Jurnal Ilmiah KURSOR
Menuju Solusi Teknologi Informasi
Vol. 12, No. 3, Juli 2024

ISSN 0216 – 0544
e-ISSN 2301– 6914

STUDENT ACADEMIC PERFORMANCE PREDICTION FRAMEWORK WITH FEATURE SELECTION AND IMBALANCED DATA HANDLING

[a]Vivi Nur Wijayaningrum, [b]Annisa Puspa Kirana, [c]Ika Kusumaning Putri

[a,b,c] Department of Information Technology, Politeknik Negeri Malang
Jalan Soekarno Hatta 9, Malang, East Java
E-mail: vivinurw@polinema.ac.id

Abstract

Various factors cause the low scores of students in practicum courses. If these factor cannot be identified, more and more students will drop out of the study due to lo scores, especially Vocational College students who do not have the opportunity t improve their scores in the short semester. Students with the potential to drop out mus be identified as soon as possible because the number of dropouts can have an impac on a university's accreditation value. In this study, the prediction of student academi performance was carried out using a framework consisting of imbalanced data handlin using SMOTE and feature selection using Random Forest, as well as the application o Multi-Layer Perceptron (MLP) for the formation of a classification model. The ML architecture consists of some neurons in the input layer, two hidden layers with fiv neurons each, and two neurons in the output layer. SMOTE succeeded in selecting te significant parameters from 22 initial parameters, which produced the most accurat predictions. According to the test results, the proposed framework offers the bes accuracy of 0.8889 and an F1-Score of 0.9032. These results prove that the proposed framework can be used as an alternative for the Department to take action to prevent students from dropping out.

Key words: Classification, Drop Out, Random Forest, SMOTE.

INTRODUCTION

A concerning issue in many higher education institutions around the world is student dropout rates [1]. Academic difficulties, physical and mental health issues, demanding part-time jobs, and other reasons are among the many causes of student dropouts [2]. Learning activities will unavoidably need to be done online starting in mid-2020 due to numerous policies issued by various governments throughout the world. As a result, many students struggle to adapt during the learning process to understand the material [3]. Students who reside in underdeveloped areas experience issues with erratic internet connections during online learning and occasionally simply no connection at all [4]. On the other hand, due to a lack of resources at home and a communication gap with instructors brought on by online learning, students are also unable to engage in their learning activities [5-7]. As a result, the dropout rate is getting higher.

The community and students are also impacted by the number of student dropout cases, apart from the study program and the college [8]. The impact received by a study program when a student drops out is the effect of the accreditation value of the study program [9]. A university's study program with a low dropout rate and a high student graduation rate is rated as having a higher quality than one with

a high dropout rate [10][11]. The value of accreditation of study programs and universities will have an impact on public perception so that it affects new student registration [12], academic reputation, teaching quality and employability of lecturers [13]. Therefore, study programs and universities are trying to improve their quality to increase their accreditation value.

Because of academic policies that stipulate that students receive dropout status if they receive a final grade of "D" for more than one practicum course in a semester, vocational universities are concerned about dropout rates. Since vocational colleges are focused on improving student performance and skills prior to entering the world of work [14], as opposed to other universities that have a higher proportion of theoretical courses than practicum courses, practicum courses play a significant role in lecture hours. In fact, students who are unable to meet the requirements outlined in the Academic Manual regarding grades will ultimately have to drop out of vocational universities, as they do not offer the option for low-achieving students to retake or improve on certain courses in the following semester.

Quite a few first-year students at vocational universities fail and end up dropping out for various reasons behind it. Most first-year informatics engineering students fail when taking the Programming Fundamental Practicum course. It is a primary course in the first semester that must be mastered by students in the field of informatics engineering to produce logical, creative, and critical thinking [15]. They are expected to have basic skills in computer programming languages [16]. However, not many students have difficulty learning it due to various factors. A weak programming foundation is typically possessed by new students who do not come from computer science-focused schools and who have little experience with programming languages [17]. Learning computer programming languages is more complicated than other fields of science because it requires skills to design algorithms, write program code, and understand program code syntax [18]. Therefore, to reduce the failure rate of students in the Programming Fundamental Practicum course, it is important to predict students' academic performance. Thus, students dropping out can also be prevented from an early age.

The dropout problems are still experienced by various universities to this day. Previous researchers have tried various settlement efforts, for example, looking for factors that influence the occurrence of students dropping out or predicting the possibility of dropping out early [19]. Various parameters are used to predict the possibility of students dropping out, including student profiles such as age, gender, student demographics, admission selection path, student daily attendance reports during lectures, student grades, student participation in extra activities, social network interactions, background parents' education, and occupation, number of siblings, and several other factors [20-24].

Several algorithms have been employed in their application to predict the likelihood of student dropouts using different student information-related input parameters. Harwati [25] separately applied the Support Vector Machine (SVM) and the Naïve Bayes algorithm to estimate dropout rates. The research conclusion shows that Naïve Bayes performs better than SVM, with an accuracy of 80.67%, compared to SVM, which is 60%. In addition to these two algorithms, the Decision Tree is also widely used by researchers to predict student dropouts. According to Mutrofin [26], the Decision Tree algorithm works best when the classification is carried out on classes with a balanced distribution. In their study, the C4.5 algorithm was applied to data that had missing values and without amputation, which resulted in an accuracy of 96.87%. In another research, Random Forest was used by Utari [27] to solve a similar case. The imbalanced data is overcome by using synthetic minority oversampling (SMOTE) when preprocessing the data, then Random Forest is applied for classification and produces an accuracy of 93.43%. Meanwhile, Mengash [23] also compared some algorithms to predict student academic performance. The test results in this research state that Artificial Neural Network (ANN) is superior to Decision Tree, SVM, and Nave Bayes, with an accuracy of 79.22% for ANN, 75.91% for Decision Tree, 75.28% for SVM, and 73.61% for Nave Bayes. Similar results were also obtained by Sa'ad [28] in their research to predict student dropouts,

which found that Extreme Learning Machine, one of the ANN algorithms, outperformed SVM with an accuracy of 72%, while the accuracy of SVM was 63%.

In addition to the suitability of algorithm selection, the quantity of features and the presence of outlier data can affect how well the problem is solved [29]. Several selected relevant features can enhance the algorithm's performance by decreasing the model complexity so that accuracy increases [27]. Feature selection reduces computation time, which not only improves data quality but also expedites the data collection process [30] because the data dimensions used when modeling is smaller [31].

Many parameters are also involved in the probability of a student dropping out based on student grades. Wild and Heuling [32] stated that low college enrollment scores, low Grade Point Average (GPA), and the type of courses taken each semester would influence student study results. Morampudi [33] involves important parameters such as Cumulative Learning Assessment, mid-term exam grades, laboratory assignment grades, performance, attendance, and theory grades involving semester grades, mid-term grades, and assignment grades to analyze student performance.

This research predicts students' academic performance in the Programming Fundamental Practicum course by entering some parameters and student data from Learning Management System, Student Academic Information System, and student questionnaires about learning activities and backgrounds. Referring to earlier research, MLP was effectively used to predict the academic performance of students using 18 parameters, yielding an F1-Score of 0.9032 and an accuracy of 0.8235 when combining 60 training and 40 test data. Meanwhile, the highest accuracy and F1-Score obtained with k-fold cross-validation are 0.8091 and 0.8099, respectively, for k values of 8 and 9. To enhance the performance evaluation results of the MLP algorithm in predicting student academic performance, this research proposes feature selection based on the parameters used as determinants of prediction results. By using feature selection, parameters that affect student academic performance can be identified so that the resulting accuracy is higher. Thus, the prediction results can be applied to prevent early dropouts due to low grades in practicum courses. The novelty of this research lies in its architecture to identify and prioritize the most influential parameters affecting student performance, as well as imbalanced data handling on the dataset used, with the hope that the accuracy value can be higher than in previous research..

MATERIAL AND METHODS

The workflow of the proposed system framework in this study is shown in Figure 1. The data set was collected from student questionnaires, the Learning Management System (LMS), and the Student Academic Information System (SIAKAD). Splitting the data into training and test sets is then carried out in the second stage. There are two scenarios used in the research. A specific percentage of training and test sets are used in the first scenario. While the second scenario uses k-fold cross-validation, which consists of training, validation, and test folds. To guarantee that the data is evenly distributed, k-fold cross-validation is applied. In the third stage, Synthetic Minority Oversampling (SMOTE) is used to handle imbalanced data. In this case, it is only applied to training sets in the data set to generate minority class data into several majority data. Meanwhile, in k-fold cross validation, imbalanced data handling is not carried out because the data is randomly distributed through k-fold. The fourth stage involves applying feature selection to select a set of features that influence the prediction results, which are indicated with the highest accuracy, after the data is prepared for use. The data set containing the optimal feature subset can then be utilized to create a MLP prediction (classification) model in the following step. "Potentially Drop Out" and "Not Potentially Drop Out" are the two classes that result from this prediction. Lastly, the accuracy and F1-Score are then used as evaluation metrics to gauge the quality of the results.

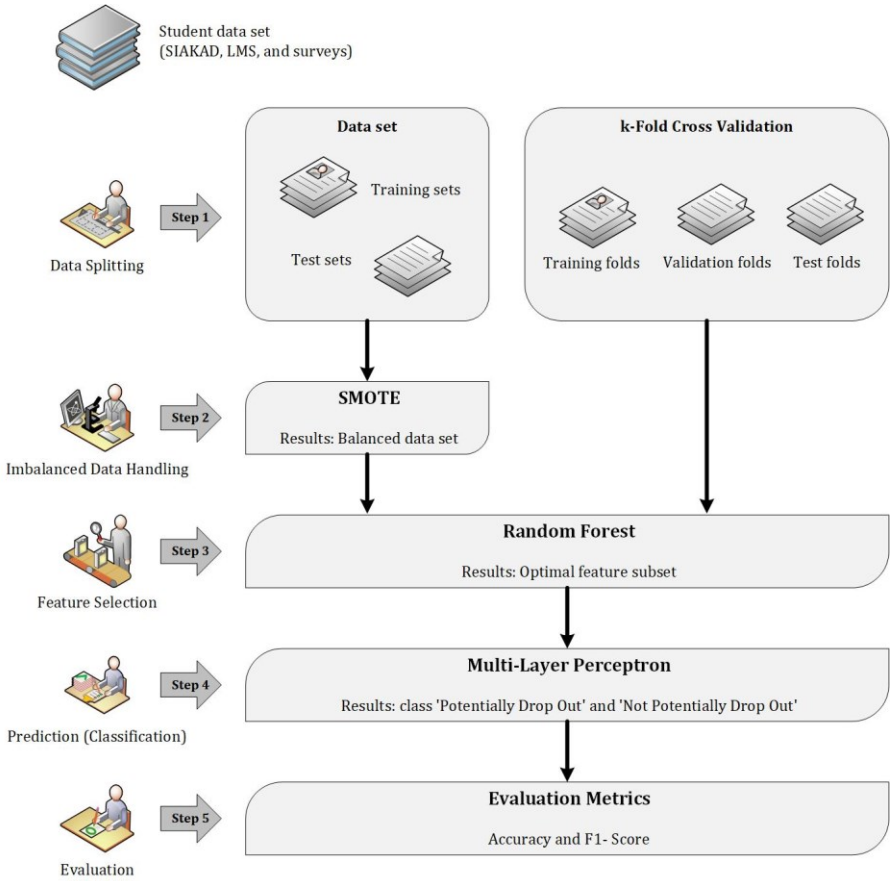

Fig 1. Workflow of the proposed system framework

Data Set

In this research, the primary data used is related to background information and student activities during the Programming Fundamental Practicum lecture. The three data sources used are SIAKAD, LMS, and student surveys. The data from SIAKAD is general information and the educational background of students. Through LMS, data related to student activities during the learning process can be obtained, such as participation in assignments and exams, as well as activeness in the learning process. Meanwhile, student surveys are used to obtain information related to resources to support the learning process, student habits and activities outside of lecture hours, and other conditions that are considered to influence the learning process.

84 students attended the Programming Fundamental Practicum lecture and were used as objects in this study, consisting of 16 students in "Potentially Drop Out" and 68 students in "Not Potentially Drop Out". The students come from five different classes, namely D3 Informatics Management (Class 1B and 1F), and D4 Informatics Engineering (Class 1B, 1C, and 1H) with different lecturers.

The criteria used were chosen based on previous research which shows that factors other than grades also influence student learning outcomes. Tables 1 to 3 show the sample data used, sourced from LMS, SIAKAD, and student surveys.

Table 1. Information Sourced from LMS

Student	Completeness of assignment submission	Late Submission of Assignments and Exams
1	100%	11.76%
2	100%	11.76%
3	100%	17.65%
…	…	…
83	66.67%	50%
84	100%	12.5%

Table 2. Information Sourced from SIAKAD

Student	Study program	Age	Gender	School origin	Pathway Admission	Source of tuition fees	Attendance
1	Informatics Management	19	Male	Senior High School	PMDK-PN	No scholarships	100%
2	Informatics Management	19	Male	Senior High School	UMPN	No scholarships	100%
3	Informatics Management	20	Female	Senior High School	Mandiri	No scholarships	100%
4	Informatics Management	20	Male	Vocational High School	UMPN	No scholarships	92.31%
…	…	…	…	…	…	…	…
83	Informatics Engineering	20	Male	Vocational High School	Mandiri	No scholarships	100%
84	Informatics Engineering	19	Male	Vocational High School	Mandiri	No scholarships	100%

Table 3. Information Sourced from Student Surveys

Student	College educated parents	Health condition	Ownership of learning devices	Activities outside of lectures	Study method	…	Has programming experience
1	Both parents	Has a history of illness	Owned by siblings/parents	No other activities	Study independently	…	No
2	Mother	Has no history of illness	One's own	No other activities	Study in groups with classmates	…	No
3	No one	Has no history of illness	One's own	Work	Study independently	…	No
4	No one	Has no history of illness	One's own	No other activities	Study independently		Yes
…	…	…	…	…	…	…	…
83	Both parents	Has no history of illness	One's own	No other activities	Study independently	…	Yes
84	Both parents	Has no history of illness	One's own	No other activities	Study in groups with other campus friends	…	No

Preprocessing is carried out on the dataset before moving to the next stage. This process consists of data cleaning, data imputation, and data scaling. In this study, missing values are handled using the average value of the feature. Meanwhile, data labeling is done using Label Encoding to change category features into numeric ones, such as 0, 1, 2, etc., so the data is ready to be used.

Data Splitting

Prior to further processing, data splitting is performed by separating the data into training and test sets. The percentage composition of training sets and test sets must be tested first to obtain the best predictive results. The proportion of initial training and test sets utilized in this research refers to earlier research, which used 70% training sets and 30% test sets. However, by going through multiple test scenarios, the composition of the training and test sets is also adjusted.

In addition to dividing the data set into two parts, data sharing is also done using k-fold cross-validation, which consists of training,

validation, and test. To prevent over-fitting during the evaluation of the measurement results, this procedure attempts to guarantee that the training data used to create the model is evenly distributed.

Imbalanced Data Handling

The student data used in this study tends to have students with a class of 'Not Potentially Drop Out' rather than 'Potentially Drop Out', so the use of the SMOTE algorithm is needed to overcome the imbalanced data set. At this stage, the sample dataset is rearranged using the resampling technique to decrease the influence of an imbalanced class distribution on the model development. SMOTE, in this research, operates by using interpolation techniques to create artificial data from the minority class, i.e., the 'Potentially Drop Out' class, based on k-nearest neighbors among minority classes. In this case, the SMOTE algorithm does not duplicate existing data but generates new data based on k-nearest neighbors [27].

The stages of implementing SMOTE to handle imbalanced data are as follows:

1. Randomly choose a minority sample x_0.
2. Find the minority samples' K-nearest neighbors for x_0.
3. Randomly choose one of the K-nearest neighbors of x_0 from the preceding step and assign the chosen sample as x_{knn}.
4. Using Equation (1), perform linear interpolation between x_0 and the selected neighbor x_{knn} to generate a new synthetic sample x_{syn}.

$$x_{syn} = x_i + (x_{knn} - x_i) \times \delta \quad (1)$$

   x_{syn} is the newly generated data, x_i indicates the replicated data, x_{knn} is the data that has the shortest distance from x_i, and δ is a random value in the range 0 and 1.
5. Continue steps 1-4 until M synthetic samples are produced.

Feature Selection

During the feature selection stage, Random Forest is employed to select multiple features that yield the best classification outcomes compared to utilizing all the features. The stages of the Random Forest algorithm for classification are as follows:

1. Select a random sample from the training sets.
2. For every sample that has been chosen, create a decision tree so that each decision tree can yield the classification results.
3. The voting process is carried out for each classification result using the mode (the value that appears most often).
4. Choose the classification results based on the most votes.

Prediction (Classification)

The prediction of student academic performance is done using MLP by classifying the data set into two classes, namely 'Not Potentially Drop Out' and 'Potentially Drop Out'. The MLP algorithm, which is an artificial neural network algorithm, consists of multiple hidden layers situated between the input and output layers. Without any prior knowledge about student data, this algorithm attempts to perform its function [34]. The input layer contains some neurons that are linked to neurons in the following layer. The number of neurons in this input layer is correlated with the number of parameters used to predict students' academic performance. Meanwhile, the 'Not Potentially Drop Out' and ' Potentially Drop Out' classes are the two neurons in the output layer that represent the quantity of output classes in the data set.

During its operation, MLP relies on some parameters to effectively learn and construct a model. Several parameters need to be well determined to produce an optimal solution [35]. In this study, the MLP parameter values used are shown in Table 4. These parameters are fine-tuned to ensure that the algorithm can produce the most accurate predictions.

Table 4. Parameter value of MLP

Parameter	Value
Number of layers	2
Number of neurons	5
Number of epochs	300
Learning rate	0.01
Threshold for error	0.0001
Momentum	0.9

Evaluation

After the stages of imbalanced data handling and feature selection, the performance measurement of the MLP algorithm in making predictions can be measured using evaluation metrics. In this study, two evaluation metrics used are

Accuracy and F1-Score. True Positive, False Positive, True Negative, and False Negative are the four parameters considered in the Accuracy and F1-Score computations, which involve a Confusion Matrix [36].

RESULT AND DISCUSSION

Evaluation metrics were used in several tests to measure the performance of the prediction framework using selection features and imbalanced data handling. Ten significant parameters were identified from 22 initial parameters that produced the best-accurate prediction results after the imbalanced data handling was completed with SMOTE and certain features were chosen. The parameters of the feature selection results are as follows:

- Age
- Pathway admissions
- Attendances
- Completeness of assignment submission
- Late submission of assignments and exams
- Health condition
- Ownership of learning devices
- Study method
- Study frequency
- Attendance at practical courses

The feature selection has some implications, such as reducing dimensions and noise. With many features, the complexity of the model increases, potentially leading to overfitting, increased computational costs, and reduced interpretability. By selecting only the most relevant features, dimensionality can be reduced, making the model more manageable and interpretable. Some features may also introduce noise or irrelevant information into the model, which can degrade its performance. Feature selection helps to eliminate such noise by excluding less relevant parameters, leading to a cleaner and more accurate model.

By using these ten parameters, MLP is then used to form a model using the best parameters shown in Table 4. The test was carried out using two schemes, namely testing the composition of the training and test sets as well as k-fold cross-validation. The test results from this research are compared to earlier research results (without feature selection and imbalanced data handling) [37] to assess how well the proposed framework performs with these features. In Figure 2, the accuracy values of the framework used in this study are compared to the results from previous research in each scenario where the composition of the training and test sets is being tested. While Figure 3 shows the calculation of the evaluation metrics using the F1-Score.

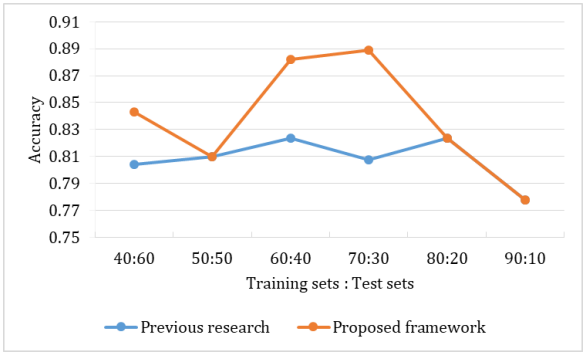

Fig 2. Graph of accuracy comparison results on data set composition

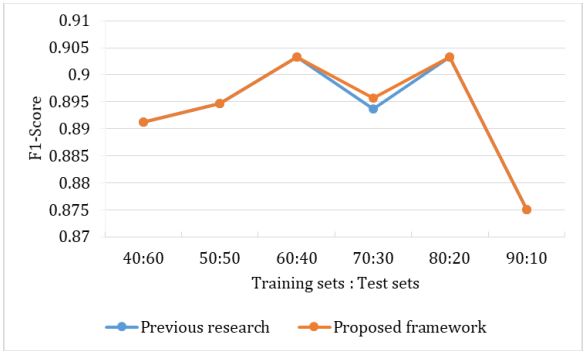

Fig 3. Graph of F1-Score comparison results on data set composition

Additionally, as indicated in Figures 4 and 5, the accuracy and F1-Score of the framework used in the study were compared with previous research to perform the k-fold cross validation.

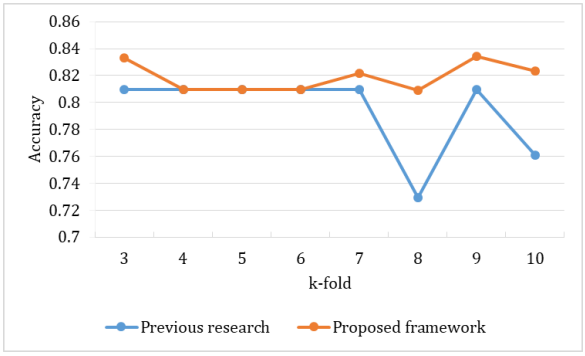

Fig 4. Graph of accuracy comparison results on k-fold cross validation

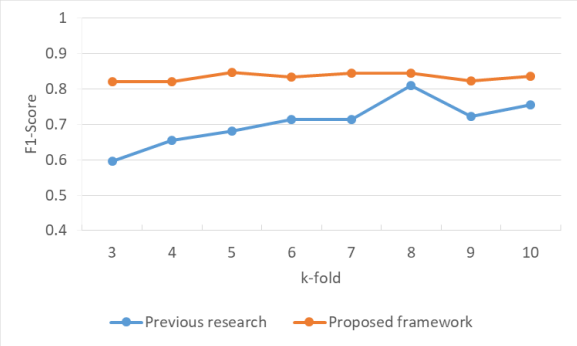

Fig 5. Graph of F1-Score comparison results on k-fold cross validation

The results of several tests have demonstrated that imbalanced data handling and feature selection enhance the predictive outcomes of student academic performance, both when using the composition of the training and test sets and during k-fold cross validation. The minority class has a higher chance of being the outcome of its initial condition due to imbalanced data handling using SMOTE in the training sets, which causes the data with the minority class to increase in proportion to the number of the majority class. Meanwhile, the high accuracy and F1-Score values compared to earlier research demonstrate the function of Random Forest in selecting features that affect prediction outcomes. As seen in Figure 5, the F1-Score of this framework can outperform the F1-Score of previous research in each k-fold test scenario.

Table 5. Performance Comparison Results

No	Techniques	Dataset	Accuracy
1	Convolutional Neural Networks [39]	MOOC	90.72%
2	Levenberg–Marquardt Backpropagation [40]	University's database	84.8%
3	Multi-Layer Perceptron [41]	University's database	77%
4	Long-short-term Memory [42]	MOOC and LMS	87%
5	Multi-Layer Perceptron [37]	SIAKAD, LMS, and surveys	82.35%
6	**Proposed framework**	**SIAKAD, LMS, and surveys**	**88.89%**

The performance comparison results between the proposed framework and several alternative techniques for predicting students' academic performance are displayed in Table 5. The proposed framework outperforms other algorithms used to solve similar problems with higher accuracy. Meanwhile, Convolutional Neural Networks (CNN) provide the highest accuracy among other algorithms. However, this algorithm requires very high resources for the training process and is very difficult to optimize [38]. In contrast, the proposed framework's feature selection can decrease processing time and resource consumption because there aren't many less significant data parameters involved in the training process, which allows for excellent accuracy to be obtained with minimal effort.

The results of the proposed system framework are also compared with previous research using the same data. The earlier research employed multiple variations of techniques to implement feature selection but did not apply techniques for balance data handling. Table 6 shows the comparison results of this study with previous research (without SMOTE).

Table 6. Results with and without SMOTE

No	Techniques	Accuracy	F1-Score
1	MLP with Chi Square [43]	81.11%	80.98%
2	MLP with Pearson Correlation Coefficient [43]	81.11%	82.09%
3	MLP with Random Forest [43]	81.11%	82.22%
4	**Proposed framework**	**88.89%**	**90.32%**

Table 6 shows the superiority of the proposed framework compared with previous research. The use of SMOTE to handle imbalanced data has a significant impact on accuracy and F1-Score. SMOTE can improve the model's recognition of patterns in minority classes, which will improve its overall generalization and capacity to generate precise predictions on new data. By handling imbalanced data using SMOTE, accuracy is higher at 88.89%, and F1-Score reaches 90.32%. Research by Kaope [44] and Maulana [45] also supports the findings of this study,

demonstrating that SMOTE can enhance algorithm performance in comparison to methods that do not address imbalanced data.

CONCLUSION

The imbalanced data handling implementation using SMOTE and feature selection using Random Forest has proven to be successful in increasing the accuracy and F1-Score in predicting student academic performance using MLP. Through a series of data set composition tests and k-fold cross-validation, this proposed framework provides the highest accuracy of 0.8889 and F1-Score of 0.9032, higher than the predicted results without any imbalanced data and feature selection.

In further research, additional data sets can be done first to overcome imbalanced data so that the distribution of classes in the data set is not too far apart. In addition, other imbalanced data handling techniques can also be used as an alternative to improve data set problems.

REFERENCES

[1] L. Bäulke, C. Grunschel, and M. Dresel, "Student dropout at university: a phase-orientated view on quitting studies and changing majors," *European Journal of Psychology of Education*, pp. 1–24, 2021, doi: 10.1007/s10212-021-00557-x.

[2] C. M. Fourie, "Risk factors associated with first-year students' intention to drop out from a university in South Africa," *J Furth High Educ*, vol. 44, no. 2, pp. 201–215, 2020, doi: 10.1080/0309877X.2018.1527023.

[3] H. S. Park and S. J. Yoo, "Early Dropout Prediction in Online Learning of University using Machine Learning," *International Journal on Informatics Visualization*, vol. 5, no. 4, pp. 347–353, 2021, doi: 10.30630/JOIV.5.4.732.

[4] I. Alqudah, M. Barakat, S. M. Muflih, and A. Alqudah, "Undergraduates' perceptions and attitudes towards online learning at Jordanian universities during COVID-19," *Interactive Learning Environments*, pp. 1–18, 2021, doi: 10.1080/10494820.2021.2018617.

[5] S. Bali and M. C. Liu, "Students' perceptions toward online learning and face-to-face learning courses," in *Journal of Physics: Conference Series*, 2018, p. 012094. doi: 10.1088/1742-6596/1108/1/012094.

[6] M. Mather and A. Sarkans, "Student Perceptions of Online and Face-to-Face Learning," *International Journal of Curriculum and Instruction*, vol. 10, no. 2, pp. 61–76, 2018.

[7] F. Ferri, P. Grifoni, and T. Guzzo, "Online learning and emergency remote teaching: Opportunities and challenges in emergency situations," *Societies*, vol. 10, no. 4, p. 86, 2020, doi: 10.3390/soc10040086.

[8] A. Tayebi, J. Gomez, and C. Delgado, "Analysis on the Lack of Motivation and Dropout in Engineering Students in Spain," *IEEE Access*, vol. 9, pp. 66253–66265, 2021, doi: 10.1109/ACCESS.2021.3076751.

[9] J. Gabalán-Coello, A. L. Balcero-Molina, F. E. Vasquez Rizo, A. Martínez-González, and G. Fonseca-Grandón, "An Analysis of Accredited Colombian Universities, Based on Performance Variables Associated with Their Quality," *J Lat Educ*, pp. 1–9, 2019, doi: 10.1080/15348431.2019.1665523.

[10] E. Sneyers and K. De Witte, "The interaction between dropout, graduation rates and quality ratings in universities," *Journal of the Operational Research Society*, vol. 68, no. 4, pp. 416–430, 2017, doi: 10.1057/jors.2016.15.

[11] Q. Kabashi, I. Shabani, and N. Caka, "Analysis of the student dropout rate at the Faculty of Electrical and Computer Engineering of the University of Prishtina, Kosovo, from 2001 to 2015," *IEEE Access*, vol. 10, pp. 68126–68137, 2022, doi: 10.1109/access.2022.3185620.

[12] J. Jacqmin and M. Lefebvre, "The effect of international accreditations on students' revealed preferences: Evidence

from French Business schools," *Econ Educ Rev*, vol. 85, p. 102192, 2021, doi: 10.1016/j.econedurev.2021.102192.

[13] A. Acevedo-De-los-Ríos and D. R. Rondinel-Oviedo, "Impact, added value and relevance of an accreditation process on quality assurance in architectural higher education," *Quality in Higher Education*, vol. 28, no. 2, pp. 186–204, 2022, doi: 10.1080/13538322.2021.1977482.

[14] M. A. S. Mustapa, M. Ibrahim, and A. Yusoff, "Engaging Vocational College Students through Blended Learning: Improving Class Attendance and Participation," *Procedia Soc Behav Sci*, vol. 204, pp. 127–135, 2015, doi: 10.1016/j.sbspro.2015.08.125.

[15] R. Kadar, S. B. Mahlan, M. Shamsuddin, J. Othman, and N. A. Wahab, "Analysis of Factors Contributing to the Difficulties in Learning Computer Programming among Non-Computer Science Students," in *2022 IEEE 12th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, IEEE, 2022, pp. 89–94. doi: 10.1109/ISCAIE54458.2022.9794546.

[16] Y. P. Huang and Y. M. Huang, "Programming language learning supported by an accredited course strategy," in *2013 IEEE 13th International Conference on Advanced Learning Technologies*, IEEE, 2013, pp. 327–329. doi: 10.1109/ICALT.2013.101.

[17] Z. Li, Z. Jie, and H. Daming, "Design and implementation of student programming profile-based teaching aids solution in introductory programming course," in *2020 15th International Conference on Computer Science & Education (ICCSE)*, 2020, pp. 383–390. doi: 10.1109/ICCSE49874.2020.9201695.

[18] A. Baist and A. S. Pamungkas, "Analysis of Student Difficulties in Computer Programming," *VOLT: Jurnal Ilmiah Pendidikan Teknik ElektroElektro*, vol. 2, no. 2, pp. 81–92, 2017, doi: 10.30870/volt.v2i2.2211.

[19] L. Dombrovskaia, J. P. del Rio, and P. Rodríguez, "Prediction of student's retention in first year of engineering program at a technological chilean university," in *2020 39th International Conference of the Chilean Computer Science Society (SCCC)*, 2020, pp. 34–37. doi: 10.1109/SCCC51225.2020.9281195.

[20] S. Sivakumar, S. Venkataraman, and R. Selvaraj, "Predictive modeling of student dropout indicators in educational data mining using improved decision tree," *Indian J Sci Technol*, vol. 9, no. 4, pp. 1–5, 2016, doi: 10.17485/ijst/2016/v9i4/87032.

[21] S. Roy and A. Garg, "Predicting academic performance of student using classification techniques," in *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*, IEEE, 2017, pp. 568–572. doi: 10.1109/UPCON.2017.8251112.

[22] V. L. Miguéis, A. Freitas, P. J. V. Garcia, and A. Silva, "Early segmentation of students according to their academic performance: A predictive modelling approach," *Decis Support Syst*, vol. 115, pp. 36–51, 2018, doi: 10.1016/j.dss.2018.09.001.

[23] H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," *IEEE Access*, vol. 8, pp. 55462–55470, 2020, doi: 10.1109/ACCESS.2020.2981905.

[24] H. P. Singh and H. N. Alhulail, "Predicting Student-Teachers Dropout Risk and Early Identification: A Four-Step Logistic Regression Approach," *IEEE Access*, vol. 10, pp. 6470–6482, 2022, doi: 10.1109/ACCESS.2022.3141992.

[25] Harwati, R. I. Virdyanawaty, and A. Mansur, "Drop out Estimation Students based on the Study Period: Comparison between Naïve Bayes and Support Vector Machines Algorithm Methods," in *IOP Conference Series: Materials Science and Engineering*, 2016, p. 012039. doi: 10.1088/1757-899X/105/1/012039.

[26] S. Mutrofin, A. M. Khalimi, E. Kurniawan, R. V. H. Ginardi, C. Fatichah, and Y. A. Sari, "Detection of

Potentially Students Drop out of College in Case of Missing Value Using C4.5," in *2019 International Conference on Sustainable Engineering and Creative Computing (ICSECC)*, IEEE, 2019, pp. 349–354. doi: 10.1109/ICSECC.2019.8907014.

[27] M. Utari, B. Warsito, and R. Kusumaningrum, "Implementation of Data Mining for Drop-Out Prediction using Random Forest Method," in *2020 8th International Conference on Information and Communication Technology (ICoICT)*, IEEE, 2020, pp. 1–5. doi: 10.1109/ICoICT49345.2020.9166276.

[28] M. I. Sa'ad, Kusrini, and M. S. Mustafa, "Student Prediction of Drop out Using Extreme Learning Machine (ELM) Algorithm," in *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)*, IEEE, 2020, pp. 1–6. doi: 10.1109/ICORIS50180.2020.9320831.

[29] I. M. S. Bimantara and I. M. Widiartha, "Optimization of K-Means Clustering Using Particle Swarm Optimization Algorithm for Grouping Traveler Reviews Data on Tripadvisor Sites," *Jurnal Ilmiah KURSOR*, vol. 12, no. 1, pp. 1–10, 2023. doi: 10.21107/kursor.v12i01.269

[30] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," *Healthcare Analytics*, vol. 2, p. 100060, 2022.

[31] S. Adi, Y. Pristyanto, and A. Sunyoto, "The best features selection method and relevance variable for web phishing classification," in *2019 International Conference on Information and Communications Technology (ICOIACT)*, IEEE, 2019, pp. 578–583. doi: 10.1109/ICOIACT46704.2019.8938566.

[32] S. Wild and L. S. Heuling, "Student dropout and retention: An event history analysis among students in cooperative higher education," *Int J Educ Res*, vol. 104, p. 101687, 2020, doi: 10.1016/j.ijer.2020.101687.

[33] M. K. Morampudi, N. Gonthina, V. D. Reddy, and K. S. Rao, "Analyzing Student Performance in Programming Education Using Classification Techniques," in *2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*, IEEE, 2022. doi: 10.1109/ASSIC55218.2022.10088377.

[34] R. Anantama, H. Suyono, and M. Aswin, "Application of Cost-Sensitive Convolutional Neural Network for Pneumonia Detection," *Jurnal Ilmiah KURSOR*, vol. 11, no. 3, pp. 101–108, 2022. doi : 10.21107/kursor.v11i3.264

[35] I. P. B. W. Brata and I. D. M. B. A. Darmawan, "Neural Network Backpropagation for Kendang Tunggal Tone Classification," *Jurnal Ilmiah KURSOR*, vol. 11, no. 2, pp. 63–74, 2021. doi : 10.21107/kursor.v11i2.258.

[36] H. A. Rosyid, A. Maulana, and U. Pujianto, "Can K-Nearest Neighbor Method Be Used To Predict Success in Indonesia State University Student Selection," *Jurnal Ilmiah KURSOR*, vol. 9, no. 4, pp. 137–144, 2018. doi : 10.28961/kursor.v9i4.186

[37] V. N. Wijayaningrum, A. P. Kirana, I. K. Putri, and T. O. Satrio, "Prediction of Student Academic Performance in Practicum Courses Based on Activity Logs and Student Background," in *2022 International Conference on Electrical and Information Technology*, IEEE, 2022, pp. 366–371. doi: 10.1109/IEIT56384.2022.9967888.

[38] F. Thiele, A. J. Windebank, and A. M. Siddiqui, "Motivation for using data-driven algorithms in research: A review of machine learning solutions for image analysis of micrographs in neuroscience," *Journal of Neuropathology and Experimental Neurology*, vol. 82, no. 7. Oxford University Press, pp. 595–610, Jul. 01, 2023. doi: 10.1093/jnen/nlad040.

[39] L. Qiu, Y. Liu, Q. Hu, and Y. Liu, "Student dropout prediction in massive open online courses by convolutional neural networks," *Soft comput*, vol. 23, no. 20, pp. 10287–10301, Oct. 2019, doi: 10.1007/s00500-018-3581-3.

[40] E. T. Lau, L. Sun, and Q. Yang, "Modelling, prediction and classification of student academic performance using

artificial neural networks," *SN Appl Sci*, vol. 1, no. 9, pp. 1–10, Sep. 2019, doi: 10.1007/s42452-019-0884-7.

[41] S. C. Tsai, C. H. Chen, Y. T. Shiao, J. S. Ciou, and T. N. Wu, "Precision education with statistical learning and deep learning: a case study in Taiwan," *International Journal of Educational Technology in Higher Education*, vol. 17, no. 12, pp. 1–13, Dec. 2020, doi: 10.1186/s41239-020-00186-2.

[42] E. Ismanto, H. A. Ghani, N. I. M. Saleh, J. Al Amien, and R. Gunawan, "Recent systematic review on student performance prediction using backpropagation algorithms," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 20, no. 3, pp. 597–606, 2022, doi: 10.12928/TELKOMNIKA.v20i3.21963.

[43] V. N. Wijayaningrum, I. K. Putri, A. P. Kirana, M. R. Mubarok, D. M. Harahap, and B. R. Hamesha, "Analisis Performa Seleksi Atribut untuk Menentukan Potensi Mahasiswa Putus Studi [Performance Analysis of Attribute Selection to Determine the Potential of Students Drop Out]," *Jurnal Informatika Polinema*, vol. 9, no. 2, pp. 237–243, 2023.

[44] C. Kaope and Y. Pristyanto, "The Effect of Class Imbalance Handling on Datasets Toward Classification Algorithm Performance," *MATRIK: Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, vol. 22, no. 2, pp. 227–238, Mar. 2023, doi: 10.30812/matrik.v22i2.2515.

[45] D. J. Maulana, S. Saadah, and P. E. Yunanto, "Kmeans-SMOTE Integration for Handling Imbalance Data in Classifying Financial Distress Companies using SVM and Naïve Bayes," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 8, no. 1, pp. 54–61, 2024, doi: 10.29207/resti.v8i1.5150.