2023 International Seminar on Application for Technology of Information and Communication (iSemantic) | 979-8-3503-3921-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/iSemantic59612.2023.10295372

Classifying Student Achievement Using K-NN Based on Feature Normalization Techniques

Yuni Yamasari
Department of Informatics
Universitas Negeri Surabaya
Surabaya, Indonesia
yuniyamasari@unesa.ac.id

Rafif Aydin Ahmad
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
5025231198@student.its.ac.id

Abstract- The existence of students' achievement classification with high performance is very important to support teachers in the evaluation of the process of learning. However, not much research has addressed this issue. Accordingly, our research concentrate on the establishment of a classification model for students' achievement with high performance. To achieve this goal, we explore three feature normalization techniques (feature standardized, feature scaled, and feature centered), which are combined with the K-NN method. The K-NN parameters are set with Euclidean distances and the number of neighbors 2, 4, and 6.Furthermore, the evaluation technique uses cross-validation with 2, 3, 5, 10, and 20 folds. The experimental results reveal that the combination of both normalization techniques and K-NN produces the best performance in terms of average accuracy, namely: Standard_K-NN and Scale_K-NN about 84.52% by the number of neighbor = 2. This shows that both techniques are more appropriate to the characteristic of our student data, which is indicated by their accuracy average outperforming the Center_K-NN method by 1.2%**.**

Keywords-classification, student, achievement, K-NN, normalization

I. INTRODUCTION

The pandemic has accelerated the use of Information Communication and Technology (ICT) in many fields, especially in the field of Education [1]. This application of ICT generates a lot of stored data that one of which is student data. On the other hand, techniques of data mining is implemented to the student data to produce the information or knowledge that we need [2]. They are for example information about student performance [3][4][5][6][7][8], and about student behavior [9] [10][11][12], etc.

In addition, information relating to student achievement is also very important to generate. For teachers, this information can help them in knowing the progress of students in learning. It allows them to identify students' strengths and weaknesses, and to plan appropriate remedial actions. In addition, teachers can find out the effectiveness of teaching. This information can be used to see whether the teaching is successful or not. If it is not, teachers can make adjustments in their approach to improving student learning outcomes with this information.

Some previous studies that concentrate on this field include the evaluation of student achievement by analyzing fuzzy characteristics [13], prediction and classification of student achievements in the field of engineering [14], prediction of student achievement based on motivation in vocational schools [15], etc.

Given the importance of this domain, of course, the existence of student achievement models that have optimal performance is indispensable. However, previous research is not much interested in this area. The few studies that have attempted to perform various techniques to improve the performance of models in this domain include this research [16], using feature extraction in student achievement clustering, and [17] using ensemble learning on student achievement prediction.

Therefore, our research focuses to improve the model performance using selecting an appropriate feature normalization technique on student data related to student achievement. We explored 3 techniques, namely: feature standardized, feature scaled, and feature centered. Then, K-NN is applied as a classification method. Further, we evaluate the cross-validation using 5 metrics, namely: AUC, F1, Precision, Accuracy, and Recall. Lastly, our paper is organized into 4 parts: the introductory paragraph, material and method, the result, and the conclusion.

II. MATERIALS AND METHOD

This chapter explains the student data used in this paper. Then, the discussion continues the steps taken to explore which normalization technique best suits the data we use as illustrated in Figure 1. The explanation is as follows:

Step 1: Student data

The student data in this paper is collected from the student's learning achievement when they join statistics courses using virtual classes. There are 66 students involved in this study. The student data comprise of 7 features as presented in Table I. All features are of numerical type.

Step 2: Feature normalization techniques

This step is a data pre-processing step. Feature normalization techniques are applied, namely: standardizing, centering, and feature scaling. The first technique is standardization (or z-score normalization). This technique is a process used to transform the data values into a common scale. To do this, subtract the average of each data value and divide the result by the standard deviation of each data value. The standardization formula is as follows:

$$x_{standardized} = \frac{x - mean}{standard\ deviation} \quad (1)$$

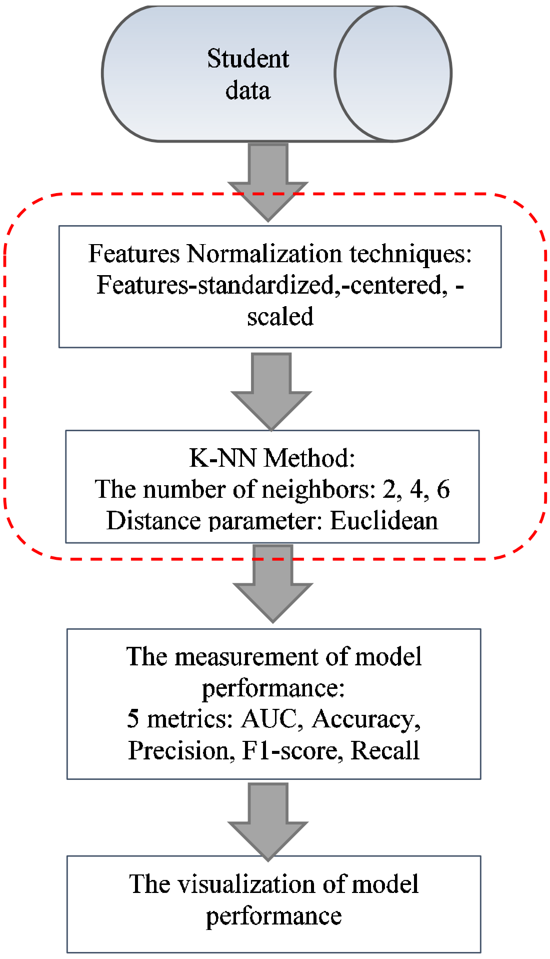
Student data
Features Normalization techniques: Features-standardized,-centered, -scaled
K-NN Method:
The number of neighbors: 2, 4, 6
Distance parameter: Euclidean
The measurement of model performance:
5 metrics: AUC, Accuracy, Precision, F1-score, Recall
The visualization of model performance

Fig. 1. The proposed method

TABLE I. STUDENT DATA

Feature	Data Type	Description
Presence	Numeric	Percentage of students who attend statistics classes
Part	Numeric	The score of students' activities during the course joining, for example: answering, questions, asking, etc
Task	Numeric	Assignment scores obtained by students in statistics courses.
MidExam	Numeric	The midterm exam scores obtained by students in statistics courses
FinExam	Numeric	The final semester exam scores obtained by students in statistics courses.
FinScorer	Numeric	The final score obtained by students in statistics courses.
Grade	Categorical	The numeric values converted to letters

Where:

- x is the data value to be standardized
- $mean$ is the average of the data values
- $standard\ deviation$ is the standard deviation of the data values
-$x\_standardized$ is the value that has been standardized

By using the standardization formula, we can convert data values that have different ranges and scales into the same scale, making it easier to analyze the data.

The second technique is feature scaling by applying min-max scaling. This technique is one of the data scaling techniques in data pre-processing. This technique rescales the data values into a certain range, which is from 0 to 1. In min-max scaling, each data value is reduced by the minimum value of all data and then divided by the difference between the maximum and minimum values. The min-max scaling formula is as follows:

$$x\_scaled = (x - min) / (max - min) \quad (2)$$

Where:

- x is the data value that will be scaled
- min is the minimum value of all data
- max is the maximum value of all data
- x_scaled is the scaled data value

Min-max scaling can be used to rescale data values that have different ranges into the same scale, making it easier to analyze the data. However, this technique is prone to outliers, as extreme data values can affect the resulting scaling range. Therefore, it is necessary to handle outliers before performing min-max scaling.

The last normalization technique in this paper is feature centering. Data centering is a data pre-processing process that aims to change data values by shifting all data so that the average value becomes zero. In data centering, the average value of all data is calculated first, then the average value is subtracted from each data value. The data-centering formula is as follows:

$$x\_centered = x - mean \quad (3)$$

Where:

- x is the data value to be centered
- $mean$ is the average of the data values
- $x\_centered$ is the centered data value

After the student data is normalized, classification methods can be applied, One of which is the K-NN method.

Step 3: Application of K-NN

K-Nearest Neighbors (K-NN) is one of the methods to classify student achievement data. K-NN is a non-parametric method, meaning there are no assumptions about data distribution. K-NN works by finding the K nearest neighbors of new student data and determining the class label by looking at the majority of the nearest neighbors. K is a parameter that must be determined before the K-NN model is applied.

TABLE II. MODEL PERFORMANCE USING 4 METRICS ON ALL METHODS

		neighbor 6				neighbor 4				neighbor 2			
	Fold	AUC	F1	Precision	Recall	AUC	F1	Precision	Recall	AUC	F1	Precision	Recall
Standar_K-NN	2	0.936	0.768	0.752	0.788	0.919	0.841	0.847	0.864	0.901	0.813	0.842	0.818
	3	0.942	0.856	0.845	0.879	0.935	0.828	0.836	0.848	0.925	0.849	0.857	0.848
	5	0.928	0.79	0.79	0.818	0.92	0.806	0.819	0.833	0.929	0.861	0.884	0.864
	10	0.925	0.792	0.78	0.818	0.919	0.777	0.79	0.803	0.925	0.849	0.857	0.848
	20	0.924	0.781	0.78	0.818	0.913	0.793	0.807	0.818	0.909	0.849	0.857	0.848
Scale_K-NN	2	0.936	0.768	0.752	0.788	0.919	0.841	0.847	0.864	0.901	0.813	0.842	0.818
	3	0.942	0.856	0.845	0.879	0.935	0.828	0.836	0.848	0.925	0.849	0.857	0.848
	5	0.928	0.79	0.79	0.818	0.92	0.806	0.819	0.833	0.929	0.861	0.884	0.864
	10	0.925	0.792	0.78	0.818	0.919	0.777	0.79	0.803	0.925	0.821	0.857	0.848
	20	0.924	0.781	0.78	0.818	0.913	0.793	0.807	0.818	0.909	0.849	0.857	0.848
Center_K-NN	2	0.931	0.715	0.695	0.742	0.913	0.777	0.765	0.803	0.902	0.791	0.789	0.803
	3	0.941	0.824	0.824	0.848	0.934	0.824	0.819	0.848	0.934	0.854	0.847	0.864
	5	0.935	0.789	0.767	0.818	0.935	0.794	0.794	0.818	0.945	0.821	0.812	0.833
	10	0.93	0.777	0.769	0.803	0.938	0.816	0.822	0.833	0.916	0.849	0.812	0.833
	20	0.928	0.777	0.769	0.803	0.936	0.816	0.822	0.833	0.916	0.821	0.812	0.833

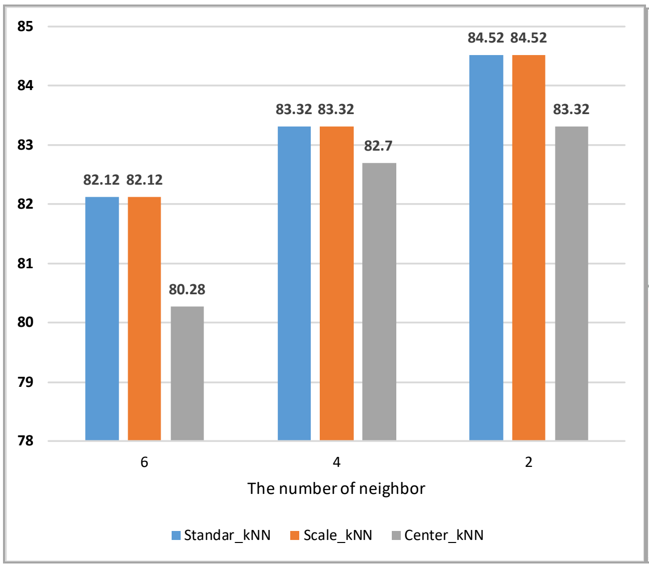

Fig.2. The comparison of accuracy average on all methods

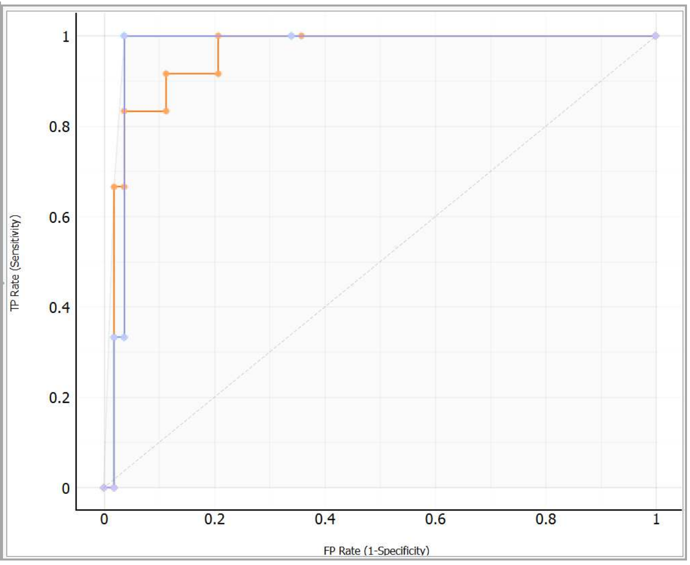

Fig.3. The ROC on all methods

The K-NN process is done by calculating the distance, in this study using Euclidean, between the new data and each existing data in the student achievement data. After the distance is calculated, K-NN will select the K closest data to the new data. Next, the model will take the majority of class labels from the nearest neighbors as the new data class label.

K-NN was applied to student data with 7 attributes. In this case, the distance between the data is calculated based on the difference between each attribute. Normalization of the attributes in the previous step is done to avoid bias on attributes that have a larger range of values compared to other attributes.

Step 4: Performance Evaluation

Performance evaluation of this model is a process to evaluate how well the model that has been built can work in performing the given task. This is critical to ensuring that the model generated can be used effectively for its intended purpose. A variety of indicators are used to assess model performance. In this paper, we use Accuracy, F1-score, Area Under the Curve (AUC), Precision, and Recall.

- Accuracy is the proportion of true predictions made by the model out of all forecasts.
- Precision is the percentage of correct predictions made by the model out of all positive predictions.
- Recall is defined as the proportion of correct predictions to the total number of positive classes that the model should have uncovered.
- AUC is a model's overall performance in distinguishing between positive and negative data classes

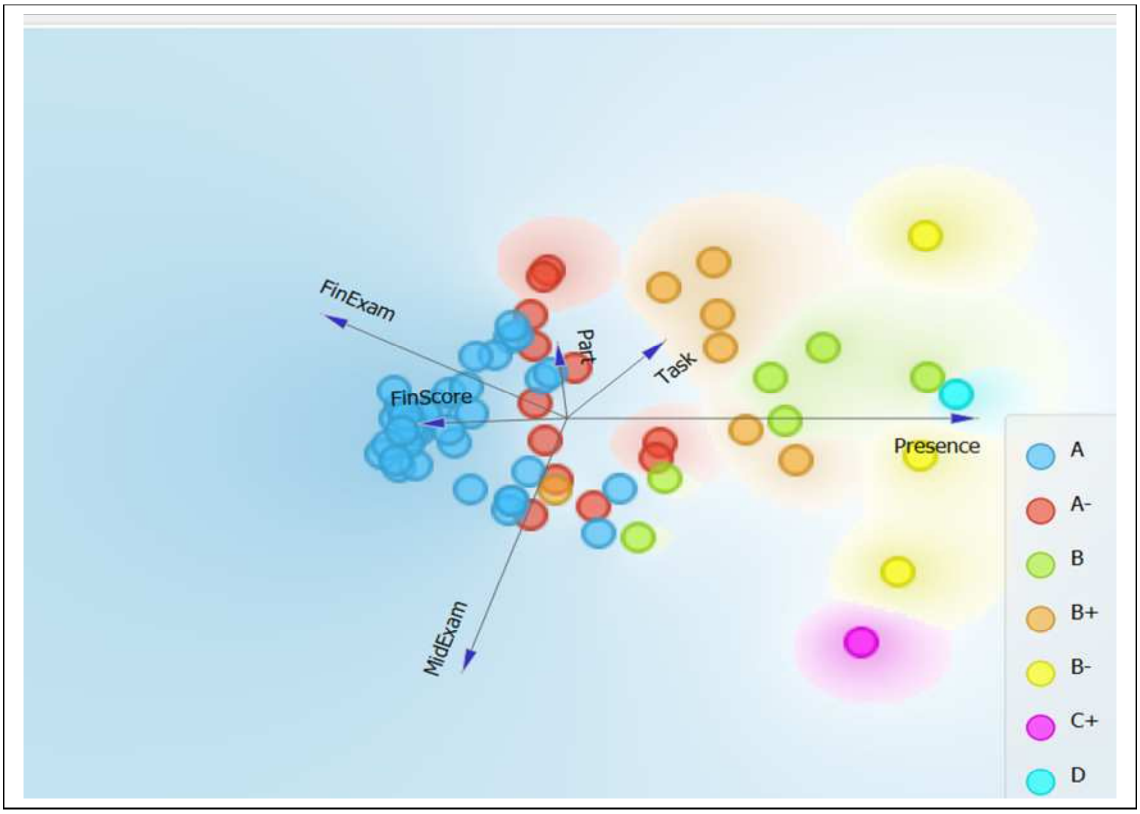

Fig.4. The visualization of the best performance model

- The harmonic mean of Precision and Recall is calculated for generating the F1-score.

It is critical to choose metrics that are suited to the task at hand and the features of the data utilized while evaluating the performance of a model. In addition, this paper uses a cross-validation technique that can help to objectively evaluate the model by dividing the student data into sections for training and testing.

Step 5: Visualization of Results

This step makes us understand the performance of the model. In this paper, we use FreeViz which allows us to visualize the relationships between variables in complex student achievement data, making it easier to understand and analyze the data. In addition, FreeViz can also select two or three variables to be plotted in three or more dimensional graphs.

III. RESULT

The proposed method is implemented in this chapter, and the model is then measured and visualized to better understand its performance.

A. The measurement result of model performance

Once the model has been built, measurements need to be taken to determine the performance model. For that, this paper applies several scenarios. Three feature normalization techniques are tested, namely: standardized features, scaled features, and centered features. Furthermore, we determine the parameters of K-NN as follows:

(a). Neighbor K is 2, 4 and 6

(b). The distance parameter is Euclidean

Meanwhile, the assessment method used in this work is cross-validation. The five measures are Area Under Curve (AUC), Precision, F1-Score, Recall, and Accuracy. Table II displays the findings for four measures: AUC, Precision, F1-Score, and Recall. Overall, the results of both approaches, Standard_K-NN and Scale_K-NN, show the same results. In terms of the widest AUC, the Center_K-NN method achieves the best results around 0.945 in fold 5 and neighbor 2. The F1, precision, and recall values for both techniques, Standard_KNN and Scale_K-NN, are 0.861, 0.884, and 0.879, respectively. This is accomplished in the scenarios fold 5 and neighbors 2; fold 5 and neighbors 2; fold 3 and neighbors 6. In contrast, the fold 2 and neighbor 2 generate the lowest AUC for both the Standard_K-NN and Scale_K-NN techniques about 0.901. The Center_K-NN technique then achieves F1, Precision, and Recall of 0.715, 0.695, and 0.742 in the fold 2 neighbors 6 cases, respectively.

The average accuracy for the accuracy measure is displayed in Fig.2. The average accuracy is attained with a precision range of 74.2 to 87.9. In the fold 2 and neighbor 6 scenarios, the Center_K-NN technique has the lowest accuracy level of 74.2. In the fold 3 and neighbor 3 test situations, however, both the Standard_K-NN and Scale_K-NN algorithms obtain the maximum accuracy level. Meanwhile, based on the average accuracy in Fig. 2, both the Standard_K-NN and Scale_K-NN methods outperform the Center_K-NN method in all scenarios. Where the highest average value is achieved by the Standard_K-NN and Scale_K-NN methods of 84.52 in neighbor 2. Conversely, the lowest average value of 80.28 occurs in the Center_K-NN method with neighbor 6.

B. The visualization of model performance

In this subchapter, the model performance is visualized. Performance visualization of models It is critical to comprehend model performance. The Receiver Operating Curve (ROC) is used in this research to plot the false positive rate (FPR) vs the true positive rate (TPR) at different decision thresholds. As illustrated in Fig.3, the ROC curve is used to assess the model's effectiveness in distinguishing between positive and negative classes in student data. ROC shows that both methods, namely: Standard_K-NN and Scale_K-NN (green and blue lines) are more dominating when compared to the Center_K-NN method (brown line). Furthermore, the Area Under Curve for both methods is 0.924 and the Center_K-NN method is 0.913. This further reinforces that both methods have better performance with an area difference of 0.11.

Finally, this paper presents a visualization of the classification results by the best-performing model as shown in Fig.4. Here, we use Freeviz (Free Visualization Tool) which is a free and open-source data visualization tool developed with Python programming language. Fig. 4 shows for students who are in the same class attract (shown by close-neighbor distance) and students who are in different classes repel (far apart). In this research, we base the class label on the letter grade which is of course related to the student's learning achievement.

IV. CONCLUSION

The student achievement model's excellent performance is critical for teachers and students since it produces more accurate information. This classification's performance can be improved by using appropriate feature normalization techniques. Both feature normalization techniques which are combined with the K-NN method, namely:featured-standardized and -scaled, are well suited to our student data. This is seen by the better accuracy level in K-NN when compared to the feature-centered method.

REFERENCES

[1] D. Yulianto and N. M. Mujtahid, "Online Assessment during Covid-19 Pandemic: EFL Teachers' Perspectives and Their Practices," *JET (Journal English Teaching)*, vol. 7, no. 2, pp. 229–242, Jun. 2021, doi: 10.33541/JET.V7I2.2770.

[2] J. Han, J., Kamber, M., & Pei, *Data Mining Concepts and Techniques*. USA: Elsevier, 2012.

[3] S. T. Ahmed, R. Al-Hamdani, and M. S. Croock, "Enhancement of student performance prediction using modified K-nearest neighbor," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 18, no. 4, pp. 1777–1783, Aug. 2020, doi: 10.12928/TELKOMNIKA.V18I4.13849.

[4] K. Deepika and N. Sathyanarayana, "Relief-F and Budget Tree Random Forest Based Feature Selection for Student Academic Performance Prediction," *Int. J. Intell. Eng. Syst.*, vol. 12, no. 1, pp. 30–39, 2019, doi: 10.22266/ijies2019.0228.04.

.

[5] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Van Erven, "Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil," *J. Bus. Res.*, vol. 94, pp. 335–343, Jan. 2019, doi: 10.1016/j.jbusres.2018.02.012.

[6] A. Qoiriah, Y. Yamasari, Asmunin, A. I. Nurhidayat, and R. Harimurti, "Exploring Automatic Assessment-Based Features for Clustering of Students' Academic Performance," *Adv. Intell. Syst. Comput.*, vol. 1383 AISC, pp. 125–134, Dec. 2020, doi: 10.1007/978-3-030-73689-7_13.

[7] A. U. Khasanah, "A review of student's performance prediction using educational data mining techniques," *J. Eng. Appl. Sci.*, vol. 13, 2018, doi: 10.3923/jeasci.2018.5302.5307.

[8] A. I. Adekitan and E. Noma-Osaghae, "Data mining approach to predicting the performance of first year student in a university using the admission requirements," *Educ. Inf. Technol.*, vol. 24, no. 2, pp. 1527–1543, Mar. 2019, doi: 10.1007/s10639-018-9839-7.

[9] B. Yang, Z. Yao, H. Lu, Y. Zhou, and J. Xu, "In-classroom learning analytics based on student behavior, topic and teaching characteristic mining," *Pattern Recognit. Lett.*, vol. 129, pp. 224–231, Jan. 2020, doi: 10.1016/j.patrec.2019.11.023.

[10] M. Jovanovic, M. Vukicevic, M. Milovanovic, and M. Minovic, "Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study," *Int. J. Comput. Intell. Syst.*, vol. 5, no. 3, pp. 597–610, Jun. 2012, doi: 10.1080/18756891.2012.696923.

[11] 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," in *2017 International Conference on Computing, Communication and Automation (ICCCA)*, May 2017, pp. 170–174, doi: 10.1109/CCAA.2017.8229794.

[12] Y. Yamasari, A. Qoiriah, N. Rochmawati, W. Yustanti, A. Wintarti, and T. Ahmad, "Clustering the Students' Behavior on the e-Learning using the Density-Based Algorithm," *Proc. - 2021 Int. Semin. Appl. Technol. Inf. Commun. IT Oppor. Creat. Digit. Innov. Commun. within Glob. Pandemic, iSemantic 2021*, pp. 22–27, Sep. 2021, doi: 10.1109/ISEMANTIC52711.2021.9573234.

[13] C. Hsieh and Y. Chen, "Analyzing the Characteristics of Fuzzy Synthetic Decision Methods on Evaluating Student ' s Academic Achievement — An Empirical Investigation of Junior High School Students in," 2009, doi: 10.1109/AICI.2009.467.

[14] N. Buniyamin, U. bin Mat, and P. M. Arshad, "Educational data mining for prediction and classification of engineering students achievement," in *2015 IEEE 7th International Conference on Engineering Education (ICEED)*, Nov. 2015, pp. 49–53, doi: 10.1109/ICEED.2015.7451491.

[15] J. N. Purwaningsih and Y. Suwarno, "Predicting students achievement based on motivation in vocational school using data mining approach," in *2016 4th International Conference on Information and Communication Technology (ICoICT)*, May 2016, pp. 1–5, doi: 10.1109/ICoICT.2016.7571880.

[16] Y. Yamasari, S. M. S. Nugroho, I. N. Sukajaya, and M. H. Purnomo, "Features extraction to improve performance of clustering process on student achievement," Feb. 2017, doi: 10.1109/ICSEC.2016.7859946.

[17] T. Fang, S. Huang, Y. Zhou, and H. Zhang, "Multi-model Stacking Ensemble Learning for Student Achievement Prediction," *Proc. - Int. Symp. Parallel Archit. Algorithms Program. PAAP*, vol. 2021-December, pp. 136–140, 2021, doi: 10.1109/PAAP54281.2021.9720454.