Journal of Engineering and Sustainable Development
http://jeasd.uomustansiriyah.edu.iq/index.php
https://doi.org/10.31272/jeasd.25.6.6

Vol. 25, No. 06, November 2021
ISSN 2520-0917

IMPROVING STUDENTS PERFORMANCE PREDICTION USING MACHINE LEARNING AND SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE

***Nibras Z. Salih[1]** **Walaa Khalaf[1]**

1) Computer Engineering Department, College of Engineering, Mustansiriyah University, Baghdad, Iraq

Received 8/3/2021 ***Accepted in revised form 16/5/2021*** ***Published 1/11/2021***

Abstract: Classification under supervision is the most common job that performed by machine learning. However, most Educators were worried about the rising evidence of student academic failures in university education. So, this study presents a supervised classification strategy of machine learning algorithm using an actual dataset contains 44 students, fourteen attributes for three previous academic years. We have proposed features that show the relationship among three main subjects which are, calculus, mathematical analysis, and control system in the education course. The objective of this study is to identify the student's failure in the control system subject and to enhance his performance by Multilayer Perceptron (MLP) algorithm. The dataset is unbalanced, which causes overfitting of the results. Synthetic Minority Oversampling Technique has applied to a dataset for obtaining balance dataset using Weka tool. Several standard metrics used to evaluate the classifier results. Therefore, the suitable results occurred after applying SMOTE with an accuracy of 76.9%.

Keywords: *Leave one out Cross-validation (LOOCV); Receiver Operating Characteristics (ROC); Precision-Recall curve (PRC); Synthetic Minority Oversampling Technique (SMOTE).*

1. Introduction

Machine learning has been one of the highest developing fields of computer science that involves the classification algorithms of important data patterns. Also, Machine learning techniques concentrate on the ability to learning and adapt to various programs. Supervised machine learning algorithms are categorized into different groups that are dependent on the expected results. Therefore, supervised algorithms generate the task that mapping inputs with the desired outcomes. Both data mining and machine learning are utilized to get various insights using appropriate algorithms [1]. Data mining (DM) techniques are utilized to improve the performance of education. DM is applied for identifying problem types, forecasting student achievement, and detecting attributes [2, 3]. Multiple data mining techniques are used to examine data collection and to find relevant data known as knowledge.

DM was already presented in the business field. Whereas, it has been proved to use for education that is designated as Education Data Mining (EDM). EDM is responsible for implementing data mining strategies for analyzing information derivation in an academic environment [4]. Classification has become one of the best data mining strategies used to categorize data. By assigning one label for each element in the dataset that is related to one class [5].

**Corresponding Author*: neranzezo@gmail.com

The ability to estimate student's failure during education courses is an important challenge in this context. Therefore, teaching strategies can be applied at an efficient level to overcome this failure [6-9]. In this paper, datasets were collected from the University of Mustansiriyah in Iraq. We have identified the main subjects including, calculus I, calculus II, mathematical analysis I, mathematical analysis II, and control system for five courses. We have proposed features that explain the relationship between these subjects and illustrate that a student's achievement in the calculus and mathematical analysis subjects affects on student's performance for studying the control system. This study aims to identify the student failure in the control system subject with their reasons. The explanations for these results will enhance students' performance in the future. Also, to prepare the student for starting a successful semester and to comprehend the fundamental concepts of these subjects.

The paper is divided into the following sections, Section II introduces the literature review, Section III present the data description and experiment design and Section IV illustrates the results and discussion and, the conclusions are presented in Section V.

SECTION II

2. Literature Review

Costa et al. presented results using an efficient EDM strategy for the early diagnosis of students who may to unsuccessful in the introduction program courses. By studying the effect of the data through pre-processing and the fine-tuning algorithm. Distance education and campus datasets were applied in public Universities from Brazil. The distance education dataset includes 262 students through 10 weeks by an online system that involves various features as semester, class, quiz, city, etc. While campus dataset includes 161 students through 16 weeks that involves different features like age, number of exercises, gender, etc. Four classifiers were utilized, SVM was the best classifier for two datasets [6].

In [7], the importance and influence of the student biography, social behavior, and student's activities were investigated. The actual dataset included 395 students and 33 attributes of mathematics subject and implemented using four classifiers which are Naïve Bayes Classifier (NBC), Multilayer Perceptron (MLP), J48, and Random Forest. Therefore, J48 was the suitable classifier that outperforms three metrics which are precision, recall, and F-measure.

Yu et al. introduced early estimation of students that were not anticipated to successfully an education course. Analysis of emotion was carried out using student's comments to detect effective information for increasing the prediction accuracy. The actual dataset contains 181 students of the computer science course in the summer program through nineteen weeks. Two algorithms were utilized which are SVM and neural network in the data mining technique. Yu et al concluded the neural network was the appropriate classifier [8].

Also, in [9] several strategies such as regression and the decision tree to forecast student's achievement and education failure were utilized. Thus, the dataset was gathered from the computer engineering that was used for predicting average scores and last semester for the learners. Thereafter, a clustering algorithm was applied for dividing the students into two categories by the k-mean classifier based on student participation in the subject of programming language.

Weka is a strong classification tool used to test and assess the accuracy of several algorithms in the field of machine learning.

Including the research, Kaur et al. detected and showed slow learners using a predictive data mining approach. Thus, the actual dataset involves 152 students with education and non-education information like the type of board, type of school, gender, etc. Five algorithms were applied including NBC, MLP, J48, Sequential Minimum Optimization (SMO), and Reptree to the student dataset. MLP was the suitable classifier of accuracy 75% and f-measure 82% [10]. Ahmed et al. constructed on teacher performance and examined causes for improving the efficiency of the education system. Thus, four classifiers were applied, including J48, NBC, MLP, and SMO. Also, the information was obtained at the University of California, which included a 5,820 importance value provided by the students that involved 28 distinct questions with five attributes. Ahmed et al. concluded the SMO and MLP were the best classifiers [3].

In [11], the ranges of personal, social, academic, and non-academic information were gathered. Analyzes showed that student achievement does not depend entirely on their academic potential, but various other variables also have an impact. NBC and decision tree algorithms were implemented to predict student performance. The information was acquired by way of a google form questionnaire that was sent to several learners during their regular studies.

Hussain et al. combined datasets with twenty-four attributes and three hundred students of social and academic information. J48, PART, Random Forest, and Bayes Net algorithms were implemented to the dataset. Therefore, the regular evaluation of the effective implementation variable had the most impact on the final semester results of the students [12].

So, several above works utilized a wide dataset that regarded a general overview of students. The datasets comprised information about the social, demographic, economics that was not immediately connected to the performance of the learners.

SECTION III

3. Dataset Description

The dataset from Mustansiriyah University was gathered during the education year in 2019. Dataset depended on the main subjects including information from three years: calculus I and calculus II in the first year. Also, the mathematical analysis I and mathematical analysis II were included in the second year. Finally, the control system was included in the first course of the third year. There are two courses in one education year each one with three attributes which are the total lecture absence, the assessment grade, and the final grade. The total lecture absence attribute contains three values which are 0, 1, and 2.

i) 0 is described as the final warning. ii) 1 is described as the initial warning. iii) 2 is described as not registered absences.

While, quizzes, mid examinations, and assignments of the students are components of the assessment grade attribute. Moreover, the highest value of the assessment grade attribute is forty and the lowest attribute value is zero. Also, the final grade attribute contains three values which are 0, 1, and 2.

i) 0 is described when the student doesn't succeed in both the first and the second trial. ii) 1 is described when the student has failed the first trial but passed at the second trial. iii) 2 is

described when the student succeeded the exam from the first trial. Table 1 shows the description of the attributes. Hence, the first attribute is the student identifier (id).

Table 1. Attributes description and possible value

Attributes	Range of Attributes	Description
Total lecture Absence	[0, 1, 2]	0 – Final warning 1 – Initial warning 2 – No Absence has Registered
Assessment grade	quizzes [0 – 10] mid examination [0 – 20] assignments [0 -10]	0 – Indicate to the lowest assessment grade 40 – Indicate to the highest assessment grade
Final grade	[0, 1, 2]	0 – Failed both trials 1– Pass in the second trial 2 – Pass in the first trial
Class	[0, 1]	0 – Pass 1 – Fail

3.1. Proposed System

The dataset contains forty-four instances, thirteen attributes, and a one-class attribute to each instance which is 0 (pass) or 1 (fail). Thus, we have generated a dataset according to the relationship between calculus and mathematical analysis subjects on one hand, and control systems on the other. Also, we have proposed features that demonstrate these subjects are shared for the basic principles such as digital and analog systems, Laplace transform, Z-transform, linear system, nonlinear system, etc. Therefore, the student's performance in the calculus and mathematical analysis subjects could affect the result of the student at the control system subject. MLP algorithm is utilized to predict the student's result using three techniques which are training set, Leave-One-Out Cross-Validation (LOOCV), and five-fold Cross-Validation (5-CV). The dataset goes through three phases, in the first phase, the classification model is trained on this dataset to find the best result which can be used at the next phase to test undefined instances. Finally, the accuracy of the MLP classifier is measured from the confusion matrix and its metrics. Figure 1 shows the scheme of the proposed system.

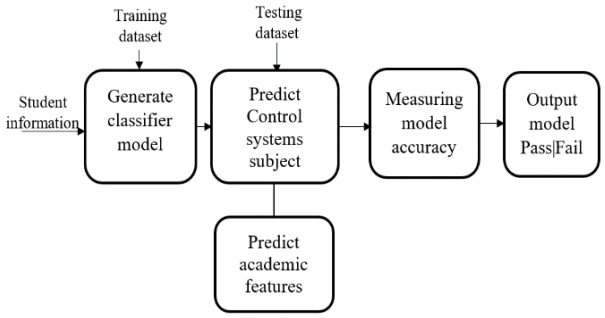

Figure 1. The proposed system structure.

Confusion Matrix includes four parameters, which are TP, TN, FP, and FN. Since TP is True Positives, TN is True Negatives, FP is False Positives, and FN is False Negatives. Therefore, for evaluating the efficiency of the selected classifier, several metrics are being utilized: sensitivity, specificity, precision, F-measure, accuracy, ROC, and PRC [12, 13].

The sensitivity is defined as the ratio of true positives by the number of positive instances, it is also known as recall which is described in equation 1.

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

The specificity is defined as the ratio of true negatives by the total number of the negative instances which is described in equation 2.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \quad (2)$$

Precision is defined as the ratio of true positives by the number of expected positive instances which is described in equation 3.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

The F-measure is the mixture of precision and recall metrics which is described in equation 4.

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

The accuracy is utilized to evaluate the classifier performance. It is defined as the ratio of true classification instances by the total number of instances which is described in equation 5.

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

The Precision-Recall curve (PRC) is utilized to assess the classification model of the noisy and unbalanced dataset. PRC displays the recall of the x-axis and precision of the y-axis for various thresholds.

Also, Receiver Operating Characteristics (ROC) curve is implemented to evaluate the classification model, it consists of two axes which are the false-positive rate on the x-axis and the true positive rate on the y-axis. The ROC curve can be used to perform a classification model using a balanced dataset for every class. Therefore, to compare the classification model as one measure can be used both ROC and PRC curves [14, 15].

Cross-Validation (CV) is designed to evaluate the classification algorithms by separating the dataset into two groups including the training set and testing set groups. So, CV is implemented to compare the classification results for various algorithms.

Moreover, k-folds CV is utilized to assess the classification performance for any algorithm of the machine learning approach. Thus, the dataset is divided into k-folds where the dataset is divided into k-folds equally. Subsequently, the training set and testing set are generated k times (iterations). Therefore, one fold of the dataset is utilized as a test set while the remaining k-1 folds are utilized as a train set.

While Leave One Out Cross-Validation (LOOCV) is a specific case of k-fold CV. In LOOCV, the number of folds is the same as the instances number. Also, LOOCV is used to assess the classifier performance of the machine learning approach when there is a limited number of instances [16, 17].

3.2. Experiment Design

Weka has been designed to organize the dataset, where it is an open-source program utilized for multiple operations in data mining [3, 10, 12]. Weka contains various tools to prepare, classify, cluster, process, and visualize the datasets. One classifier is selected in the Weka tool which is Multilayer Perceptron (MLP). MLP is the supervised algorithm that is utilized the backpropagation algorithm to train neural networks and also for classifying instances. MLP contains several layers, each layer includes a set of neurons that is connected to the next layer [18].

This study deals with an unbalanced dataset since the number of cases of one class is less than the other. Where the smaller class was known as a minority class; whereas the larger class was known as a majority class. So, to solve these issues, we have applied Synthetic Minority Oversampling Technique (SMOTE).

SMOTE is an over-sample solution for the issue of unbalanced datasets that converts an

unbalanced dataset to a balanced dataset. SMOTE is produced synthetic cases in the smaller class. Therefore, both the smaller and larger classes have been distributed using generating synthetic cases in a smaller class. Thus, this procedure is used to improve prediction performance in a smaller class.

The sample is placed in the smaller class over the lines that contain some of the nearest neighbors. In many cases, SMOTE utilized five nearest neighbors [19, 20]. Thus, over-sample increased the number of instances to keep both instances and Non-instances (synthetic cases) using replacement samples [21]. SMOTE function in the Weka tool increased the instances number by 50 % for the smaller class.

SECTION IV

4. Result and Discussion

For the average of the seven metrics, the relevant results are presented in Table 2 using the training set, LOOCV, and five times 5-CV. One classifier is applied to the dataset which is MLP to predict the performance of the students. We have implemented this experiment with the training set methodology to evaluate the classification of the prediction class. Then, CV techniques (LOOCV and five times 5-CV) have implemented to assess the classification model. We have compared the classification model results for the training set with the classification results of CV techniques.

We have found that the classification performance of the training set is higher than CV techniques. The difference between the results of the training set and CV techniques is because of the effect of overfitting in the classification model. Therefore, we have applied the SMOTE supervised filter to overcome the overfitting issues and for improving the prediction of the classification model. The results show that the MLP outperforms in terms of sensitivity, specificity, precision, and F-measure for LOOCV; while it outperforms for both ROC and PRC metrics for 5-CV.

The maximum values of the classification results are underlined in Table 2.

When there are low specificity values it is usually an effect of high sensitivity values. Thus, specificity is a significant metric since it indicates a student's failure in an academic course.

Moreover, the ROC and PRC values became higher when SMOTE filter was applied to the dataset, as shown in Table 2, which proves that the dataset turned into a balanced dataset. Figure 2 shows the classification results of the student dataset.

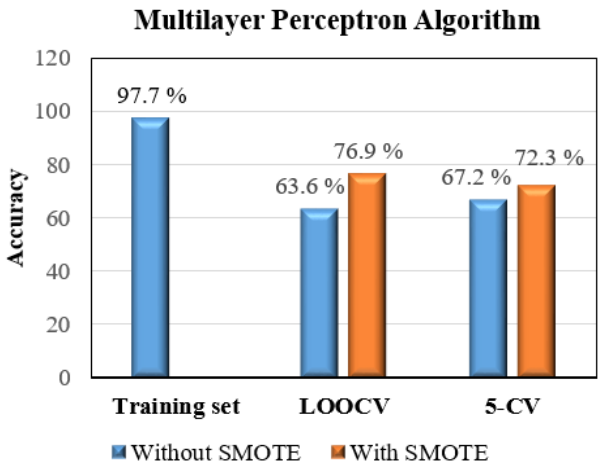

Figure 2. Classification results of the student dataset.

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

	Without SMOTE			With SMOTE	
Metrics	**Training set**	**LOOCV**	**5-CV (Mean ±Std)**	**LOOCV**	**5-CV (Mean ± Std)**
Sensitivity	0.98	0.64	0.67 ±0.01	0.77	0.72 ± 0.02
Specificity	0.96	0.70	0.77 ± 0.02	0.78	0.64 ±0.02
Precision	0.98	0.64	0.67 ± 0.02	0.77	0.73 ± 0.02
F-Measure	0.98	0.64	0.67 ± 0.02	0.77	0.72 ± 0.02
ROC	0.97	0.64	0.68 ± 0.01	0.75	0.77 ± 0.01
PRC Area	0.96	0.68	0.70 ± 0.02	0.72	0.78 ± 0.02
Accuracy	97.7 %	63.6 %	67.2 ± 1.2	76.9 %	72.3 ± 2.2

SECTION V

5. Conclusion

In this study, an exploration goals to offer a prediction to teachers that might aid them to get better learning programs at their universities. Without any socio-economic information, predicting students' achievement based on assessment grades, final grades, and lecture attendance. We have proposed features that describe the relationship for three fundamental subjects: calculus in the first year, mathematical analysis in the second year, and the control system subject for the third year. The goal of this study is to improve student performance in calculus and mathematical analysis subjects, lead to prevent students from failing in the control system subject.

Three techniques are implemented to the dataset using the Weka tool which are the training set, LOOCV, and 5-CV, with the help of MLP algorithms. We have noticed that the classification model suffered from overfitting as a result of the unbalanced student dataset. So, to address this issue and to improve the prediction of students' performance, a supervised SMOTE technique is used. Thus, we have concluded that suitable results occur after implementing SMOTE filter with an accuracy of 76.9% in LOOCV.

The conclusions indicate that the ability to anticipate a student's performance in one subject to achieve the best undergraduate grades.

In the future, we will expand the dataset to improve the prediction 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.

Conflict of Interest

The authors confirm that the publication of this article causes no conflict of interest.

6. References

1. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and Chouvarda, I. (2017). *"Machine learning and data mining methods in diabetes research*". *Computational and structural biotechnology journal*, Vol. *15*, pp. 104-116.
2. Adekitan, A. I. and Salau, O. (2019). "*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

3. Ahmed, A. M., Rizaner, A. and Ulusoy, A. H. (2016). "*Using data mining to Predict Instructor Performance*". Procedia Computer Science, Vol. 102, pp. 137–142.
4. Ketui, N., Wisomka, W., and Homjun, K. (2019). "*Using Classification Data Mining Techniques for Students Performance Prediction*", in IEEE 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), pp. 359-363.
5. Berry, M. W., Mohamed, A. and Yap, B. W (2019). "*Supervised and Unsupervised Learning for Data Science"*. New York, NY: Springer.
6. Costa, E. B., Fonseca, B., Santana, M. A., Araújo, F. F. de and Rego, J. (2017). "*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.
7. Kiu, C. (2018). "*Data Mining Analysis on Student's Academic Performance through Exploration of Student's Background and Social Activities"*, in 2018 Fourth International Conference on Advances in Computing, Communication, and Automation, Subang Jaya Malaysia, pp. 1-5.
8. Yu, L. C., Lee, C. W., Pan, H. I., Chou, C. Y., Chao, P. Y., Chen, Z. H., Lai, K. R. (2018). "*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.
9. Jacob, J., Jha, K., Kotak, P. and Puthran, S. (2015). "*Educational Data Mining techniques and their applications"*, in IEEE International Conference on Green Computing and Internet of Things, pp. 1344–1348.
10. Kaur, P., Singh, M. and Josan, G. S. (2015). *"Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector"*. Procedia Computer Science, Vol. 57, pp. 500–508.
11. Abu, A. (2016). "*Educational Data Mining & Students' Performance Prediction*". International Journal of Advanced Computer Science and Applications". Vol. 7, No. 5, pp. 212–220.
12. Hussain, S., Dahan, N. A., Ba-Alwib, F. M., and Ribata, N. (2018). "*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.
13. Tharwat, A. (2018). "*Classification assessment methods*". Applied Computing and Informatics.
14. Van Hulse, J., Khoshgoftaar, T. M., and Napolitano, A. (2009). "*An empirical comparison of repetitive undersampling techniques*". In 2009 IEEE international conference on information reuse and integration, pp. 29-34.
15. Fu, G. H., Yi, L. Z., and Pan, J. (2019). *"Tuning model parameters in class-imbalanced learning with precision-recall curve*". Biometrical Journal, Vol. 61, No. 3, pp. 652-664.
16. Yadav S, Shukla S. (2016). "*Analysis of k-Fold Cross-Validation over Hold-Out*

Validation on Colossal Datasets for Quality Classification". In: Advanced Computing (IACC) 2016 IEEE 6th International Conference on. IEEE; pp. 78–83.

17. Wong, T-T. (2015). "*Performance evaluation of classification algorithms by k-fold and leave-one-out cross-validation*". Pattern Recognition, Vol. 48, No. 9, pp.2839–2846.
18. Gaikwad, N. B., Tiwari, V., Keskar, A., and Shivaprakash, N. C. (2019). "*Efficient FPGA implementation of multilayer perceptron for real-time human activity classification*". IEEE Access, Vol. 7, pp. 26696-26706.
19. Jishan, S. T., Rashu, R. I., Haque, N., and Rahman, R. M. (2015). "*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.
20. Kaur, P., Gosain, A. (2018). "*Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise*". Singapore: Springer, pp. 23–30.
21. Zhang, y. and Trubey, P. (2019). "*Machine Learning and Sampling Scheme: An Empirical Study of Money Laundering Detection"*. Computational Economics, Vol. 54, No. 3, pp. 1043–1063.