measure a classifier's performance is accuracy. This metric is the proportion between the number of correct predictions and the total number of samples examined. Although accuracy is easy to understand, it ignores many essential factors that should be considered in assessing the performance of a classifier. In Iran Dataset, all of the accuracy results are below 60%, which reveals that none of the classifiers have achieved satisfactory and remarkable accuracy results. The Artificial Neural Network classifier has obtained 58.46% accuracy, which is the best result among all other models. Also, the worst accuracy result belongs to the Naïve Bayes with the accuracy of 46.15%. In Portugal dataset, the Random Forest, with an accuracy of 76.83%, has the best performance, and the Naïve Bayes with an accuracy of 47.72% has the worst performance.

The Recall is the probability of detection indicating the proportion of items identified correctly. It means that ANN correctly identifies 58.46% of all different students in the Iran dataset. All the classifiers' recall test results are similar to their accuracy results. Besides, precision is the portion of the relevant results. The results of this test do not include remarkable outcomes using Iran dataset. The highest precision in the Iran dataset belongs to ANN with 52.98%, and the lowest one goes to Naïve Bayes with 27.76% among all other classifiers. In the Portugal dataset, results are so much better. In fact, the highest precision goes to Random Forest with 78.44%, and the lowest one goes to Decision Tree with 58.23% among all other classifiers.

F1-score, which is the harmonic average of Precision and Recall, includes critical and indispensable results about classifiers' performance on each class. As stated, the distribution of the classes is not balanced, and the majority of the data relates only to the two of the classes. Considering both datasets, the results of the F1-score with each class reveal that predictive classifiers do not perform well with some of the classes. For example, the ANN model and Random Forest, which have the best accuracy result among others in both datasets, fail to predict one of the classes or Support Vector Machine (RBF Kernel) fails to predict two of the classes in the Iran dataset. Accordingly, the classifiers' performance is not acceptable

One of the essential results from table 5 is the overall low performance in all the classifiers. For example, in the Iran dataset, the highest accuracy among the classifiers belongs to. ANN with 58.46%, and in the Portugal dataset, Random Forest, with 46.83% has the best performance. There are lots of reasons that can reduce the initial accuracy in classification. One of the reasons could be the structure of the features. The initial results of the Portugal dataset are better than Iran dataset, and this can be because of their features' structure. Actually, Portugal dataset has more numeric features that help the models to better find the patterns in the data. Another reason could be the number of classes. As mentioned, this paper deals with two different datasets that have four classes each. To analyze the effect of the number of classes on the initial performance, we reduced the number of classes to two. The initial accuracy results of the implementation of ML models on the new datasets are shown in Table 7. The highest accuracy among the classifiers in binary classification belongs to Logistic Regression with 77.69% for

TABLE 6. Accuracy results of the classifiers based on the hold-out strategy on different balanced data.

Model	Dataset	Unbalanced Data	SMOTE	Borderline SMOTE	Random Over Sampler	SMOTE ENN	SVM SMOTE	SMOTE Tomek
Artificial Neural Network	Iran	**58.46%**	47.69%	43.84%	43.84%	40.76%	48.46%	45.38%
	Portugal	67.97%	61.64%	60.37%	64.17%	65.44%	66.70%	62.91%
XG-Boost	Iran	57.69%	48.46%	48.46%	48.46%	49.23%	48.46%	42.30%
	Portugal	69.24%	67.97%	73.03%	66.70%	66.21%	69.24%	66.70%
Support Vector Machine (RBF Kernel)	Iran	56.15%	49.23%	52.30%	48.46%	53.07%	54.61%	50.76%
	Portugal	69.11%	59.11%	65.44%	60.37%	59.11%	66.70%	62.91%
K-Nearest-Neighbor	Iran	54.61%	42.30%	42.30%	45.38%	43.84%	45.38%	36.92%
	Portugal	75.56%	64.17%	64.17%	65.44%	59.11%	64.91%	62.91%
Random Forest	Iran	54.61%	**55.38%**	**55.38%**	**56.15%**	**56.92%**	**58.46%**	**51.53%**
	Portugal	**76.83%**	**71.77%**	75.56%	**69.24%**	**73.03%**	**76.30%**	**70.50%**
Logistic Regression	Iran	53.07%	44.61%	45.38%	43.07%	43.07%	46.92%	41.53%
	Portugal	69.24%	56.58%	55.31%	52.78%	54.05%	62.91%	59.11%
Decision Tree	Iran	48.46%	38.46%	40.00%	45.38%	44.46%	40.76%	40.76%
	Portugal	56.58%	70.50%	**76.83%**	66.70%	52.91%	64.17%	56.70%
Naïve Bayes	Iran	46.15%	39.23%	42.30%	41.53%	38.46%	44.61%	43.84%
	Portugal	47.72%	51.51%	57.84%	46.45%	52.78%	55.31%	51.51%

TABLE 7. Accuracy results of the classifiers based on the binary classification.

Models	Iran Dataset	Portugal Dataset
Artificial Neural Network	66.15%	78.10%
XG-Boost	74.61%	93.29%
Support Vector Machine	75.38%	85.69%
K-Nearest-Neighbor	67.69%	81.89%
Random Forest	74.61%	88.22%
Logistic Regression	77.69%	88.22%
Decision Tree	65.38%	77.10%
Naïve Bayes	66.15%	80.63%

the Iran dataset and XG-Boost with 93.29% for the Portugal dataset. The results reveal that ML models have so much better performance while using binary classification, and the accuracies are increased significantly. In fact, decreasing the number of classes has a great effect on the performance of the models. The high number of classes increases the complexity; therefore, models need a high number of samples to better find the patterns in multi-classification problems. This means that given a fixed number of samples, a greater number of classes will lead to poorer results.

As mentioned, this paper works on two different datasets in a multi-classification problem and tries to determine the effect of the imbalanced data problem and discover the best resampling method and classifier. The results obtained from imbalanced data in the multi-classification problem indicate that machine learning algorithms do not give accurate results using imbalanced datasets; also, most of the classifiers cannot predict all the target classes. Therefore, solving the imbalanced data problem is notably necessary. Table 6 represents the accuracy obtained by each machine learning technique on both balanced datasets using six different resampling models.

The accuracy result achieved using an imbalanced dataset is not acceptable. Accuracy can be a useful measure if data has the same number of samples per class. However, with an imbalanced set of samples, accuracy is not helpful at all because the model predicts the value of the majority classes for all predictions. The results of the accuracy-test using the balanced dataset are not significantly improved. It is logical that most of the classifiers are predicting with lower accuracy results on the balanced data because they are considering all classes. Since the imbalanced data problem is handled by resampling methods, the accuracy can be trustable now.

Table 8 indicates the results of the Recall and Precision tests at the same time. It should be pointed out that the results of the recall test are the same as the accuracy, but some of the machine learning models have notable advancement in their precision results. For instance, in the Iran dataset, the Support Vector Machine achieved the result of 44.80% with the precision test using imbalanced data, while the result is increased up to the 57.31% with balanced data using SVM-SMOTE method. Moreover, in the Portugal dataset, the XG-Boost obtained the result of 64.85% with the precision test using imbalanced data, while the result is increased up to the 76.32% with balanced data using

TABLE 8. Recall and Precision results of the classifiers based on the hold-out strategy on different balanced data.

Model		SMOTE		Borderline SMOTE		Random Over Sampler		SMOTE ENN		SVM SMOTE		SMOTE Tomek	
		Iran	Portugal	Iran	Portugal	Iran	Portugal	Iran	Portugal	Iran	Portugal	Iran	Portugal
Artificial Neural Network	Precision	49.29%	64.00%	45.07%	61.46%	43.63%	64.95%	40.76%	68.57%	51.35%	66.68%	47.61%	62.70%
	Recall	47.69%	61.64%	43.84%	60.37%	43.84%	64.17%	42.16%	65.44%	48.46%	66.70%	45.38%	62.91%
XG-Boost	Precision	46.99%	70.57%	48.09%	76.32%	48.17%	71.41%	50.34%	68.57%	47.71%	69.84%	41.69%	71.07%
	Recall	48.46%	67.97%	48.46%	73.03%	48.46%	66.70%	49.23%	66.21%	48.46%	69.24%	42.30%	66.70%
Support Vector Machine	Precision	51.74%	62.55%	55.37%	66.51%	53.87%	64.10%	56.10%	61.10%	57.31%	66.39%	56.31%	64.81%
	Recall	49.23%	59.11%	52.30%	65.44%	48.46%	60.37%	53.07%	59.11%	54.61%	66.70%	50.76%	62.91%
K-Nearest-Neighbor	Precision	49.13%	76.77%	48.14%	66.33%	48.56%	63.82%	50.21%	64.52%	51.43%	62.94%	44.83%	65.62%
	Recall	42.40%	64.17%	42.30%	64.17%	45.38%	65.44%	43.84%	59.11%	45.38%	62.91%	36.92%	62.91%
Random Forest	Precision	52.15%	72.68%	51.77%	76.08%	54.69%	70.19%	57.63%	75.01%	54.71%	73.60%	48.95%	70.64%
	Recall	55.38%	71.77%	55.38%	75.56%	56.15%	69.24%	56.92%	73.03%	58.46%	76.30%	51.53%	70.50%
Logistic Regression	Precision	48.18%	64.21%	49.00%	60.80%	48.53%	58.85%	48.44%	59.88%	49.46%	66.07%	45.85%	66.45%
	Recall	44.61%	56.58%	45.38%	55.31%	43.07%	52.78%	43.07%	54.05%	46.92%	62.91%	41.53%	59.11%
Decision Tree	Precision	38.38%	59.59%	41.25%	76.42%	45.65%	67.22%	47.77%	68.48%	41.47%	65.67%	43.46%	68.94%
	Recall	38.46%	70.50%	40.00%	76.83%	45.38%	66.70%	44.46%	62.91%	40.76%	64.17%	40.76%	66.70%
Naïve Bayes	Precision	28.70%	72.73%	30.92%	69.89%	30.87%	52.17%	49.35%	66.91%	48.74%	68.24%	31.24%	70.39%
	Recall	39.23%	51.50%	42.30%	57.84%	41.53%	46.45%	38.46%	52.78%	44.61%	55.31%	43.84%	51.51%

Borderline SMOTE method. As mentioned, to better analyze the recall and precision tests, it is more beneficial to use the F1-score.

It should be regarded that the classifiers do not achieve an excellent result with the F1-score while using imbalanced data, and classifiers do not perform well with all the classes. This is an essential problem that should be solved by handling the imbalanced data problem. After using different resampling methods and solving the imbalanced data problem, the results show that classifiers do not ignore any classes, and all four classes are predicted and analyzed in both datasets. This is one of the most significant reasons for using balanced data. For example, the Artificial Neural Network model ignores one of the classes while using imbalanced datasets. However, after solving the imbalance problem, this model considers all the classes. Table 9 presents the results of the F1-score for all applied machine learning models.

B. SHUFFLE 5-FOLD CROSS-VALIDATION RESULTS

This paper utilizes the shuffle 5-fold cross-validation, which splits the dataset into five subsets and uses one of the five subsets as the test set and the other four subsets as the training set every time and then repeats the hold-out method five times. Table 10 shows the achieved average accuracy results and variance of implementing machine learning models using this type of validation method.

The results of shuffle 5-fold cross-validation are more trustable and acceptable because of the way that this strategy works. The results show that after solving the imbalanced data problem, there is a slight improvement in some of the models' accuracies. The obtained results from the balanced dataset using SVM-SMOTE is significantly better than other datasets. In the Iran dataset, the Random Forest classifier achieved 73 % with the accuracy, which is acceptable and better than other models. Also, In the Portugal dataset, the Random Forest has reached an impressive accuracy of 81.27%, which is the best performance among all the models in all situations. Regarding the performance of classifiers using other resampling methods, it can be noted that Random Forest has achieved excellent results in almost all the balanced datasets.

C. STATISTICAL TEST RESULTS

Various resampling methods provide different balanced data and classifiers have different performances while using different balanced data. Therefore, it is so hard to find the best resampling method to achieve the best results from machine learning models.

Statistical significance tests assist in dealing with the challenge of choosing the best resampling method. As declared, this paper uses the accuracy data collected by shuffle 5-fold cross-validation for each resampling method based on different machine learning models. It is worth noting that the

TABLE 9. F1-score results of the classifiers based on the hold-out strategy on different balanced data.

Model		SMOTE		Borderline SMOTE		Random Over Sampler		SMOTE ENN		SVM SMOTE		SMOTE Tomek	
		Iran	Portugal	Iran	Portugal	Iran	Portugal	Iran	Portugal	Iran	Portugal	Iran	Portugal
Random Forest	Poor	10%	38%	13%	32%	11%	36%	33%	33%	21%	50%	11%	40%
	Medium	64%	45%	64%	55%	65%	48%	68%	52%	66%	45%	60%	47%
	Good	55%	44%	54%	42%	56%	39%	55%	47%	60%	44%	50%	39%
	Excellent	52%	34%	56%	58%	53%	57%	44%	32%	56%	24%	56%	52%
XG-Boost	Poor	12%	39%	21%	40%	22%	33%	42%	27%	10%	31%	10%	27%
	Medium	53%	33%	56%	42%	52%	40%	52%	41%	50%	38%	49%	32%
	Good	48%	48%	46%	55%	48%	43%	48%	42%	53%	48%	41%	51%
	Excellent	53%	44%	47%	32%	50%	43%	49%	47%	55%	34%	43%	27%
Support Vector Machine	Poor	30%	18%	21%	27%	18%	18%	26%	22%	31%	17%	29%	24%
	Medium	57%	37%	60%	43%	56%	40%	63%	38%	61%	44%	57%	42%
	Good	49%	29%	54%	34%	52%	31%	55%	25%	56%	37%	54%	34%
	Excellent	44%	18%	51%	20%	48%	15%	42%	14%	55%	25%	49%	18%
K-Nearest-Neighbor	Poor	31%	32%	40%	21%	26%	23%	29%	18%	31%	8%	29%	14%
	Medium	58%	36%	52%	41%	59%	44%	59%	34%	62%	37%	51%	45%
	Good	29%	43%	31%	35%	38%	42%	33%	33%	35%	40%	26%	31%
	Excellent	37%	17%	38%	29%	35%	18%	41%	22%	37%	29%	31%	24%
Naïve Bayes	Poor	30%	30%	35%	39%	36%	20%	30%	19%	32%	34%	35%	30%
	Medium	57%	20%	61%	19%	59%	54%	53%	20%	62%	15%	62%	19%
	Good	49%	33%	41%	40%	37%	34%	49%	38%	48%	40%	43%	34%
	Excellent	44%	8%	49%	11%	50%	8%	51%	10%	49%	10%	53%	18%
Decision Tree	Poor	24%	46%	14%	41%	36%	40%	30%	30%	18%	52%	21%	36%
	Medium	46%	51%	49%	60%	54%	44%	57%	36%	52%	35%	48%	48%
	Good	40%	30%	40%	41%	42%	33%	42%	41%	41%	31%	41%	33%
	Excellent	34%	00%	32%	36%	32%	42%	32%	38%	28%	37%	35%	38%
Artificial Neural Network	Poor	30%	8%	15%	7%	20%	17%	18%	27%	27%	16%	44%	27%
	Medium	57%	41%	55%	31%	57%	42%	50%	41%	58%	47%	51%	44%
	Good	45%	41%	42%	46%	37%	37%	42%	42%	52%	38%	45%	36%
	Excellent	42%	8%	38%	18%	33%	15%	24%	47%	28%	18%	30%	26%
Logistic Regression	Poor	19%	23%	24%	19%	22%	18%	36%	21%	26%	13%	26%	29%
	Medium	52%	26%	52%	22%	50%	15%	48%	14%	54%	41%	49%	30%
	Good	40%	39%	41%	39%	40%	39%	38%	40%	42%	42%	33%	39%
	Excellent	52%	51%	56%	41%	51%	37%	49%	10%	54%	41%	52%	39%

TABLE 10. Accuracies results of the classifiers based on the shuffle 5-fold cross-validation on the different balanced datasets.

Model	Unbalanced	SMOTE	Borderline SMOTE	Random Over Sampler	SMOTE ENN	SVM SMOTE	SMOTE Tomek
Artificial Neural Network	**53.86 ± 6 %**	46.46 ± 4 %	44.76 ± 4 %	46.76 ± 2 %	48.00 ± 2 %	67.84 ± 3 %	44.15 ± 2 %
	65.46 ± 5 %	66.04 ± 3 %	67.83 ± 5 %	68.08 ± 3 %	63.52 ± 4 %	69.06 ± 2 %	63.76 ± 2%
XG-Boost	52.49 ± 5 %	52.92 ± 4 %	54.43 ± 9 %	51.53 ± 3 %	52.76 ± 3 %	72.33 ± 3 %	50.15 ± 3 %
	76.97 ± 5 %	72.64 ± 3 %	73.66 ± 3 %	68.32 ± 4 %	64.28 ± 5 %	70.86 ± 4 %	72.63 ± 2%
Support Vector Machine	51.72 ± 5 %	46.92 ± 4 %	46.76 ± 4 %	44.61 ± 3 %	47.69 ± 3 %	67.56 ± 2 %	44.30 ± 3 %
	75.89 ± 2 %	65.78 ± 3 %	67.30 ± 3 %	65.51 ± 4 %	67.30 ± 2 %	69.07 ± 3 %	63.49 ± 2 %
K-Nearest-Neighbor	51.58 ± 6 %	38.61 ± 3%	39.38 ± 2 %	42.15 ± 3 %	42.61 ± 1 %	68.70 ± 3 %	38.30 ± 1 %
	73.58 ± 3 %	55.63 ± 4%	65.02 ± 3%	70.35 ± 2%	65.01 ± 3%	78.82 ± 4%	62.47 ± 2%
Random Forest	34.93 ± 9 %	**55.38 ± 1 %**	**57.07 ± 0 %**	**52.46 ± 1 %**	**52.92 ± 3 %**	**73.00 ± 3 %**	**52.76 ± 1 %**
	78.19 ± 3 %	**77.70 ± 3 %**	**77.19 ± 2 %**	77.46 ± 1 %	**76.96 ± 2 %**	**81.27 ± 2 %**	**77.97 ± 1 %**
Logistic Regression	49.89 ± 5 %	37.69 ± 4 %	37.84 ± 2 %	36.61 ± 3 %	42.46 ± 4 %	59.35 ± 3 %	36.46 ± 4 %
	77.41 ± 3 %	59.70 ± 4 %	64.01 ± 3 %	**78.42 ± 2 %**	64.00 ± 5 %	70.35 ± 4 %	58.42 ± 3 %
Decision Tree	49.87 ± 5 %	46.71 ± 1 %	44.61 ± 3 %	44.15 ± 1 %	44.61 ± 3 %	64.60 ± 2 %	47.07 ± 2 %
	63.71 ± 1 %	68.06 ± 3 %	72.39 ± 4 %	70.09 ± 3 %	69.32 ± 2 %	67.56 ± 4 %	67.57 ± 5 %
Naïve Bayes	39.37 ± 9 %	33.53 ± 4 %	35.23 ± 5 %	31.76 ± 8 %	33.38 ± 2 %	54.49 ± 4 %	35.69 ± 5 %
	46.03 ± 4 %	43.10 ± 2 %	46.15 ± 4 %	62.94 ± 3 %	48.42 ± 3 %	56.63 ± 2 %	41.07 ± 4 %

TABLE 11. The Anderson-Darling normality test results.

Datasets	Mean	Standard Deviation	Number of samples	P-value
Iran Dataset	48.01 %	10.50 %	48	**0.024**
Portugal Dataset	66.52 %	8.88 %	48	**<0.005**

TABLE 12. The Friedman test results.

Datasets	Degrees of freedom	Chi-Square	P-Value
Iran Dataset	5	24.80	**0.000**
Portugal Dataset	5	11.74	**0.039**

normality assumption should be checked before applying the ANOVA test. The Anderson-Darling normality test results on shuffle 5-fold cross-validation indicate that the p-value is less than 0.05 ($\alpha = 0.05$) for both datasets; therefore, the null hypothesis is rejected, and the ANOVA test cannot be used. Table 11 reveals the results of the Anderson-Darling normality test for both datasets.

Since the ANOVA normality assumption is violated, the Friedman test is applied for comparing the resampling methods instead of the ANOVA test in both datasets. Table 12 displays the results of the Friedman test.

These results show that the p-value of both datasets is less than the significance level ($\alpha = 0.05$). Therefore, the null hypothesis is rejected, and it can be concluded that at least one of the resampling methods has a different effect on both datasets.

TABLE 13. Additional information from Friedman test results.

Iran Dataset				Portugal Dataset			
Rank	Resampling Methods	Median	Sum of Ranks	Rank	Resampling Methods	Median	Sum of Ranks
1	**SVM-SMOTE**	**66.58**	**48.0**	**1**	**SVM-SMOTE**	**69.56**	**37.0**
2	SMOTE-ENN	46.37	30.5	**2**	Random Over Sampler	68.50	35.0
3	Borderline-SMOTE	46.05	29.0	**3**	Borderline-SMOTE	67.11	33.5
4	SMOTE	45.87	28.0	**4**	SMOTE	65.17	23.0
5	SMOTE-Tomek	44.07	18.0	**5**	SMOTE-ENN	65.53	21.5
6	Random Over Sampler	44.02	14.5	**6**	SMOTE-Tomek	64.01	18.0
	Overall	**48.83**			**Overall**	**66.65**	

Table 13 exposes the results of the median and sum of ranks derived from the Friedman test in both datasets. The midpoint of the dataset is named the median.

The data points of each resampling method are split equally above and below the midpoint value. Furthermore, the overall median is the midpoint of all data points. The

median response for the SVM-SMOTE method is considerably higher than the overall median in both datasets. Moreover, the result of the sum of ranks for the SVM-SMOTE method is better than other resampling methods in both datasets. These results confirm that the SVM-SMOTE method might be more efficient than the other methods.

VII. CONCLUSION

The recent improvements in numerous areas have led to the collection of a considerable amount of data. Today, educational institutions collect information about students. One of the main challenges for these institutions is analyzing and predicting their students' performance. Educational Data mining is a robust analytical method that can be used to discover significant and meaningful knowledge from educational data; however, it can face some difficulties such as imbalanced educational data problems in predicting students' performance.

This study intends to show the effect of imbalanced data problem and find the best resampling method among the different methods of handling the imbalanced data problem, namely Borderline SMOTE, Random Over Sampler, SMOTE, SVM-SMOTE, SMOTE-ENN, and SMOTE-Tomek. It should be noted that two different datasets related to students' performance are used, the difference between multiclass and binary classification, and structures of the features are considered. Several classifiers are applied to inform a better conclusion of resampling methods. All the classifiers are first performed using the random hold-out method on the imbalanced dataset. The results show that classifiers do not have acceptable predictions while imbalanced data, and they cannot predict some of the classes at all. Moreover, the obtained results using different evaluation metrics indicate that the few numbers of classes will lead to better performance with machine learning models. Also, more numeric features help the models to have better performance. Using the random hold-out method on different balanced data generated by various resampling methods determine that the performance of some classifiers is improved, and all the classes are predicted, so the classifiers' performance is satisfactory. Moreover, shuffle 5-fold cross-validation is used to achieve more reliable results with accuracy. The results of this validation method indicate that classifiers have a varying performance on different balanced data for both datasets; therefore, selecting the best resampling method is not easy. However, it seems that classifiers work better on the data balanced by the SVM-SMOTE method in both Iran and Portugal datasets. This paper used the Friedman test to choose the best resampling method. The results of this test confirm that the performance of SVM-SMOTE is better than other resampling methods. Also, the Random Forest model has achieved the best results among other classifiers while using the SVM-SMOTE resampling method.

This study can be developed in many ways, and it is possible to perform future work in the following directions. New ensemble and hybrid classifiers can be introduced for having a better comparison and also achieving higher performance. Additionally, feature selection methods as a way of improving models' results can be performed to get a better perspective on the significant features.

REFERENCES

[1] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 135–146, Jul. 2007.
[2] R. Ghorbani and R. Ghousi, "Predictive data mining approaches in medical diagnosis: A review of some diseases prediction," *Int. J. Data Netw. Sci.*, vol. 3, no. 2, pp. 47–70, 2019.
[3] D. Kabakchieva, "Student performance prediction by using data mining classification algorithms," *Int. J. Comput. Sci. Manage. Res.*, vol. 1, no. 4, pp. 686–690, 2012.
[4] 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, Mar. 2019.
[5] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. V. Erven, "Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil," *J. Bus. Res.*, vol. 94, pp. 335–343, Jan. 2019.
[6] R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *J. Educ. Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.
[7] E. Chandra and K. Nandhini, "Knowledge mining from student data," *Eur. J. Sci. Res.*, vol. 47, no. 1, pp. 156–163, 2010.
[8] A. B. El Din Ahmed and I. S. Elaraby, "Data mining: A prediction for student's performance using classification method," *World J. Comput. Appl. Technol.*, vol. 2, no. 2, pp. 43–47, 2014.
[9] M. M. A. Tair and A. M. El-Halees, "Mining educational data to improve students' performance: A case study," *Int. J. Inf. Commun. Technol. Res.*, vol. 2, no. 2, pp. 1–7, 2012.
[10] H. A. A. Hamza and P. Kommers, "A review of educational data mining tools & techniques," *Int. J. Educ. Technol. Learn.*, vol. 3, no. 1, pp. 17–23, 2018.
[11] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 40, no. 6, pp. 601–618, Nov. 2010.
[12] C. Márquez-Vera, A. Cano, C. Romero, and S. Ventura, "Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data," *Int. J. Speech Technol.*, vol. 38, no. 3, pp. 315–330, Apr. 2013.
[13] S. T. Karamouzis and A. Vrettos, "An artificial neural network for predicting student graduation outcomes," in *Proc. World Congr. Eng. Comput. Sci.*, 2008, pp. 991–994.
[14] R. M. de Albuquerque, A. A. Bezerra, D. A. de Souza, L. B. P. do Nascimento, J. J. de Mesquita Sá, and J. C. do Nascimento, "Using neural networks to predict the future performance of students," in *Proc. Int. Symp. Comput. Educ. (SIIE)*, Nov. 2015, pp. 109–113.
[15] S. A. Naser, I. Zaqout, M. A. Ghosh, R. Atallah, and E. Alajrami, "Predicting student performance using artificial neural network: In the faculty of engineering and information technology," *Int. J. Hybrid Inf. Technol.*, vol. 8, no. 2, pp. 221–228, Feb. 2015.
[16] T. Devasia, T. P. Vinushree, and V. Hegde, "Prediction of students performance using educational data mining," in *Proc. Int. Conf. Data Mining Adv. Comput. (SAPIENCE)*, Mar. 2016, pp. 91–95.
[17] Z. Kovacic, "Early prediction of student success: Mining students' enrolment data," in *Proc. Informing Sci. IT Educ. Conf. (InSITE)*, New Zealand, 2010.
[18] A. Acharya and D. Sinha, "Early prediction of students performance using machine learning techniques," *Int. J. Comput. Appl.*, vol. 107, no. 1, pp. 37–43, 2014.
[19] F. Marbouti, H. A. Diefes-Dux, and K. Madhavan, "Models for early prediction of at-risk students in a course using standards-based grading," *Comput. Educ.*, vol. 103, pp. 1–15, Dec. 2016.
[20] M. Hlosta, Z. Zdrahal, and J. Zendulka, "Ouroboros: Early identification of at-risk students without models based on legacy data," in *Proc. 7th Int. Learn. Anal. Knowl. Conf.*, 2017, pp. 6–15.
[21] V. Kumar and M. L. Garg, "Comparison of machine learning models in student result prediction," in *Proc. Int. Conf. Adv. Comput. Netw. Inform.* Singapore: Springer, 2019, pp. 439–452.

[22] D. Kabakchieva, "Predicting student performance by using data mining methods for classification," *Cybern. Inf. Technol.*, vol. 13, no. 1, pp. 61–72, Mar. 2013.
[23] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," in *Proc. 5th Annu. Future Bus. Technol. Conf.*, A. Brito and J. Teixeira, Eds. Porto, Portugal: 2008, pp. 5–12.
[24] M. Hussain, W. Zhu, W. Zhang, S. M. R. Abidi, and S. Ali, "Using machine learning to predict student difficulties from learning session data," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 381–407, Jun. 2019.
[25] Y.-H. Hu, C.-L. Lo, and S.-P. Shih, "Developing early warning systems to predict students' online learning performance," *Comput. Hum. Behav.*, vol. 36, pp. 469–478, Jul. 2014.
[26] A. Mueen, B. Zafar, and U. Manzoor, "Modeling and predicting students' academic performance using data mining techniques," *Int. J. Mod. Educ. Comput. Sci.*, vol. 8, no. 11, p. 36, 2016.
[27] D. Thammasiri, D. Delen, P. Meesad, and N. Kasap, "A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition," *Expert Syst. Appl.*, vol. 41, no. 2, pp. 321–330, Feb. 2014.
[28] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data preprocessing for supervised leaning," *Int. J. Comput. Sci.*, vol. 1, no. 2, pp. 111–117, 2006.
[29] R. Longadge and S. Dongre, "Class imbalance problem in data mining review," 2013, *arXiv:1305.1707*. [Online]. Available: http://arxiv.org/abs/1305.1707
[30] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS Int. Trans. Comput. Sci. Eng.*, vol. 30, no. 1, pp. 25–36, 2006.
[31] B. W. Yap, K. A. Rani, H. A. A. Rahman, S. Fong, Z. Khairudin, and N. N. Abdullah, "An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets," in *Proc. 1st Int. Conf. Adv. Data Inf. Eng. (DaEng)*. Singapore: Springer, 2013, pp. 13–22.
[32] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 39, no. 2, pp. 539–550, Apr. 2009.
[33] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 4, pp. 463–484, Jul. 2012.
[34] H. Li, J. Li, P.-C. Chang, and J. Sun, "Parametric prediction on default risk of Chinese listed tourism companies by using random oversampling, isomap, and locally linear embeddings on imbalanced samples," *Int. J. Hospitality Manage.*, vol. 35, pp. 141–151, Dec. 2013.
[35] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[36] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Proc. Int. Conf. Intell. Comput.* Berlin, Germany: Springer, 2005, pp. 878–887.
[37] Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification," *IEEE Trans. Syst., Man, Cybern. B. Cybern.*, vol. 39, no. 1, pp. 281–288, Feb. 2009.
[38] D. Guan, W. Yuan, Y.-K. Lee, and S. Lee, "Nearest neighbor editing aided by unlabeled data," *Inf. Sci.*, vol. 179, no. 13, pp. 2273–2282, Jun. 2009.
[39] T. Elhassan and M. Aljurf, "Classification of imbalance data using Tomek link (T-link) combined with random under-sampling (RUS) as a data reduction method," *J. Informat. Data Min.*, vol. 2, no. 2, pp. 1–12, 2016.
[40] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 20–29, 2004.
[41] S. Aksoy and R. M. Haralick, "Feature normalization and likelihood-based similarity measures for image retrieval," *Pattern Recognit. Lett.*, vol. 22, no. 5, pp. 563–582, Apr. 2001.
[42] K. K. Pal and K. S. Sudeep, "Preprocessing for image classification by convolutional neural networks," in *Proc. IEEE Int. Conf. Recent Trends Electron., Inf. Commun. Technol. (RTEICT)*, May 2016, pp. 1778–1781.
[43] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: A tutorial," *Computer*, vol. 29, no. 3, pp. 31–44, Mar. 1996.
[44] G. Zhang, B. E. Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks:: The state of the art," *Int. J. Forecasting*, vol. 14, pp. 35–62, Mar. 1998.
[45] P. Cunningham and S. J. Delany, "k-Nearest neighbour classifiers," *Multiple Classifier Syst.*, vol. 34, pp. 1–17, Mar. 2007.
[46] J. Wang, P. Neskovic, and L. N. Cooper, "Improving nearest neighbor rule with a simple adaptive distance measure," *Pattern Recognit. Lett.*, vol. 28, no. 2, pp. 207–213, 2007.
[47] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: A tutorial," *Computer*, vol. 29, no. 3, pp. 31–44, Mar. 1996.
[48] I. A. Basheer and M. Hajmeer, "Artificial neural networks: Fundamentals, computing, design, and application," *J. Microbiol. Methods*, vol. 43, no. 1, pp. 3–31, Dec. 2000.
[49] M. Zięba, S. K. Tomczak, and J. M. Tomczak, "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction," *Expert Syst. Appl.*, vol. 58, pp. 93–101, Oct. 2016.
[50] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *J.-Jpn. Soc. Artif. Intell.*, vol. 14, nos. 771–780, p. 1612, 1999.
[51] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Dept. Comput. Sci. Inf. Eng., Univ. Nat. Taiwan, Taipei, Taiwan, Tech. Rep., 2003, pp. 1–12.
[52] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, 1999.
[53] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Trans. Syst., Man, Cybern.*, vol. 21, no. 3, pp. 660–674, 1991.
[54] W. Du and Z. Zhan, "Building decision tree classifier on private data," in *Proc. IEEE Int. Conf. Privacy, Secur. Data Mining*, vol. 14. Darlinghurst, NSW, Australia: Australian Computer Society, 2002, pp. 1–8.
[55] D. R. Cox, "The regression analysis of binary sequences," *J. Roy. Stat. Soc., B, Methodol.*, vol. 20, no. 2, pp. 215–232, Jul. 1958.
[56] D. W. Hosmer, Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. Hoboken, NJ, USA: Wiley, 2013.
[57] I. Rish, "An empirical study of the naive Bayes classifier," in *Proc. Workshop Empirical Methods Artif. Intell.*, vol. 3, no. 22, pp. 41–46, 2001.
[58] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Jan. 2006.
[59] R. A. Fisher, *Statistical Methods and Scientific Inference*. Oxford, U.K.: Hafner Publishing Co, 1956.
[60] H. W. Lilliefors, "On the Kolmogorov-Smirnov test for normality with mean and variance unknown," *J. Amer. Stat. Assoc.*, vol. 62, no. 318, pp. 399–402, Jun. 1967.
[61] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. Amer. Stat. Assoc.*, vol. 32, no. 200, pp. 675–701, Dec. 1937.
[62] M. Friedman, "A comparison of alternative tests of significance for the problem of *m* rankings," *Ann. Math. Statist.*, vol. 11, no. 1, pp. 86–92, Mar. 1940.

RAMIN GHORBANI received the B.S. degree in industrial engineering from Yazd University, Iran, in 2017, and the M.S. degree in system optimization (data mining in healthcare) from the Iran University of Science and Technology, Iran, in 2019. Since 2018, he has been a Research Assistant. His research interests include artificial intelligence, machine learning, and data analysis, with a passion for health informatics and bioinformatics.

ROUZBEH GHOUSI received the Ph.D. degree from the Iran University of Science and Technology, in 2013. He initiated his work as a Faculty Member with SIE, since 2015. He is currently an Assistant Professor of industrial engineering with the Iran University of Science and Technology. His research vision is concentrated mainly on safety and healthcare engineering, human reliability, sustainable supply chain management, and data science. He teaches human reliability analysis, diagnosis of production systems and services, human factor engineering, health, safety, and environment management system, time, and work-study.