Indonesian Journal of Electrical Engineering and Computer Science
Vol. 22, No. 3, June 2021, pp. 1708~1715
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v22.i3.pp1708-1715

Prediction of student's performance through educational data mining techniques

Nibras Z. Salih, Walaa Khalaf
Department of Computer Engineering, Mustansiriyah University, Baghdad, Iraq

Article Info

Article history:

Received Mar 8, 2021
Revised May 14, 2021
Accepted May 19, 2021

Keywords:

Classification algorithms
Cross-validation
Imbalance datasets
Synthetic minority
Oversampling technique

ABSTRACT

Many educators have worried about the failures of students through academic education. Thus, a variety of predictions have been applied to general information including culture, social, and economic information which wasn't related to student performance. We have gathered an actual dataset from three years of academic stages of Mustansiriyah University in Iraq. The dataset consists of academic information without any socioeconomic data, it includes forty-four undergraduate students with thirteen attributes. We have proposed a model that explains the correlation between two main subjects which are, mathematics, and control systems. This study aimed to identify student failure of the control systems subject in the third year depending on the academic features of the mathematics subjects in the first and second years. Three algorithms were applied to the dataset including Naïve Bayes, support vector machine, and multilayer perceptron. Since the dataset was imbalanced, this leads to appear overfitting problem in the results so the synthetic minority oversampling technique was utilized to solve this problem. Our results show that the support vector machine algorithm proves an efficient classification after applied synthetic minority oversampling technique. The accuracy of the classifiers was measured from the confusion matrix using the Waikato environment for knowledge analysis (WEKA) tool and its related metrics.

Corresponding Author:

Nibras Z. Salih
Department of Computer Engineering
Mustansiriyah University
Baghdad, Iraq
Email: EGma020@uomustansiriyah.edu.iq, neranzezo@gmail.com

1. INTRODUCTION

The troubling indices of students' academics failure for educational courses at universities have been concerned educators. Therefore, researchers show that during their course tasks, students encounter many difficulties in a manner that many of them could be failing during their studies or leaving the academic courses [1]. The method of searching for essential data from an extensive dataset to evaluate the data from various destinations is known as data mining. Data mining technologies have been used to enhance education's performance by connecting a knowledge gap, predicting student performance, and identifying features [2]-[4].

Many techniques for data mining are used to evaluate a dataset and to find useful information called knowledge. Data mining was introduced in business sectors only, but it is proven to be applied for education and hence is defined as educational data mining (EDM) [5]. EDM is concerned with the execution of data mining strategies to inspect several academics data for knowledge derivation. Also, EDM explores new techniques that lead to a better understanding of the learning environment. The increasing attention in academic work is expanded to recognize key factors which influence the student's success. Therefore, educators should analyze students'

results to have a stronger means of studying. Moreover, the strategies of EDM are designed to provide additional advantages for educators to better student activities. The extraction from large datasets of essential patterns is known as knowledge discovery in databases (KDD) [5], [6].

One of the main challenges in this sense is the ability to predict students' failure during educational courses in the future at an efficient level so that pedagogical strategies can be performed to prevent student's failure [1], [7]-[10]. Therefore, classification is one of the most famous tools of data mining used to classify data to the groups so each element of the dataset is allocated to one class by assigning a label regarding its class. The idea behind this tool is to separate the dataset into the training set and test set, the training collection includes the data that is trained using a certain algorithm then used as a reference for this classification [11]. Mathematical methods were used to characterize the data by certain models working as a classifier, such as decision tree, Bayesian classification, neural networks, support vector machine, and classification based on association [6], [7].

Waikato environment for knowledge analysis (WEKA) has been considered a powerful classification tool used for educational data mining to analyze and evaluate the accuracy of various algorithms. WEKA is commonly implemented in machine learning depends on the Java language that involves various algorithms and different techniques as classification, clustering, and regression [3], [12]. In this paper, we have identified two main subjects which are, mathematics and control systems for five courses of academic stages. Also, we have proposed a model that explains the correlation between these subjects and illustrates that the results of the student in the first subject will affect his results in the second subject. This correlation depends on three academic features of the mathematics subject which are total lecture attendance, assessment grade, and final grade.

The goal of this paper is to predict students' success or failure in the control systems subject, and the reasons behind it. Therefore, finding these reasons will support the students to improve their performance, preparing the student to start a good semester, and understand the basic concepts that lead to a successful education course. Three classification models are applied to predict student performance: Naïve Bayes classifier (NBC), support vector machine (SVM), and multilayer perceptron (MLP). In this research, seven metrics are used to evaluate the efficiency of the selected classifiers: recall, precision, specificity, F-measure, receiver operating characteristics (ROC), precision-recall curve (PRC) area, and accuracy. The paper is divided into the following sections, section 2 presents related work, section 3 describes the data structure and the proposed method, section 4 shows experiment design, Section 5 illustrates the results and discussion while the conclusions are in section 6.

2. RELATED WORKS

Some works have analyzed the prediction of students' failure during their education courses. Costa *et al.* [1] discussed the results of the effective EDM approach to early detection for students who were expected to fail introductory courses through studying the impact of data pre-processing and fine-tuning algorithm tasks that concerning the efficiency of the classifiers. Two datasets were implemented which were distance studies data that consist of 262 students for 10 weeks and campus data that contain 161 students for 16 weeks. Hence, four algorithms were utilized on these datasets as decision tree, neural network, SVM, and NBC. The datasets consist of several features like age, town, student registration and semester, civil status, and discipline status. Costa *et al.* [1] found the SVM was an appropriate classifier with an F-measure of 92% for distance studies courses while with an F-measure of 83% for campus courses [1].

Ching-Chieh Kiu analyzed the significance and effect of student history, social behaviors of students, and the accomplishment of student coursework in forecasting student academic performance in the Mathematic subject. The dataset contains 395 instances and 33 attributes that were divided into three groups. The first group of the data was student background such as the job of father, guardian of a student, and gender of student. The second group was the social data such as home internet access, outing with friends, and the status of current health. And the third group was the coursework data that involves the grading period [7].

An early prediction of students fails who were expected not to succeed in an academic course was implemented by sentiment testing to detect effective data and to improve predictive accuracy. The emotional analysis was provided by a student's comments that identify an achievement student by graphic techniques. The dataset included 181 students of computer science subjects for nineteen weeks that involved completing the homework, participation in the class, attendance, and the student's emotion. SVM and convolution neural network (CNN) were applied through the data mining techniques. CNN was the best classifier with an F-measure of 0.78% during the 9th week [8]. Also, Asif *et al.* [9] used data mining techniques to evaluate students' undergraduate, since the authors focused on two factors of student performance which are the prediction of students' academic performance at the end of four years study program, to analyze and combine standard progressions with prediction outcomes Asif *et al.* [9] detected two student categories which were low and high student activities.

Including the research, Jacob *et al.* [10] implemented multiple techniques like regression and decision tree to effectively predict the students' performance and academic failure. The dataset was collected from computer engineering and applied various regression analyses and decision trees to predict the average grade and the final semester of the students. Thus, grouping was applied in learning styles by dividing students into academic

strength groups and weaknesses groups that depend on the student activities of programming language subject using k-mean algorithm [10].

P. Kaur, M. Singh, and G. S. Josan [12] recognized and displayed slow learners among students through a predictive data mining model. The student academic dataset was tested and implemented using five classifiers including multilayer perception, Naïve Bayes, J48, sequential minimum optimization (SMO), and REPTree. The dataset contains 152 students that were included academic and non-academic features such as student's gender, and the computer at home, medium of instruction, the student having a cell phone. P. Kaur, M. Singh, and G. S. Josan [12] indicated that the MLP was the appropriate classifier with an accuracy of 75%. Ahmed *et al.* [3] focus on the performance of the teacher and a discussion of the reasons that affect the performance of the students to enhance the efficiency of the education system by applying four classifiers, which are J48, NB, MLP, and SMO. The dataset was gathered from California University that was contained a 5,820-assessment score supplied by the student, it involves 28 specific questions with 5 attributes. Amjad Abu Saa [13] collected different datasets including a variety of personal, social, and academic data. The authors concluded that the achievement of the student is not completely dependent on their academic efforts, although several other variables have great influences. Four algorithms of the decision tree and the Naïve Bayes algorithm were applied. The datasets were obtained through an online survey conducted in Google forms and circulated to various students in their ordinary studies. It was included various features such as the student's gender, nationality category, and teaching language in the university.

Al-Shehri *et al*. [14] introduced the SVM and K-nearest neighbor classifiers on a student data collection to estimate the student grade in the final exam of Mathematic subject. The student was assessed for two periods and combined with the third period to achieve the final score. The dataset consists of 395 instances and 33 attributes, it was included different features such as family size, father's job, and mother's job. Hussain *et al*. [4] presented the student performance based on three various collages of Assam in India. The datasets consist of 300 students with twenty-four attributes of social, demographic, and academic. Four classification strategies were used in this research, including J48, PART, random forest, and Bayes network classifiers. The continuous appraisal process of the internal assessment variable had the greatest effect on the students' final semester outcomes. The dataset was included different features as father and mother qualifications, size of the family, and gender [4]. Most researchers of the above-mentioned works have used a large dataset concerning a general description of students that included social, demographic, personal, and socio-economy information that wasn't related directly to students' performance.

3. DATA DESCRIPTION

During the academic year 2019-2020, the dataset was collected from Mustansiriyah University in Iraq. Dataset has pertained depend on two main subjects which are mathematics and control systems subjects. Mathematics subject involves Calculus I and Calculus II from the first stage, Mathematical analysis I, and Mathematical analysis II from the second stage. Whereas, control systems subject (class) from the third stage, either 0 refers to pass or 1 refers to fail. One academic stage consists of two courses, each course of the mathematics subject has three academic features: The total lectures attendance, assessment grade, and the final grade. The total lecture attendance feature has a value of 0, 1, or 2 (0 indicates the last warning, 1 indicates the first warning, and 2 indicates no absence has been recorded). The assessment grade involves quizzes, mid-term exams, and student assignments. The maximum assessment grade value is 40 while the minimum grade value is 0. The final grade feature has a value of 0, 1, or 2 (0 indicates that the student hasn't passed either the first or the second attempt, 1 indicates that the student has failed in the first attempt but passed in the second, and 2 indicates that the student passed the exam from the first attempt. Twelve attributes description of mathematics subjects and one class attribute of the control systems subjects are shown in Table 1.

Table 1. Attributes description for two main subjects of mathematics and control systems

Stage	Attributes	Range of atributes	Description
First stage	Attendance course 1	[0,1,2]	0 – more than 10% absence 1 – less than 10% absence 2 – no absence
	Attendance course 2		
Second stage	Attendance course 1		
	Attendance course 2		
First stage	Assessment grade course 1	Quizzes [0-10] Mid-term exam [0-20] Assignments [0-10]	0 – refers to the lowest assessment grade 40 – refers to the highest assessment grade
	Assessment grade course 2		
Second stage	Assessment grade course 1		
	Assessment grade course 2		
First stage	Final grade course 1	[0,1,2]	0 – failed both attempts 1 – pass from the second attempt 2 – pass from the first attempt
	Final grade course 2		
Second stage	Final grade course 1		
	Final grade course 2		
Third stage	Final results course 1 (Class)	[0,1]	0 – pass and 1 - fail

Calculus I, and calculus II for the first stage, mathematical analysis I and mathematical analysis II for the second stage, and control systems of the first course for the third stage. Also Figure 1 shows the dataset representation in the WEKA tool using attribute relation file format (ARFF) extension

```
@relation student_data
@attribute std_id  numeric
@attribute Attendance-11 numeric
@attribute Attendance-12 numeric
@attribute Attendance-21 numeric
@attribute Attendance-22 numeric
@attribute Assessmentgrade-11 numeric
@attribute Finalgrade-11 numeric
@attribute Assessmentgrade-12 numeric
@attribute Finalgrade-12 numeric
@attribute Assessmentgrade-21 numeric
@attribute Finalgrade-21 numeric
@attribute Assessmentgrade-22 numeric
@attribute Finalgrade-22 numeric
@attribute class {0,1}
@data
1,2,2,2,2,14,0,20,2,28,2,17,2,0
2,2,2,2,2,34,2,22,2,24,2,19,2,0
```

Figure 1. Data representation using ARFF format

Dataset consists of 44 instances (students), 12 attributes, and one class attribute value for each instance. The lower classes have 17 instances known as 0 (pass) while the higher classes have 27 instances known as 1 (fail). We have constructed the dataset depends on the correlation between mathematics subjects from one side, and the control systems from the other side. We have suggested that these subjects share the basic concepts for three years of academic courses. The basic concepts that share in these subjects including analog and digital systems, Laplace, z-transform, optimization techniques, linear, and nonlinear systems. Consequently, the student's results in the first subject of academic courses will affect his results in the second subject.

Three classifiers are utilized to predict the student's result including NBC, SVM, and MLP. In the first phase, each classifier is applied to the student dataset using a specific technique in the WEKA tool such as leave one out cross validation (LOOCV), five-folds cross-validation (5-CV), or training set technique. Then, the dataset is classified into the training set and testing set in the classification technique. The training set includes data that has been trained with a particular classifier and used as a classification reference which can be implemented in the test set. While in the testing set, the undefined instances that are not labeled (without classes) are tested to predict student performance in the control systems subject.

Seven metrics are applied in the third phase to measure the accuracy of the model using the confusion matrix in the WEKA tool which are recall, precision, specificity, F-measure, ROC, PRC area, and accuracy. Finally, the final output results are presented based on the average results for each metric. Figure 2 shows the framework of the proposed system.

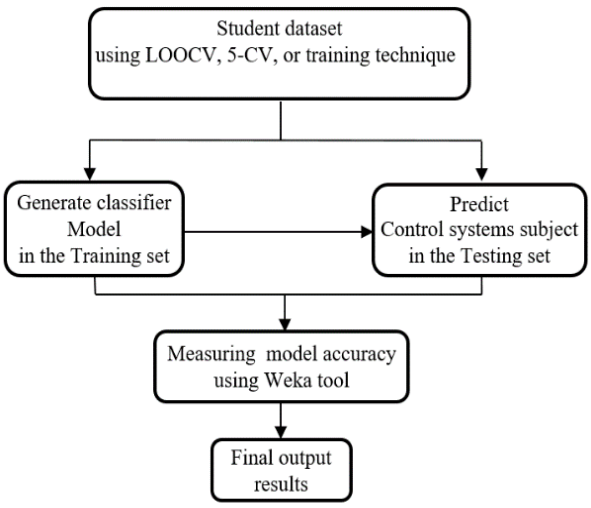

Figure 2. The framework of the proposed system structure

The confusion matrix consists of TP, TN, FP, and FN, where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives. Seven metrics are used to evaluate the efficiency of the selected classifier: Recall, precision, specificity, F-measure, ROC, PRC area, and accuracy [4], [15]. The recall described true positives states divided by positive states expressed as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (1)$$

The Precision identifies true positives states divided by expected positive states expressed as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

Specificity described true negatives states divided by the total number of negative states expressed as:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \quad (3)$$

F-measure is a mixture of precision and recall measurement expressed as:

$$\text{F-measure} = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (4)$$

PRC is the precision-recall curve used to evaluate the classifier performance for imbalanced and noisy datasets. The receiver operating characteristics (ROC) curve is used to evaluate the performance of the classifier includes two-axes: false positive rate on the x-axis and true positive rate (Recall) on the y-axis. Therefore, ROC and PRC curves are used to compare the performance model as a single metric. The ROC is used to realize the performance of a classifier on a balanced dataset at each class while the PRC represents the change of the precision with the recall for different thresholds of the imbalanced dataset [16], [17]. Accuracy is used to measure the performance of the classifier and described as true classification states divided by the total number of states expressed as:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (5)$$

Cross-validation aims to assess learning algorithms by dividing the data into two sets which are the training set and testing set. Also, it is used to compare the results of different classifiers. *k*-folds cross-validation (CV) is used to evaluate the performance of any classifier in machine learning. The data are randomly separated into *k*-folds, since the dataset is split into *k* equally folds, thereafter *k* iterations of the training and testing are carried out so that at every iteration a various fold of the dataset is kept for testing whereas the remaining (*k*-1) folds are applied for the training set. A particular situation of *k*-fold cross-validation is the leave one out cross-validation (LOOCV) where the number of folds is proportional to the total number of instances. Leave one out cross-validation has been used to evaluate the efficiency of any machine learning algorithm when the number of instances is limited [18], [19].

4. EXPERIMENT DESIGN

WEKA is implemented to classify the dataset since it's an open-source machine learning software used for data mining tasks [3], [12], [14]. WEKA includes many tools for data preprocessing, classification, and clustering. Three classification algorithms are selected from WEKA: NBC, MLP, and SVM to predict the performance of the student. Naïve Bayes classifier (NBC) is a simple supervised classification method that depends on a presumption of the class conditional independence. NBC is assumed that all attributes provided in a dataset are independent based on the Bayes rule of conditional probability [20].

Multilayer perceptron (MLP) is a supervised classifier implemented for neural network training and to classify instances by a backpropagation algorithm which uses a gradient descent technique for minimizing mean square error in the input vector through the desired and actual outputs. MLP comprises several layers of neurons and each neuron excluding the input neurons has activation functions where every layer is attached to the next layer [21]. Support vector machine (SVM) is a supervised machine learning algorithm that was applied by Vapnik, using for both classification, and regression. This classifier has the potential to minimize errors of the classifier and to maximize the graphical margin [22]. In this study, we have applied linear SVM to the dataset.

For the imbalanced and noisy dataset, an overfitting problem could appear which can be excluded by following some statistical technique [23]. In this paper, an imbalanced dataset is presented because the instances number of one class is smaller than the other one. The lower classes have 17 instances but the higher classes have 27 instances. The smaller class is called the minority class while the bigger class is called the majority class.

Synthetic minority oversampling technique (SMOTE) is the oversampling method used to solve the imbalanced dataset problem. SMOTE transforms an imbalanced dataset and produces balanced datasets. The majority and minority classes are distributed using SMOTE by generating synthetic instances in the minority class, this technique is used to enhance prediction performance in the minority class. In the minority class, the sample is positioned across the line segments that include one or more of the k-nearest neighbors. The SMOTE is usually used by five closest neighbors [24], [25].

Oversampling increases the number of occurrences to retaining both occurrences and non-occurrences by using sampling with replacement [26]. So, the two classes became similar when this technique was applied. SMOTE supervised filter has been implemented using two main parameters which are percentage and nearest neighbors in WEKA. The lower classes have increased by 50% (based on option-P 50.0 in WEKA) and adjusting nearest neighbors to obtain the best results (based on option-K in WEKA).

5. RESULT AND DISCUSSION

The relevant results are illustrated in Tables 2 and 3 for the average of the following seven metrics: Recall, precision, specificity, F-measure, ROC, PRC area, and accuracy. Three classifiers are applied which are NBC, MLP, and SVM to predict the performance of the student. In the first case, we perform this experiment using the training set technique to assess the classifier on how well the class of cases is trained to predict and then applied cross-validation techniques (LOOCV and 5-CV). Therefore, to understand the performance of each classifier, the results of the above-mentioned metrics for the training set, LOOCV, and five-times, 5-CV are shown in Table 2.

We compare the result of the predictive model for the training set, LOOCV, and 5-CV. We found that the performance of the classifier on the training set is better than LOOCV and 5-CV, which means that the predictive model suffers from overfitting, which led to a large difference between the results of the training set and cross-validation techniques. So SMOTE supervised filter is used to overcome the overfitting problem and to enhance the prediction performance.

Table 3 shows the results for three classifiers (NBC, MLP, and SVM) using LOOCV and five times, 5-CV after a supervised SMOTE filter is applied. The best results for this experiment have been underlined. We have noticed the SVM outperforming in terms of sensitivity, precision, F-measure, PRC area, ROC, and accuracy for LOOCV, and NBC outperforms in terms of sensitivity, specificity, F-measure, and accuracy for 5-CV. The low specificity values are usually a result of high sensitivity values, so the specificity plays an important role because it identifies the student's failure in the academic course.

Table 2. The classification results: using the training set, LOOCV, and 5-CV

	Using Training-set			LOOCV			5-CV (Mean ± std)		
Metrics	NBC	MLP	SVM	NBC	MLP	SVM	NBC	MLP	SVM
Sensitivity	0.773	0.977	0.864	0.659	0.636	0.659	0.636 ± 0.022	0.672 ± 0.012	0.636 ± 0.032
Specificity	0.740	0.962	0.888	0.740	0.703	0.777	0.695 ± 0.03	0.769 ± 0.016	0.74 ± 0.074
Precision	0.791	0.979	0.864	0.656	0.636	0.650	0.638 ± 0.02	0.666 ± 0.015	0.6312 ± 0.02
F-Measure	0.776	0.977	0.864	0.657	0.636	0.652	0.637 ± 0.021	0.668 ± 0.015	0.629 ± 0.031
ROC	0.852	0.969	0.856	0.590	0.643	0.624	0.603 ± 0.023	0.677 ± 0.012	0.606 ± 0.03
PRC Area	0.874	0.960	0.815	0.610	0.679	0.601	0.617 ± 0.015	0.705 ± 0.017	0.587 ± 0.02
Accuracy	77.27 %	97.72 %	86.36 %	65.90 %	63.63 %	65.90 %	63.636 ± 2.27	67.27 ± 1.24	63.636 ± 3.21

Table 3. The classification results using SMOTE filter

	LOOCV			5-CV (Mean ± std)		
Metrics	NBC	MLP	SVM	NBC	MLP	SVM
Sensitivity	0.712	0.769	0.788	0.738 ± 0.029	0.723 ± 0.022	0.738 ± 0.04
Specificity	0.666	0.777	0.740	0.703 ± 0.026	0.636 ± 0.017	0.673 ± 0.03
Precision	0.716	0.769	0.793	0.742 ± 0.03	0.734 ± 0.024	0.748 ± 0.05
F-Measure	0.711	0.769	0.788	0.738 ± 0.029	0.721 ± 0.022	0.737 ± 0.04
ROC	0.736	0.754	0.790	0.746 ± 0.023	0.774 ± 0.014	0.741 ± 0.04
PRC Area	0.703	0.717	0.730	0.720 ± 0.015	0.780 ± 0.018	0.69 ± 0.058
Accuracy	71.153 %	76.923 %	78.846 %	73.846 ± 2.91	72.30 ± 2.19	73.84 ± 4.42

6. CONCLUSION AND FUTURE WORK

In this article, the exploration aims to provide a prediction for study program managers and educators which might assist them in obtaining better educational programs at their academy. Predicting students' performance depending on marks and course attendance without any socioeconomic data. The datasets were collected from three studies years of academic stages of Mustansiriyah University in Iraq, it consists of 44 students and 13 attributes that included five courses. We have proposed a model that explains the correlation between two basic subjects which are mathematics of the first and second years and control systems of the third year. The study aims to improve student performance by analyzing academic features of mathematics courses to avoid student failure of the control systems course. Thus, this prediction leads to guide the student for improving their academic features of mathematics courses in the first and second years of their studies.

The results show that is the potential to predict the student's result of one subject in the university program to obtain good undergraduate marks with plausible accuracy. With the assistance of NBC, MLP, and SVM algorithms, three techniques are applied to the dataset including the training set, LOOCV, and 5-CV using the WEKA tool. We found the predictive model suffers from overfitting because an imbalanced dataset was utilized, therefore a supervised SMOTE approach is implemented to overcome this problem and to enhance the prediction of the students' performance. We conclude that the best classifier result has appeared after applied SMOTE technique which is the SVM for LOOCV. The future work is to enlarge the dataset to strengthen the generalizability of the prediction. Also, we will study the correlation between other subjects such as communication system and digital signal processing, and their impact on student's performance.

ACKNOWLEDGMENTS

My sincere appreciation and thanks to the University of Mustansiriyah for the guidance and support. Also, all thanks and appreciation to those who helped me and gave me scientific advice in this research.

REFERENCES

[1] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses," *Computers in Human Behavior*, vol. 73, pp. 247-256, August 2017, doi: 10.1016/j.chb.2017.01.047.

[2] A. I. Adekitan and O. Salau, "The impact of engineering students' performance in the first three years on their graduation result using educational data mining," *Heliyon*, vol. 5, no. 2, pp. e01250, February 2019, doi: 10.1016/j.heliyon.2019.e01250.

[3] A. M. Ahmed, A. Rizaner, and A. H. Ulusoy, "Using data mining to predict instructor performance," *Procedia Computer Science*, vol. 102, pp. 137-142, 2016, doi: 10.1016/j.procs.2016.09.380.

[4] S. Hussain, N. A. Dahan, F. M. Ba-Alwib, and N. Ribata, "Educational data mining and analysis of students' academic performance using WEKA," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 9, no. 2, pp. 447-459, February 2018, doi: 10.11591/ijeecs.v9.i2.pp447-459.

[5] N. Ketui, W. Wisomka, and K. Homjun, "Using Classification Data Mining Techniques for Students Performance Prediction," *2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON)*, 2019, pp. 359-363, doi: 10.1109/ECTI-NCON.2019.8692227.

[6] N. Kumar and S. Khatri, "Implementing WEKA for medical data classification and early disease prediction," *2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT)*, 2017, pp. 1-6, doi: 10.1109/CIACT.2017.7977277.

[7] Ching-Chieh Kiu, "Data Mining Analysis on Student's Academic Performance through Exploration of Student's Background and Social Activities," *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*, 2018, pp. 1-5, doi: 10.1109/ICACCAF.2018.8776809.

[8] L. C. Yu, *et al*., "Improving early prediction of academic failure using sentiment analysis on self-evaluated comments," *Journal of Computer Assisted Learning*, vol. 34, no. 4, pp. 358-365, 2018, doi: 10.1111/jcal.12247.

[9] R. Asif, A. Merceron, S. Ali, and N. Haider, "Analyzing undergraduate students' performance using educational data mining," *Computers and Education*, vol. 113, pp.177-194, Oct. 2017, doi: 10.1016/j.compedu.2017.05.007.

[10] J. Jacob, K. Jha, P. Kotak, and S. Puthran, "Educational Data Mining techniques and their applications," *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, 2015, pp. 1344-1348, doi: 10.1109/ICGCIoT.2015.7380675.

[11] M. W. Berry, A. Mohamed, and B. W. Yap, "Supervised and Unsupervised Learning for Data Science," *Springer Nature*, New York, NY, 2019.

[12] P. Kaur, M. Singh, and G. S. Josan, "Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector," *Procedia Computer Science*, vol. 57, pp. 500-508, 2015, doi: 10.1016/j.procs.2015.07.372.

[13] A. A. Saa, "Educational Data Mining & Students' Performance Prediction," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 5, pp. 212-220, 2016.
[14] H. Al-Shehri, *et al*., "Student performance prediction using Support Vector Machine and K-Nearest Neighbor," *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, 2017, pp. 1-4, doi: 10.1109/CCECE.2017.7946847.
[15] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp.168-192, January 2021, doi: 10.1016/j.aci.2018.08.003.
[16] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano, "An empirical comparison of repetitive undersampling techniques," *2009 IEEE International Conference on Information Reuse & Integration*, 2009, pp. 29-34, doi: 10.1109/IRI.2009.5211614.
[17] G. H. Fu, L. Z. Yi, and J. Pan, "Tuning model parameters in class-imbalanced learning with precision-recall curve," *Biometrical Journal*, vol. 61, no. 3, pp. 652-664, Dec. 2018, doi: 10.1002/bimj.201800148.
[18] S. Yadav and S. Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification," *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, 2016, pp. 78-83, doi: 10.1109/IACC.2016.25.
[19] Tzu-TsungWong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross-validation," *Pattern Recognition*, vol. 48, no. 9, pp. 2839-2846, Sept. 2015, doi: 10.1016/j.patcog.2015.03.009.
[20] S. S. Athani, S. A. Kodli, M. N. Banavasi, and P. G. S. Hiremath, "Student academic performance and social behavior predictor using data mining techniques," *2017 International Conference on Computing, Communication and Automation (ICCCA)*, 2017, pp. 170-174, doi: 10.1109/CCAA.2017.8229794.
[21] N. B. Gaikwad, V. Tiwari, A. Keskar, and N. C. Shivaprakash, "Efficient FPGA Implementation of Multilayer Perceptron for Real-Time Human Activity Classification," *IEEE Access*, vol. 7, pp. 26696-26706, 2019, doi: 10.1109/ACCESS.2019.2900084.
[22] M. Gaudioso, W. Khalaf, and C. Pace, "On the Use of the SVM Approach in Analyzing an Electronic Nose," *7th International Conference on Hybrid Intelligent Systems (HIS 2007)*, 2007, pp. 42-46, doi: 10.1109/HIS.2007.16.
[23] S. E. Roshan and S. Asadi, "Improvement of Bagging performance for classification of imbalanced datasets using evolutionary multi-objective optimization," *Engineering Applications of Artificial Intelligence*, vol. 87, p. 103319, January 2020, doi: 10.1016/j.engappai.2019.103319.
[24] S. T. Jishan, R. I. Rashu, N. Haque, and R. M. Rahman, "Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique," *Decision Analytics*, vol. 2, no. 1, pp. 1-25, March 2015, doi: 10.1186/s40165-014-0010-2.
[25] P. Kaur, A. Gosain, "Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise," *Springer*, Singapore, pp. 23-30, 2018, doi: 10.1007/978-981-10-6602-3_3.
[26] Y. Zhang and P. Trubey, "Machine Learning and Sampling Scheme: An Empirical Study of Money Laundering Detection," *Computational Economics*, vol. 54, no. 3, pp. 1043-1063, October 2019, doi: 10.1007/s10614-018-9864-z.

BIOGRAPHIES OF AUTHORS

Nibras Z. Salih received the B.Sc. degree in Control and System engineering in 2008 from the University of Technology, Baghdad-Iraq and, He is an M.Sc. student at the University of Mustansiriyah, Baghdad-Iraq, since 2019, Thesis title was "Investigating Multiple instances learning classifiers for improved data classification".

Walaa Khalaf received the B.Sc. degree in electrical engineering in 1996 from Mustansiriyah University, Baghdad-Iraq and, in 2001, the M.Sc. degree in electronics and telecommunication engineering from the same University, and the Ph.D. from the University of Calabria, Italy. He is currently a professor of Operations Research at the Mustansiriyah University.