Predicting Students Performance Using Supervised Machine Learning Based on Imbalanced Dataset and Wrapper Feature Selection

Sadri Alija [1], Edmond Beqiri [2*], Alaa Sahl Gaafar [3], Alaa Khalaf Hamoud [4]
[1] Faculty of Business and Economics, South East European University, North Macedonia.
[2] University of Peja "Haxhi Zeka" – Peja, Kosovo.
[3] Department of Educational Planning, Directorate of Education in Basrah, Iraq.
[4] Department of Computer Information Systems, University of Basrah, Iraq.
Email: s.aliji@seeu.edu.mk, edmond.beqiri@unhz.eu, alaasy.2040@gmail.com, alaa.hamoud@uobasrah.edu.iq.

Keywords: supervised machine learning, feature selection, wrapper, particle swarm optimization, info gain, SMOTE

Received: November 14, 2022

For learning environments like schools and colleges, predicting the performance of students is one of the most crucial topics since it aids in the creation of practical systems that, among other things, promote academic performance and prevent dropout. The decision-makers and stakeholders in educational institutions always seek tools that help in predicting the number of failed courses for the students. These tools can help in finding and investigating the factors that led to this failure. In this paper, many supervised machine learning algorithms will investigate finding and exploring the optimal algorithm for predicting the number of failed courses of students. An imbalanced dataset will be handled with Synthetic Minority Oversampling TEchinque (SMOTE) to get an equal representation of the final class. Two feature selection approaches will be implemented to find the best approach that produces a highly accurate prediction. Wrapper with Particle Swarm Optimization (SPO) will be applied to find the optimal subset of features, and Info Gain with ranker to get the most correlated individual features to the final class. Many supervised algorithms will be implemented such as (Naïve Bayes, Random Forest, Random Tree, C4.5, LMT, Logistic, and Sequential Minimal Optimization algorithm (SMO)). The findings show that the wrapper filter with SPO-based SMOTE outperforms the Info-Gain filter with SMOTE and improves the performance of the algorithms. Random Forest outperforms the other supervised machine learning algorithms with (85.6%) in TP average rate and Recall, and (96.7%) in ROC curve.

Povzetek: Opisana je metoda za napovedovanje uspeha študentov s pomočjo strojnega učenja.

1 Introduction

High-quality universities always require a great record of their students and the students are the main resource for them. The main concern for the universities is the performance of the students which is the base stone for building the top rate graduates and post-graduate students who will be the leaders of the nations and take responsibility of the economic and social growth of the society. Moreover, the main concerns for market employers are the performance of universities and students' academic performance due to its direct effect on the employment process and then employee productivity. So, the employers' demands are met by the graduated students who exert efforts in their academic journey. Student performance is measured by the learning assessment and the curriculum according to Usamah et al [1].

It is frequently important to be able to predict the behavior of future students to enhance the design of the curriculum and prepare the interventions for academic guidance and support. Machine learning (ML) is useful in this situation. ML approaches examine datasets, extract information, and then organize that information for eventual use. The primary goals of ML are to identify and extract patterns from recorded data by using a variety of techniques and algorithms [2]. Numerous algorithms exist and are used with educational data, including supervised algorithms such as Decision Tree (DT) and Naive Bayes (NB), and unsupervised algorithms such as K-Nearest Neighbor (KNN), and Neural Network (NN). Such algorithms forecast patterns, upcoming trends, and behaviors, enabling businesses to make informed, proactive decisions mining. This paper's major goal is to predict student performance using Supervised ML based on an imbalanced dataset and wrapper feature selection. The following section sheds light on related previous studies, then followed by the methodology and the concluded points, and future work.

2 Literature review

High quality universities always require the great record of their students where the students are the main resource for them. The main concern for the universities is the performance of the students which is the base stone for building the top rate graduates and post-graduates students who will be the leaders of the nations and take the responsibilities of the economic and social growth of the

The concept of data mining techniques can be implemented and applied in the educational field to improve our comprehension of the learning process, with a particular emphasis on the identification, extraction, and evaluation of factors linked to students' learning processes [3]. ML algorithms enable users to categorize and summarize associations discovered throughout the mining process as well as examine data from different perspectives. Bhardwaj and Pal in [4] explore the performance of the students by taking a sample of 300 undergraduate students' row records from the department of computer application from different institutions in Dr. R. M. L. Awadh University, India. The Bayesian classifiers are utilized on 17 features where the researchers found that there is a strong correlation between student action and other factors such as (living location, the academic background of the mother, senior secondary exam, the status, and the annual outcome of the student's family).

Next, in the same university, Pandey and Pal [5] selected 600 students to implement the model based on Bayes classifier to classify the background qualification, category, and language. While Hijazi and Naqvi in [6] have selected 300 students (75 female, and 225 male) from different colleges in Pakistan's Punjab University to explore and investigate student performance. Based on the linear regression, they found that there are many factors that affected the student's performance such as the attitude toward the class they attend, the time spent in studying after college, the mothers' ages, the income of their families, and the educational level of their mothers (where the performance is strongly affected by it). Khan in [7], explored the performance by building a model based on a clustering approach using 400 rows of student data from Aligarh Muslim University's senior secondary school in Aligarh, India. The main goal of the study is to determine the predictive value of different measures such as personality, cognition, and demographic variables that affect success at a higher level of secondary school. The outcomes of the study found that females with socioeconomic status scored higher performance, whereas males with low socioeconomics had higher performance in the science stream.

In the next case study [8], Kovacic implemented a data mining model for determining the educational enrollment data in New Zealand to predict the performance of the students. Chi-square automatic interaction detection (CHAID) and Classification and Regression (CART) algorithms are utilized to categorize the successful and failed students. The algorithms did not produce promising accuracies where they predicted the results with (59.4, and 60.5 respectively). The other case study is implemented by Galit [9] where the learning behavior is examined to predict the students' outcomes and alert the students to the critical status before the final exam. The final study [10] is proposed by Al-Radaideh, where the model is implemented to predict the students' final grades in C++ course for the students enrolled in the Yarmouk university in 2005, in Jordan. NB, DT (ID3, and C4.5) are utilized to predict the grades where the DT has outperformed the NB in prediction.

In our proposed model, the problem of imbalanced dataset is handled and the effect of handing this problem is observed by implementing different machine learning algorithms (supervised and unsupervised). The effect of handling imbalanced dataset is also observed by implementing feature selection which has the direct effect on the result accuracies.

3 Methodology

The model implementation framework is depicted in Figure 1, which consists of five steps starting with data preprocessing and ending with the model evaluation. The step of attribute feature selection (FS) is implemented by a single FS and a subset FS to find the effect of each step on the result accuracies. SMOTE filter is applied then, where it is followed by implementing supervised ML algorithms.

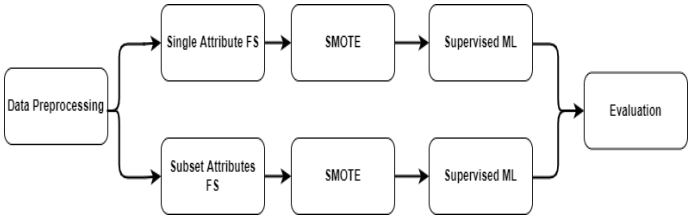

Figure 1: Model framework

3.1 Dataset reliability

A questionnaire is adopted in this study to build the model where Google Forms is used to build the questionnaire and collect undergraduate students' answers from both of Faculty of Contemporary Sciences and Technologies (CST) and the Faculty of Business and Economics (FBE) in South East European University (SEEU) in North Macedonia (RNM). The aim of this study is to find the optimal DT in predicting student performance based on the conceptual framework that was implemented by researchers in [11]. The aim of the framework is to find the hidden patterns that may affect and correlate with the performance of the students and provide suggestions to enhance and improve the performance. Many questions related to many factors are found in the questionnaire, such as academic behavior, health, finance, time planning, self-development, social relationships, and achieving goals. The questionnaire in [11] lists the factors and the questions related to each question, where the answer for most of the questions was on a 5-point Likert scale (from

1 to 5) which represented the formal answers (from "Strongly Disagree" to "Strongly Agree").
The dataset of the questionnaire involves 141 rows of respondents. The dataset reliability is required to measure the overall consistency of the dataset. The measure of reliability which describes consistency can be confirmed to have a high level if it produces similar results under consistent conditions. The most frequent measure in statistics is the coefficient alpha, which is used to calculate the internal consistency of the independent variables of the study. The coefficient's alpha for the dataset is 0.93. This value indicates an excellent internal consistency of the dataset reliability [12][13]. The applied tool for this model is Weka 3.8.5 and the system specifications are (RAM 8GB, HARD 35.5GB free, OS Win7 Pro).

Table 1: Dataset reliability

Number of Respondents	Number of Features	Coefficient's Alpha	% of Respondents
141	58	0.93	100%

3.2 Feature selection (FS)

FS approach can be considered as a form of data reduction where features are reduced and only the correlated features remain. The main goal of FS methods is finding the optimal subset of features or the highly correlated features that have a direct effect or may affect the final class(s). Due to the number of attributes in our dataset (57), it is required to find the most correlated attributes or features that can be utilized in the next steps to get more accurate results in classification [14]. Two approaches are applied in our model (Wrapper with Particle Swarm Optimization (PSO)) and (Info-Gain Attribute Evaluator).

- **Wrapper method**

The Wrapper method evaluates the subset of attributes according to the classifier performance for both supervised algorithms (such as DT, SVM, and NB) and unsupervised algorithms (such as clustering). For each subset, the evaluation process is repeated while the search strategy determines the subset generation. The wrapper method is slower than the filter in finding good subsets because it depends on resource demands for the algorithm of modeling. Due to using real modeling algorithms, the wrapper method is proven empirically to produce better feature subsets [15].

- **Particle swarm optimization (PSO)**

Kennedy and Eberhart in 1995 proposed one of the evolutionary computation techniques based on social behavior such as fish schooling and bird flocking. The basic idea behind PSO underlines that the population-social interaction optimizes knowledge where the thinking is personal and social. The solutions are represented by particles, while particles are represented by vectors that have positions in the search space. Each vector xi=(xi1,xi2,…xiD) Where D is the search space dimensionality. To search for the optimal solutions, the particles move in the search space. According to that, each particle has a velocity that can be represented by vi where vi takes the values (vi1,vi2,….,viD). The particle updates its location and velocity during the movement, and this update is performed according to the neighbors and their own experience. Two values of positions are recorded, the best which represents the best previous personal position of the particle, and gbest is the best-obtained position by the population. The following equation is used to update the position and velocity:

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1} \quad (1)$$

$$v_{id}^{t+1} = w * v_{id}^{t} + c_1 * r_1 * (p_{id} - x_{id}^{t}) + c_2 * r_2 * (p_{gd} - x_{id}^{t}) \quad (2)$$

Where t is the tth iteration in the evolutionary process while d represents the d dimension in the search space where d belongs to D. The weight w it controls the previous velocity impact on the current velocity impact. The acceleration constants c_1, c_2 are random values in the range (0 to 1), p_{id} and p_{gd} represent the elements of *pbest*, *gbest* alternatively in the dimension dth. *vmax* is the maximum velocity where $v_{id}^{t+1} \in$ [−*vmax*, *vmax*]. The algorithm will stop when the predefined criterion of fitness is met with a good fitness value or a predefined number of maximum iterations [16][17].

- **Info gain**

In this feature selection evaluator, the information of each class is estimated to evaluate the attribute. The method used in this evaluator is minimum description-length-based discretization where the attributes are binarized or discretized. In this method, the missing values are either regarded as separate values or distributed the values among other values according to the frequencies. As the value of the feature is absent, the decrease in entropy is measured. For the multiclass attribute, the InfoGain evaluator has reported the best performance. The generalized form of the nominal values is taken from the nominal attribute. Info Gain is measured by the decrease of X entropy that is caused by Y which is represented by:

$$IG(X|Y) = H(X) - H(X|Y) \quad (3)$$

According to this measurement, (Y) feature can be considered as more correlated to (X) feature if (IG(X | Y) > IG(Z | Y). IG normalized the values that fall within the range (0 to1), where (1) value indicates that the predicted value is completely correct and (0) value indicates that (X) feature is independent of (Y) feature. For the nominal and continuous features, the Entropy can be applied in order to determine the correlation between continuous and nominal features [18][19][20][21].
The Wrapper filter with SPO is applied to find and explore the most correlated subsets of features that make the highly accurate results for each supervised algorithm. Wrapper as a subset of attributes evaluator is applied for each supervised classifier individually. In this step, different subsets of features are found for each classifier where the SPO is selected as a search method to improve the speed of search for features subsets. In order to find

the effect of wrapper evaluator, Info Gain evaluator is applied to find the features with high correlations with the final class and to find how wrapper and Info Gain affect the result algorithms accuracies of the algorithms. Table 2 shows the most correlated features (subset) after applying wrapper with SPO for each algorithm and Info Gain with Ranker.

Table 2: Selection of attributes

Feature Evaluator	Attributes
Wrapper (Random Forest) with SPO	1,5,6,7,8,9,10,12,13,14,16,17,18,27,33,36,44,49,52,53,56
Wrapper (NB) with SPO	5,8,14,18,25,31,42,48
Wrapper (Logistic) with SPO	2,4,5,6,11,13,17,31,35,48,51,52,53,54,57
Wrapper (SimpleLogistic) with SPO	1,4,5,6,8,9,11,15,17,23,26,27,28,31,32,34,42,44,46,50,52,53,55
Wrapper (SMO) with SPO	4,5,14,15,17,24,31,32,35,42,45,47,54,55,56,57
Wrapper (LMT) with SPO	1,2,4,5,6,7,8,9,11,14,15,17,19,20,21,23,25,26,27,28,32,34,41,42,44,49,52,53,55
Wrapper (J48) with SPO	5,7,13,22,23,24,26,31,35,42,45,46,52
Wrapper (Random Tree) with SPO	5,15,27,33,35,43,44,45,46,48,49
Info Gain with Ranker	5,57,19,18,21,17,20,22,15,23,26,25,24,16,14,28,7,4,3,2,6

3.3 Synthetic minority over-sampling technique (SMOTE)

The dataset is said to be imbalanced if the classes in the final class are not equally represented [22]. If the final class has the classes (1,2, and 3) and the representations of the classes are (10% for 1, 15% for 2 and 75% for 3) then the dataset is imbalanced. The imbalanced datasets are found in almost all sectors starting from the medical sector [23], telecommunications management [24], fraudulent telephone calls [25], and text classification [26]. The SMOTE approach creates "synthetic" examples, to oversample the minority classes or by replacing the samples. This approach has been inspired and proven its success by the recognition process of handwritten characters [27]. The generation of synthetic examples is performed based on the operating in the feature space rather than the data space. The data space will face certain operations to generate the training data. The process of oversampling is performed by taking each minority class attribute of the final class attribute and introducing new examples (synthetic) along the line segments which join all k classes if they are nearest neighbors. The selection of the k nearest neighbors is performed randomly according to the oversampling amount required. The synthetic samples generation is implemented by taking the difference between each sample with its neighbors, then the result difference is multiplied by a random number between 0 and 1; then the result obtained is added to the feature vector. This process effectively forces to make the minority class more generally, see Figure 2 [28].

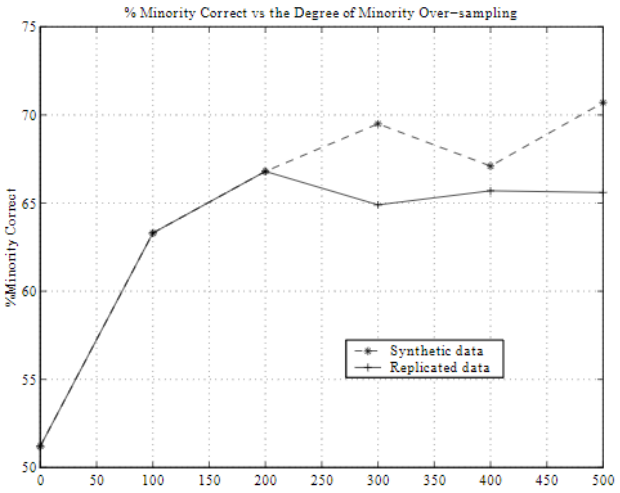

Figure 2: Comparison of number of minority correct for replicated oversampling and SMOTE for a dataset [28].

In our imbalanced dataset, the percentage of classes' representation is shown in Table 3. Class (3) takes only (4.3%) of the overall dataset, followed by classes (1, and 2) respectively with (21.3%, and 21.9%). The SMOTE filter in our model will be implemented on the classes (1,2, and 3) to make the dataset balanced and to get reliable performances of the algorithms. The SMOTE filter is applied to get equal representations of all classes.

Table 3: Classes representation

Class	Number of Rows	% of Representation
0	74	52.5%
1	30	21.3%
2	31	21.9%
3	6	4.3%

3.4 Supervised machine learning (ML)

In the proposed model, many supervised ML algorithms have been implemented to find the accurate algorithm for predicting the number of failed courses for the students. The algorithms fall in approaches such as (decision tree (DT) (Random Forest, Random Tree, LMT, and J48), naïve Bayes (Naïve Bayes, and Bayes Net), Logistic (Logistic and Simple Logistic), and Support Vector Machine (SMO)). DT is one of the supervised ML approaches that aim to build a training model to be used in predicting the final class attribute [29]. DT classifiers are widely used in different sectors and have proved their accuracies in the fields of education [11], [30][31], healthcare [32], wireless sensor networks [33], image processing [34][35], and disaster management [36][37]. There are many types of algorithms and the most used algorithms are (Random Forest, CART, Iterative Dichotomies 3 (ID3), and Successor of ID3 (C4.5 or J48) [38][39]. DT is used in the field of classification (predicting the categorical values) and regression (predicting the continuous values) [40]. Random Forest (which was proposed by L. Breiman in 2001) is a general-purpose regression and classification approach that works on the principle of aggregating the predictions by calculating the predictions averages and shows excellent

performance when the variables numbers is larger than the number of the observations [41]. In logistic model trees (LMT), logistic regression is utilized to select the attributes in a natural way by using stage-wise fitting. The logistic model in this approach is built on leaves by refining the leaves incrementally at the higher level of the tree [42].

SVM is an ML algorithm that falls under the supervised learning algorithm [43], as it is one of the data-based algorithms used to solve classification problems. It is considered one of the most important algorithms to accomplish that task (solving classification problems) [44]. Support Vector Machine has a vector support processing approach in which many questions are answered depending on the understanding and knowledge of the problem and how to design it. Moving to the real world, we find that the Support Vector Machine algorithm was used to find solutions to many problems in this world, including face recognition, detection, hand lines, and others [45]. In order to understand the SVM algorithm, it is necessary to understand its main terminology, the maximum-margin hyperplane, the separating hyperplane, the soft margin, and the kernel function [43]. SVM can be classified into two types: Linear SVM, and Non-Linear SVM. Linear SVM is an algorithm used when the data can be separated into two groups in a linear way by using a straight line where the data can be called as linearly separable, in addition to that the classifier is described as SVM classifier. Non-Linear SVM is an algorithm used when the data cannot be separated in a linear manner, and thus a straight line cannot be used to separate the data into two categories. To compensate for this, another thing called the kernel trick is used, through which we define the data in a higher dimension to be separated using some mathematical functions.

Regression is considered a simple type of ML algorithm. It is considered a supervised learning algorithm. These algorithms are used in a wide range to find a relationship between the continuous predictor and response variables. It is considered a way to measure the relationships between response variables and continuous predictors [46]. An example of this is the linear regression algorithm, which is one of the supervised learning algorithms, where this algorithm simulates the mathematical relationship between variables. It attempts to find relationships between independent variables (input data) and dependent variables (result, and forecast). It works to find continuous or numerical variables by predicting that as it assumes that the relationships between the predicted variables and the goal to be reached are linear, such as sales, age, and product price. The regression may be linear or curvilinear, so it must pass through all data points to reach the target prediction so that if the measurement is made between the data points and the regression line, the result is minimal.

In order to solve classification problems, a logistic regression algorithm was built, which is one of the supervised learning algorithms, where the results are always binary, not devoid of one of the two values, either 1 or 0, success or failure, rain or no rain; its working principle is probability. Logistic regression is used in the analysis of binary outcomes, or as it is said that they are two-level, or whose levels are opposite [47]. A characteristic of logistic regression is that its predictions are deterministic and have the ability to adapt to multiple predictions. This is necessary for the analysis of observational data when adjustment is useful to avoid differences in the totals to be compared [48]. Logistic regression is used to reach the highest weighting of a variable in the event that there is more than one variable. Thus, it is similar in terms of multiple linear regression and is inconsistent with it that the response variable has only a binomial, and as a result, each variable is considered to have an impact on the likelihood ratio of any expected event. Hence, it has the advantage that it can avoid confusing influences by analyzing the correlations of all variables at the same time [49].

NB is considered one of the supervisor learning algorithms; it is based on Bayes' rule together with additional to strong assumptions attributes that are categorically and conditionally independent [50]. Then it is used for solving classification problems. This algorithm assumes conditional independence of traits; so it is rarely true in the real world, which has made the competitive performance of this algorithm a lot of attention and surprising [51]. The Naïve Bayes algorithm is used in a wide range of applications, including article classification and spam filtering. Naïve Bayes Classifier is able to build ML model through which we get fast predictions. The hypothesis states that the independence between every two features, so the naïve Bayes classifier calculates the probability of belonging to a certain class. As a product of simple probabilities resulting from assumed Naïve independence. The hypothesis states that there is independence between each of the two features, so the Naïve Bayes classifier computes the probability of a particular instance belonging to a particular class. If we assume that the described is described by a vector x of attributes and the target of the class is the element y, then we can express the conditional probability $p(y|x)$ as the product of the simple probabilities resulting from the assumed naïve Bayes independence [52].

Bayesian networks are considered probabilistic models that depend mainly on non-periodic direct graphs. These models are causal relationships between their variables, and their structure represents the combination of previous knowledge and target data. They are also called belief networks as they belong to probabilistic graphical models, and knowledge can be represented in uncertain domains through the use of their graphical structures. It is observed by looking at its graphs, where nodes represent random variables, while arrows between nodes (variables) represent probabilistic dependencies. In most cases, generally accepted statistical methods are used to estimate these conditional dependencies. Hence, we can say that Bayesian networks combine graphs and statistics as well as computer science and probability theory [53]. Also, Bayesian networks are used to perform causal logic and predict risks. In addition, there are many advantages if we compare it with the methods used in regression methods [54]. One of Bayes Network's products is the modeling language in addition to the inference algorithms associated with random domains. Experiments have proven a lot of

success when used in medium-sized applications. But if Bayesian networks are used in areas that are relatively complex or large domains, then these networks will use the task of modeling, which is somewhat similar to programming using logic circuits [55].

3.5 Model evaluation

The evaluation process of algorithms is performed based on the confusion matrix, see Figure 3. The class value of True Negative (TN) is the predicted class as (NO) and it is (NO), while the class value of False Positive (FP) is the class when it is predicted as (YES) and it is (NO). False Negative (FN) class value is the class when it is predicted as (NO) and it is (YES) while True Positive (TP) class value is the class when it is predicted as (YES) and it is (YES).

	Predicted (YES)	Predicted (NO)
Actual (YES)	True Positive (TP)	False Negative (FN)
Actual (NO)	False Positive (FP)	True Negative (TN)

Figure 3: Confusion matrix.

Based on the above matrix, the performance criteria are:

$$Sensevity\ or\ TP = \frac{TP}{TP+FN} \quad (4)$$

$$Specificity\ or\ FP = \frac{FP}{FP+TN} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

The sensitivity or recall is a measurement of the truly predicted cases and measures the relevance of TP with FN. The more the TP rate, the more accurate the predicted cases and the more accurate the classification algorithm. The specificity or FP rate is the false alarm rate that measures the incorrectly predicted cases. The more FP, the more predicted incorrect cases. The precision represents the relevant cases among the predicted cases [29]–[31].

Table 4: Algorithms performance after wrapper with SPO.

Algorithm	TP Rate	FP Rate	Precision	Recall
LMT	0.766	0.078	0.762	0.766
Random Forest	**0.856**	**0.049**	**0.857**	**0.856**
Random Tree	0.697	0.100	0.695	0.697
NB	0.717	0.094	0.711	0.717
Logistic	0.727	0.091	0.729	0.727
Simple Logistic	0.773	0.075	0.770	0.773
SMO	0.757	0.081	0.752	0.757
J48	0.796	0.067	0.796	0.796

Table 4 lists the performance evaluation of supervised algorithms after implementing Wrapper with SPO. The algorithms are implemented after removing the uncorrelated features where the Wrapper base classifier is the supervised algorithm. Then, the SMOTE filter is applied to get equal representations of classes for the final class. RF algorithm outperforms the other supervised algorithms with (85.6% in TP rate and Recall), (4.9% in FP rate) and (85.7%) in precision. C4.5 (J48) algorithm comes in the second rank with (79.6% in TP rate and Recall), (6.7% in FP rate), and (79.6%) in Precision. NB comes in the last rank with (71.7% in TP rate and Recall), (9.4%) in FP rate, and (71.1%) in Precision.

Table 5: Algorithms performance after info gain evaluator.

Algorithm	TP Rate	FP Rate	Precision	Recall
LMT	0.750	0.083	0.749	0.750
Random Forest	**0.836**	**0.054**	**0.835**	**0.836**
Random Tree	0.701	0.099	0.696	0.701
NB	0.678	0.107	0.678	0.678
Logistic	0.737	0.087	0.735	0.737
Simple Logistic	0.707	0.097	0.701	0.707
SMO	0.734	0.088	0.730	0.734
J48	0.753	0.082	0.750	0.753
Bayes Net	0.750	0.083	0.753	0.750

Table 5 depicts the performance criteria of supervised ML algorithms after implementing Info Gain. The algorithms are implemented after removing the uncorrelated features (36 features), then the SMOTE filter is applied to get equal representations of classes for the final class. RF algorithm outperforms the other supervised algorithms with (83.6% in TP rate and Recall), (5.4% in FP rate) and (83.5%) in precision. C4.5 (J48) algorithm comes in the second rank with (75.3% in TP rate and Recall), (8.2% in FP rate), and (75%) in Precision. NB comes in the last rank with (67.8% in TP rate and Recall), (10.7%) in FP rate, and (67.8%) in Precision.

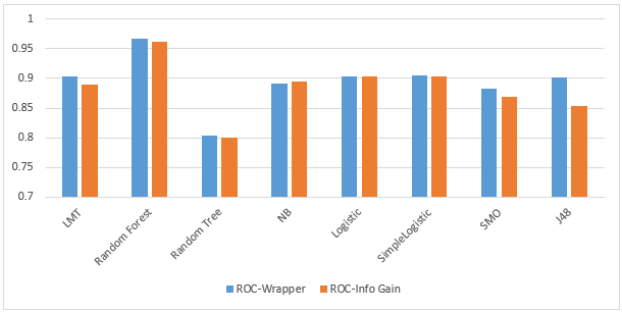

Figure 4: ROC of algorithms with wrapper and info gain.

One of the performance criteria that determines the optimal classifiers is the Receiver Operating Characteristic (ROC) curve, where ROC is considered one of the standard techniques that summarize classifier performance over a range of tradeoffs between TP and FP error rates [32][28]. As much as the ROC is closer to 1, as much as the classifier is accurate. Based on Figure 4, the RF classifier is the optimal classifier among all other classifiers with (96.7%) ROC when the wrapper with SPO

is implemented. The ROC is (96.1%) for the same classifier when Info Gain is implemented. The figure shows that ROCs for all algorithms are enhanced after implementing a wrapper evaluator with SPO. NB is the only classifier that has (89.1%) ROC when implementing wrapper and (89.5%) with Info Gain Evaluator.

4 Conclusions and future works

The imbalanced dataset faced many techniques and approaches to solve the minority and majority class problems related to the final class. In our model, the imbalanced dataset has multi-values in the final class which is required to handle this problem using SMOTE filter. In our model, the step of feature selection is performed two ways, the first one is by applying wrapper evaluator with SPO as a search method to find subsets of attributes that may affect and be correlated with the final class, and the second one by applying Info Gain as an evaluator with ranker as a search method to find the features with most correlation with the final class. After finding the most correlated features or feature subsets using evaluators, the uncorrelated features are removed and the SMOTE filter is applied to produce a balanced dataset and to make the multi-values classes equally represented. Many supervised ML algorithms are applied such as (NB, RF, Random Tree, LMT, J48, Logistic, Simple Logistic, and SMO). The performance evaluation of the algorithms shows that using the wrapper with the classifiers and SPO as a search method outperforms the Info-Gain evaluator. RF algorithm outperforms other algorithms in predicting students' performance and the number of failed courses. The model can be updated by predicting the students' status whether will fail or pass the final class. The features will be explored and investigated using different filters and classifiers to find the features with the most correlations with students' failure.

References

[1] U. Bin Mat, N. Buniyamin, P. M. Arsad, and R. A. Kassim, "An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention," in 2013 IEEE 5th International Conference on Engineering Education: Aligning Engineering Education with Industrial Needs for Nation Development, ICEED 2013, 126-130, 2014. https://doi.org/10.1109/iceed.2013.6908316.

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," AI Mag., vol. 17, no. 3, 37-37, 1996.

[3] A. El-Halees, "Mining Students Data To Analyze Learning Behavior : a Case Study Educational Systems," Work, 2008.

[4] A. B. E. D. Ahmed and I. S. Elaraby, "Data Mining: A prediction for performance improvement using classification," World J. Comput. Appl. Technol., vol. 2, no. 2, 2014. https://doi.org/10.13189/wjcat.2014.020203

[5] U. K. Pandey and S. Pal, "Data Mining: A prediction of performer or underperformer using classification," arXiv Prepr. arXiv1104.4163, 2011.

[6] S. M. M. Syed Tahir Hijazi & Raza Naqvi, "Factors affecting students' performance: A case of private colleges," Bangladesh e-Journal Sociol., vol. 3, no. 1, pp. 1–10, 2006.

[7] Z. N. Khan, "Scholastic Achievement of Higher Secondary Students in Science Stream," J. Soc. Sci., vol. 1, no. 2, 2005. https://doi.org/10.3844/jssp.2005.84.87

[8] Z. J. Kovacic, "Early Prediction of Student Success: Mining Students Enrolment Data," in Proceedings of the 2010 InSITE Conference, 2010. https://doi.org/10.28945/1281

[9] G. (Univ T. A. Ben-Zadok, R. (Univ T. A. Mintz, A. (Univ T. A. Hershkovitz, and R. (Univ T. A. Nachmias, "Examining online learning processes based on log files analysis: A case study," Res. Reflections Innov. Integr. ICT Educ. Proc. Fifth Intertnational Conf. Multimdeia ICT Educ., no. 2, 2009.

[10] Q. A. Al-Radaideh, E. M. Al-Shawakfa, and M. I. Al-Najjar, "Mining student data using decision trees," in International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan, 2006.

[11] A. K. Hamoud, A. S. Hashim, and W. A. Awadh, "Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis," Int. J. Interact. Multimed. Artif. Intell., 2018. https://doi.org/10.9781/ijimai.2018.02.004

[12] B. Carson, "The transformative power of action learning," Chief Learn. Off. Retrieved, 2017.

[13] U. Sekaran and R. Bougie, Research methods for business: A skill building approach. john wiley & sons, 2016.

[14] B. Remeseiro and V. Bolon-Canedo, "A review of feature selection methods in medical applications," Computers in Biology and Medicine, vol. 112. 2019. https://doi.org/10.1016/j.compbiomed.2019.103375

[15] Y. Kim, W. N. Street, and F. Menczer, "Evolutionary model selection in unsupervised learning," Intell. Data Anal., vol. 6, no. 6, 2002. https://doi.org/10.3233/ida-2002-6605

[16] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach," IEEE Trans. Cybern., vol. 43, no. 6, 2013. https://doi.org/10.1109/tsmcb.2012.2227469

[17] Y. Shi and R. Eberhart, "Modified particle swarm optimizer," in Proceedings of the IEEE Conference on Evolutionary Computation, ICEC, 1998. https://doi.org/10.1109/icec.1998.699146

[18] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," in Proceedings, Twentieth International Conference on Machine Learning, 2003, vol. 2.

[19] E. Frank, M. A. Hall, and I. H. Witten, "The WEKA Workbench Data Mining: Practical Machine Learning Tools and Techniques," Morgan Kaufmann, Fourth Ed., 2016. https://doi.org/10.1016/b978-0-12-374856-0.00010-9

[20] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in Proceedings of the 13th International Joint Conference on Artificial Intelligence, 1993.

[21] H. Liu, F. Hussain, C. L. Tan, and M. Dash, "Discretization: An enabling technique," Data Min. Knowl. Discov., vol. 6, no. 4, 2002.

[22] F. Provost and T. Fawcett, "Robust classification for imprecise environments," Mach. Learn., vol. 42, no. 3, 2001.

[23] A. S. Desuky, A. H. Omar, and N. M. Mostafa, "Boosting with crossover for improving imbalanced medical datasets classification," Bull. Electr. Eng. Informatics, vol. 10, no. 5, 2021. https://doi.org/10.11591/eei.v10i5.3121

[24] J. Xiao, L. Xie, C. He, and X. Jiang, "Dynamic classifier ensemble model for customer classification with imbalanced class distribution," Expert Syst. Appl., vol. 39, no. 3, 2012. https://doi.org/10.1016/j.eswa.2011.09.059

[25] C. Lu, S. Lin, X. Liu, and H. Shi, "Telecom fraud identification based on ADASYN and random forest," in 2020 5th International Conference on Computer and Communication Systems, ICCCS 2020, 2020. https://doi.org/10.1109/icccs49078.2020.9118521

[26] C. Padurariu and M. E. Breaban, "Dealing with data imbalance in text classification," in Procedia Computer Science, 2019, vol. 159. https://doi.org/10.1016/j.procs.2019.09.229

[27] T. M. Ha and H. Bunke, "Off-line, handwritten numeral recognition by perturbation method," IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 5, 1997. https://doi.org/10.1109/34.589216

[28] 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. 3 https://doi.org/10.1613/jair.953 21–357, 2002.

[29] M. A. Kumar and A. J. Laxmi, "Machine Learning Based Intentional Islanding Algorithm for DERs in Disaster Management," IEEE Access, vol. 9, 2021. https://doi.org/10.1109/access.2021.3087914

[30] A. K. Hamoud, "Selection of Best Decision Tree Algorithm for Prediction and Classification of Students' Action," Am. Int. J. Res. Sci. Technol. Eng. Math., vol. 16, no. 1, pp. 26–32, 2016.

[31] A. S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student performance prediction model based on supervised machine learning algorithms," in IOP Conference Series: Materials Science and Engineering, 2020, vol. 928, no. 3, p. 32019. https://doi.org/10.1088/1757-899x/928/3/032019

[32] T. Saba, I. Abunadi, M. N. Shahzad, and A. R. Khan, "Machine learning techniques to detect and forecast the daily total COVID-19 infected and deaths cases under different lockdown types," Microsc. Res. Tech., vol. 84, no. 7, 2021. https://doi.org/10.1002/jemt.23702

[33] I. A. Najm, A. K. Hamoud, J. Lloret, and I. Bosch, "Machine Learning Prediction Approach to Enhance Congestion Control in 5G IoT Environment," Electronics, vol. 8, no. 6, p. 607, May 2019. https://doi.org/10.3390/electronics8060607

[34] J. Chen, Y. Lian, and Y. Li, "Real-time grain impurity sensing for rice combine harvesters using image processing and decision-tree algorithm," Comput. Electron. Agric., vol. 175, 2020. https://doi.org/10.1016/j.compag.2020.105591

[35] I. S. Masad, A. Al-Fahoum, and I. Abu-Qasmieh, "Automated measurements of lumbar lordosis in T2-MR images using decision tree classifier and morphological image processing," Eng. Sci. Technol. an Int. J., vol. 22, no. 4, 2019. https://doi.org/10.1016/j.jestch.2019.03.002

[36] S. Khatoon et al., "Development of social media analytics system for emergency event detection and crisismanagement," Comput. Mater. Contin., vol. 68, no. 3, 2021. https://doi.org/10.32604/cmc.2021.017371

[37] H. Li, D. Caragea, C. Caragea, and N. Herndon, "Disaster response aided by tweet classification with a domain adaptation approach," J. Contingencies Cris. Manag., vol. 26, no. 1, 2018. https://doi.org/10.1111/1468-5973.12194

[38] Y. Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," Shanghai Arch. Psychiatry, vol. 27, no. 2, 2015.

[39] N. Mahdi Abdulkareem and A. Mohsin Abdulazeez, "Machine Learning Classification Based on Radom Forest Algorithm: A Review," Int. J. Sci. Bus., vol. 5, no. 2, 2021.

[40] S. M. Rasoolimanesh, M. Wang, J. L. Roldán, and P. Kunasekaran, "Are we in right path for mediation analysis? Reviewing the literature and proposing robust guidelines," J. Hosp. Tour. Manag., vol. 48, 2021. https://doi.org/10.1016/j.jhtm.2021.07.013

[41] G. Biau and E. Scornet, "A random forest guided tour," Test, vol. 25, no. 2, 2016. https://doi.org/10.1007/s11749-016-0481-7

[42] N. Landwehr, M. Hall, and E. Frank, "Logistic Model Trees," Mach. Learn., vol. 59, no. 1, pp. 161–205, 2005. https://doi.org/10.1007/s10994-005-0466-3

[43] W. S. Noble, "What is a support vector machine?" Nature Biotechnology, vol. 24, no. 12. 2006. https://doi.org/10.1038/nbt1206-1565

[44] T. Joachims, "Svmlight: Support vector machine," SVM-Light Support Vector Mach. http//svmlight.joachims.org/, Univ. Dortmund, vol. 19, no. 4, 1999.

[45] S. Ghosh, A. Dasgupta, and A. Swetapadma, "A study on support vector machine based linear and

non-linear pattern classification," in Proceedings of the International Conference on Intelligent Sustainable Systems, ICISS 2019, 2019. https://doi.org/10.1109/iss1.2019.8908018

[46] K. Park, R. Rothfeder, S. Petheram, F. Buaku, R. Ewing, and W. H. Greene, "Linear regression," in Basic Quantitative Research Methods for Urban Planners, 2020. https://doi.org/10.4324/9780429325021-12

[47] A. J. Scott, D. W. Hosmer, and S. Lemeshow, "Applied Logistic Regression.," Biometrics, vol. 47, no. 4, 1991. https://doi.org/10.2307/2532419

[48] B. R. Kirkwood and J. A. C. Sterne, Essential Medical Statistics. 2003.

[49] S. Sperandei, "Understanding logistic regression analysis," Biochem. Medica, vol. 24, no. 1, 2014. https://doi.org/10.11613/bm.2014.003

[50] G. I. Webb, E. Keogh, and R. Miikkulainen, "Naïve Bayes.," Encycl. Mach. Learn., vol. 15, pp. 713–714, 2010. https://doi.org/10.1007/978-0-387-30164-8_576

[51] H. Zhang, "The optimality of naive Bayes," Aa, vol. 1, no. 2, p. 3, 2004.

[52] W. Lou, X. Wang, F. Chen, Y. Chen, B. Jiang, and H. Zhang, "Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes," PLoS One, vol. 9, no. 1, p. e86703, 2014. https://doi.org/10.1371/journal.pone.0086703

[53] J. Pearl, "Bayesian networks," 2011.

[54] P. Arora, D. Boyne, J. J. Slater, A. Gupta, D. R. Brenner, and M. J. Druzdzel, "Bayesian networks for risk prediction using real-world data: a tool for precision medicine," Value Heal., vol. 22, no. 4, pp. 439–445, 2019. https://doi.org/10.1016/j.jval.2019.01.006

[55] D. Koller and A. Pfeffer, "Object-oriented Bayesian networks," arXiv Prepr. arXiv1302.1554, 2013.

[56] A. Khalaf et al., "Supervised Learning Algorithms in Educational Data Mining: A Systematic Review," Southeast Eur. J. Soft Comput., vol. 10, no. 1, pp. 55–70, 2021.