Effects of Training Data on Prediction Model for Students' Academic Progress

Susana Limanto[1], Joko Lianto Buliali[2]*, Ahmad Saikhu[3]
Informatics Department, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia[1, 2, 3]
Department of Informatics Engineering, Universitas Surabaya, Surabaya, Indonesia[1]

***Abstract*—The ability to predict students' academic performance before the start of the class with credible accuracy could significantly aid the preparation of effective teaching and learning strategies. Several studies have been conducted to enhance the performance of prediction models by emphasizing three key factors: developing effective prediction algorithms, identifying significant predictor variables, and developing preprocessing techniques. Importantly, none of these studies focused on the effect of using different types of training data on the performance of prediction models. Therefore, this study was conducted to evaluate the effects of differences in training data on the performance of a prediction model designed to monitor students' academic progress. The findings showed that the performance of the prediction model was strongly influenced by the heterogeneity of the values of the predictor variables, which should accommodate all the existing possibilities. It was also discovered that the application of training data with different characteristics and sizes did not improve the performance of the prediction model when its heterogeneity was not representative.**

Keywords*—Decision tree; effects of training data; heterogeneity; prediction; students' academic performance

I. INTRODUCTION

The prediction of students' academic performance enables higher education institutions to improve the quality of their graduates. This is because it helps to prepare strategies and allocate resources to assist students at risk of academic failure. However, the benefits of these strategies can be minimal when the prediction results are not provided early or when they have poor performance. It is also important to note that the earlier detection of students having the potential to fail academically can lead to immediate implementation of necessary actions by relevant stakeholders [1]. Therefore, several studies have been conducted on early detection of students' academic performance.

Early detection models were divided into two types which include those developed to screen superior students and those designed to deal with students having the potential to experience academic failure. The first type was commonly used to evaluate prospective students' future academic performance to make informed decisions about their admission [2], [3]. Meanwhile, the models to determine students with the potential to experience academic failure were applied to increase the number of students who graduate on time. Some of those were applied at the beginning of the semester using demographic and academic data of the previous semester [4], [5]. However, the accuracy of these models was found to be lower than the early detection models which were not applied at the beginning of the semester. This was because there were several predictor variables such as test scores that have a significant influence on the model can only be obtained after the course has started [6]. The effort to solve this problem led [6] to develop a two-stage prediction model. The first prediction was made at the beginning of the semester to obtain a list of students with the potential to fail in order to take immediate action while the second was after the midterm exam to monitor their academic progress.

Several studies have been conducted to improve the performance of prediction models with a focus on three aspects which include developing prediction algorithms, obtaining different predictor variables with significant influence on the model, and developing pre-processing techniques. The prediction algorithm was generally developed by integrating specific techniques. For example, an algorithm that integrates Collaborative Filtering and Artificial Immune Systems was developed by [7] to predict student grades based on recommended courses. The results showed that the developed algorithm provided very accurate results. OKC algorithm was also developed by [8] to improve the performance of prediction models using imbalanced datasets without resampling. This algorithm was a hybrid of One-class support vector machine, K-nearest neighbor, and Classification and regression tree algorithms (abbreviated ad OKC). Research [9] performed an ensemble of seven prediction methods with a majority voting technique to improve the accuracy of this prediction model.

Scholars have been striving to identify the factors influencing the performance of existing prediction models in order to enhance their accuracy. This has led to the application of demographic and academic data as variable predictors to predict students' academic performance. The academic data can be divided into secondary and higher education data. Moreover, efforts are currently being made to determine other factors having the potential to strongly influence student academic performance in order to improve the prediction model performance. For example, [10] applied students' motivation, social, and managerial aspects to complement demographic and academic data as predictor variables. The managerial and social aspects were also used by [11] to develop a prediction model. Meanwhile, psychological aspects such as common talents were utilized by [3] to predict the future performance of prospective students. The study also implemented predictor variable weighting to improve the performance of the model. Other studies were also observed to have used psychological aspects in the form of personality to predict academic performance [12]–[15].

Preprocessing has also been applied to improve prediction model performance [11]. It was discovered that [11] added feature selection before the model formation process to eliminate variables with little effect on the performance and to increase the accuracy. A similar attempt was made by [16] through the combination of different feature selection techniques and data transformation to obtain the best prediction model performance. Apart from feature selection, [9], [17], [18] also applied resampling to overcome imbalanced classes to improve model performance.

Several researchers studied the effect of training data to improve the performance of predictive models [19]–[21]. Research conducted by [19] shows that if there is harmonization in the dataset, classifier performance can be improved. So, it is important to choose the appropriate training and testing data for harmonization. The training data order also greatly affected the performance of the classifier [21]. On exactly the same training dataset, classifier performance can vary from 10% to 100%.

These findings showed that several attempts had been made to improve the performance of prediction models but there was a need for more evidence on the effect of using training data with different characteristics on this performance. Therefore, this study was conducted to evaluate the effect of differences in training data on the performance of the prediction model developed to monitor students' academic progress. This article was divided into four parts which include the introduction section followed by the methodology, analysis of the results, and conclusion of the research.

II. Research Methodology

This research was conducted using the simulation method through the steps shown in Fig. 1. Data were collected through the questionnaires distributed to students at a private higher education institution in Surabaya and academic data retrieved from the higher education information system. The data collected were prepared according to the trial scenario to be implemented in the subsequent process. This was followed by the development of the prediction model according to the existing scenario. Finally, the model's performance was evaluated with a focus on the differences in the training data used in developing the model.

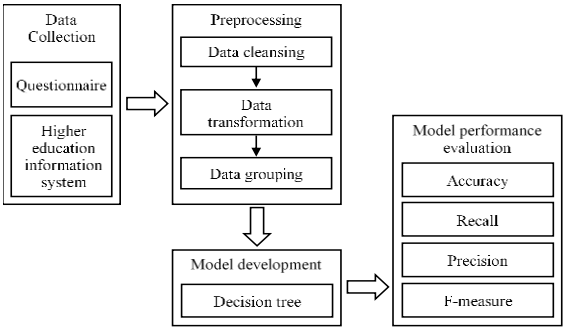

Fig. 1. Research method.

A. Participants and Datasets

The dataset was compiled from 246 private higher education students in Surabaya, Indonesia. The focus was on three different academic years from 2019-2020 to 2021-2022 and three programs including Informatics Engineering, Business Information Systems, and Multimedia. The data obtained were integrated into the student's academic data to form 3,169 instances after the preprocessing stage. Student academic data was collected from 87 subjects grouped in 22 knowledge areas.

TABLE I. Predictor Variable Used

Predictor Variable	Data Type	Details
Program (X_1)	Nominal	Informatics Engineering, Business Information Systems, and Multimedia
Sex (X_2)	Nominal	Male, female
Age (X_3)	Numeric	16, 17, …
Father's academic background (X_4)	Ordinal	Not in school, Elementary school, Middle school, High school, Undergraduate, Masters, Doctoral degree
Mother's academic background (X_5)	Ordinal	Not in school, Elementary school, Middle school, High school, Undergraduate, Masters, Doctoral degree
College entrance path (X_6)	Nominal	Regular, Collaboration, non-academic achievement scholarships, academic achievement scholarships
The specialization chosen while in high school (X_7)	Nominal	Natural Sciences, Social Sciences, Languages, others
High school city (X_8)	Nominal	Surabaya, Java outside Surabaya, outside Java
Average Math score at 12th grade (X_9)	Numeric	66.5 - 99
Average English score at 12th grade (X_{10})	Numeric	70-100
The Average score at 12th grade (X_{11})	Numeric	73.6-97.93
Cumulative credits (X_{12})	Numeric	Total credits that have been collected since the first semester
CGPA (X_{13})	Numeric	0-4
Number of courses taken in the semester (X_{14})	Numeric	3-9
Number of course participants (X_{15})	Numeric	4-108
Previous course grade (X_{16})	Ordinal	A-E
Length of time repeat (X_{17})	Numeric	How many semesters before, last time taking the course
Position difference (X_{18})	Numeric	The difference between the course positions taken in the curriculum and the student's current position
Pass percentage (X_{19})	Numeric	The percentage of passing of this course in previous class by the same lecturer
Prerequisite grade (X_{20})	Ordinal	A-E
Prerequisite time taken (X_{21})	Numeric	How many semesters before, was the last time this prerequisite course was taken

This study used two-stage prediction. The first was applied for the early detection of student academic performance in a course. The predictions were made before the start of lectures using 21 predictor variables as shown in Table I. Meanwhile, the target class was divided into two, Pass or Fail. The results obtained can be used by students, lecturers, and higher

education stakeholders to devise appropriate strategies to ensure the success of students that have been predicted to fail at the end of the semester. The second stage was applied after lectures had been received for half a semester and the midterm exam scores had been released. This led to the addition of two predictor variables including the midterm exam scores (range: 0 – 100) (X_{22}) and the number of students' absence from lectures during the first half of the semester (range: 0 – 7) (X_{23}). The results obtained can be used by lecturers to monitor students' academic progress.

B. Preprocessing

This stage was used to remove irrelevant data, overcome missing values, transform data to the format required by the algorithm to be used, and prepare data according to the research scenarios. The first time, irrelevant data such as students' ID, course code, and other attributes not needed as well as those considered to be incomplete and impossible to complete were removed from the dataset. Example of those considered incomplete include are the data related to courses opened for the first time in the program without any record of the pass percentage from lecturers in the previous period.

The data were later transformed by discretizing the continuous data to form categories in order to increase the accuracy of the prediction model [16]. This was followed by the conversion of the categorical data into a numeric form. It was done because the prediction will be carried out using the Decision Tree algorithm with Python programming which requires data to be in a numerical form. Moreover, the application of numeric data had the ability to make some Machine Learning algorithms run efficiently [5].

Finally, the structure of data was prepared to be appropriate for the four scenarios: to evaluate the effect of using training data from certain knowledge areas, different admission years, specific courses, and different training data sizes. The details of each scenario are described in the evaluation section.

C. Model Development and Evaluation

The Decision Tree method was used to develop the prediction models. The method is more popular than the others [1], [22], [23]. Its application was due to the fact that the resulting model tree is usually easy to understand and the conversions are directly in the form of IF-THEN rules.

The dataset consists of 2941 instances for the Pass target class known as major data and 228 instances for the Fail target class classified as minor data. The existence of imbalance in the number of each target class can reduce the performance of the prediction model [8], [24], [25].

The evaluation was conducted to reduce the effect of the target class imbalance using ten-fold stratified cross-validation technique. The data were divided into ten sections with each having a balanced proportion for each target class. It is pertinent to note that one section was used as test data while nine sections were applied as training data. The test process was repeated ten times to avoid deviations using the test data from each section.

The evaluation was conducted using four different scenarios run in different ways as indicated in the following explanations:

1) First scenario: The dataset consists of 22 knowledge areas. The knowledge areas used refer to the Computer Science Curricula 2013 from the ACM – IEEE Computer Society [26]. The dataset was divided into 22 sub-datasets with each containing instances from a particular knowledge area. A total of five knowledge areas with the highest number of instances were used as indicated in the statistics presented in the following Table II. Moreover, a sub-dataset will be compiled which is a combination of those five sub-datasets for comparison.

TABLE II. FIRST SCENARIO SUB-DATASET STATISTICS

Knowledge Areas	Number of Major Class	Number of Minor Class
Programming Languages (PL)	820	85
Computational Science (CN)	628	95
Algorithm and Complexity (AL)	664	58
Software Development Fundamentals (SDF)	471	23
Discrete Structures (DS)	371	88

2) *Second scenario:* The dataset was divided into three sub-datasets with the first containing instances of students that entered higher education in 2019, the second for those in 2020, and the third for those in 2021 as indicated in the following Table III. Moreover, a sub-dataset containing combined instances from the three sub-datasets was compiled for comparison.

TABLE III. SECOND SCENARIO SUB-DATASET STATISTICS

Admission Year	Number of Major Class	Number of Minor Class
2019	226	32
2020	2073	150
2021	642	46

3) *Third scenario:* The dataset was divided into several sub-datasets with each containing instances from specific courses. The five courses with the highest number of instances and the most significant number of minor data were used as indicated in Table IV. Moreover, a sub-dataset containing instances from different courses was compiled for comparison.

TABLE IV. THIRD SCENARIO SUB-DATASET STATISTICS

Courses	Number of Major Class	Number of Minor Class
Algorithm and Programming (AP)	194	39
Statistics (S)	143	64
Discrete Mathematics (DM)	180	18
Web Programming (WP)	120	23
Computer Network (CN)	126	9

4) Fourth scenario: The dataset was divided into eight sub-datasets with 50, 100, 500, 1000, 1500, 2000, 2500, 3000, and 3169 instances respectively. These instances were selected randomly from the dataset using a simple random sampling technique.

The trial conditions from the first to the third scenarios were balanced by ensuring the number of instances in the comparison sub-dataset was equal to the number of training data in the other sub-datasets. The instances in the comparison sub-dataset were selected randomly from the dataset using a stratified random sampling technique while the test data used were the same as those applied for each sub-dataset.

Accuracy was used as the performance measure and it was determined as the ratio of the correctly predicted number of instances to the total number of instances, as indicated in (1).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

True Positive (TP) is the number of instances predicted to fail from instances with the failed target class. True Negative (TN) is the number of instances predicted to pass from instances with the target class passed. False Positive (FP) is the number of instances predicted to fail from instances with the target class passed. False Negative (FN) is the number of instances predicted to pass from instances with the failed target class.

The classes of each predictor and target variable were not distributed equally and this means applying only the accuracy measure can cause confusion [5], [16]. Therefore, F-measure was introduced and it involved calculating the harmonic average between recall and precision values, as indicated in (2), (3), and (4).

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

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

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

III. Results and Discussion

A total of five analyses were conducted in the first scenario as indicated in Tables V and VI. It was discovered that there was an increase in the prediction model's performance from the first to the second stage. This showed that the addition of appropriate predictor variables increased the accuracy as well as the performance of the model for unbalanced data distribution without increasing the amount of minor data. Furthermore, a higher increase in the F-measure value compared to the accuracy value indicated a rapid increment in the predictive ability of minor data compared to the major data.

The prediction model performance of most comparison sub-datasets was found to be better than for specific knowledge areas but the variation was relatively small. This showed that the academic performance of students in certain knowledge area was more clustered. As a result, heterogeneity in the training data in each sub-dataset needed to be distributed appropriately. The combination of different kinds of knowledge areas in the comparison sub-dataset complemented its heterogeneity, thereby, increasing its ability to accommodate heterogeneity in the test data.

Tables VII and VIII show the predicted results for the four models formed according to the second scenario. The pattern of the prediction model's performance from the first to the second stage was found to be the same as the first scenario. It was also discovered that there was a significant increase in the F-measure but not as much as in the first scenario.

The test results for the sub-dataset with a particular admission year showed better performance than the comparison sub-dataset. This means that the heterogeneity in the training data of each sub-dataset was well distributed to accommodate the one in the test data. Meanwhile, the heterogeneity of the datasets for each admission year was found to be different and their combination to form a prediction model reduced the performance compared to the application of datasets from a particular admission year. Meanwhile, the difference was minimal. Based on these results, it can be concluded that student academic performance varies for each year of admission. As a result, the prediction of student performance with a certain admission year will be better if the training data was taken from the same admission year.

TABLE V. First Scenario Test Results (Accuracy)

Knowledge Areas	Accuracy (%)		Comparison Models	
	1st Stage	*2nd Stage*	*1st Stage*	*2nd Stage*
PL	87.85	92.26	89.28	**91.49**
CN	87.68	87.84	87.56	89.90
AL	87.95	93.21	88.50	93.34
SDF	92.53	93.34	**91.11**	94.56
DS	81.27	84.30	81.93	84.31

TABLE VI. First Scenario Test Results (F-measure)

Knowledge Areas	F-measure (%)		Comparison Models	
	1st Stage	*2nd Stage*	*1st Stage*	*2nd Stage*
PL	15.25	56.34	27.41	57.35
CN	46.44	47.74	**44.18**	56.37
AL	16.76	54.82	21.05	56.97
SDF	10.67	37.57	8.33	47.38
DS	42.55	58.60	43.07	**54.43**

TABLE VII. Second Scenario Test Results (Accuracy)

Admission Year	Accuracy (%)		Comparison Models	
	1st Stage	*2nd Stage*	*1st Stage*	*2nd Stage*
2019	84.45	84.85	**80.18**	**83.37**
2020	92.58	93.66	**91.36**	**91.99**
2021	92.88	94.19	**91.42**	**93.17**

TABLE VIII. Second Scenario Test Results (F-measure)

Admission Year	F-measure (%)		Comparison Models	
	1st Stage	*2nd Stage*	*1st Stage*	*2nd Stage*
2019	30.71	42.45	**24.84**	**34.17**
2020	34.03	47.83	**25.71**	**35.12**
2021	37.45	57.43	**19.08**	**37.77**

TABLE IX. THIRD SCENARIO TEST RESULTS (ACCURACY)

Courses	Accuracy (%)		Comparison Models	
	1st Stage	*2nd Stage*	*1st Stage*	*2nd Stage*
AP	74.69	87.10	76.41	**85.36**
S	66.26	76.45	70.60	**72.55**
DM	85.32	90.97	**83.87**	**86.39**
WP	84.71	86.00	**76.95**	88.76
CN	88.90	94.12	91.15	95.60

TABLE X. THIRD SCENARIO TEST RESULTS (F-MEASURE)

Courses	F-measure (%)		Comparison Models	
	1st Stage	*2nd Stage*	*1st Stage*	*2nd Stage*
AP	19.18	63.92	21.21	**43.07**
S	42.42	60.70	46.83	**51.74**
DM	19.00	49.33	19.00	**41.38**
WP	57.57	53.00	**43.65**	63.00
CN	6.67	25.00	24.00	40.00

The test data results for the datasets from specific courses presented in Tables IX and X were observed to be similar to those of the previous evaluation such that the second stage was found to perform better than the first stage, except for the F-measure of the Web course Programming. However, the decrease in the F-measure value of the Web Programming course, 8%, was much lower than the average increase, 178%, for the other course sub-datasets and 68% in the comparison sub-dataset. The average increase in the accuracy value of the five sub-datasets was also estimated at 8%-9% and this was much lower than the F-measure value.

The test results for each particular course sub-dataset did not show any particular pattern when compared with the comparison sub-dataset. In some courses, the model's performance was better than the comparison sub-dataset and this means the heterogeneity in the training data of each sub-dataset was not properly distributed, thereby, indicating the inability to accommodate the heterogeneity in the test data. Moreover, the combination of different kinds of courses in the comparison sub-dataset complemented the heterogeneity in the comparison sub-dataset and this allowed the accommodation of heterogeneity in the test data.

In the fourth test scenario, nine types of analysis were performed using different dataset sizes as indicated in Fig. 2 and Fig. 3. It was discovered that the performance of the prediction model in the second stage was better than the first stage, except for the accuracy of the dataset with 2000 instances. This was observed to have a similar pattern as the first scenario trial.

There was no pattern showing that an increment in the dataset's size led to an increase in the accuracy value. Even though, the most extensive dataset size provided the highest accuracy and F-measure in this trial. This was observed to be in line with the research conducted by [5]. It also showed that a small dataset could provide a credible accuracy rate as long as it had the ability to identify key indicators.

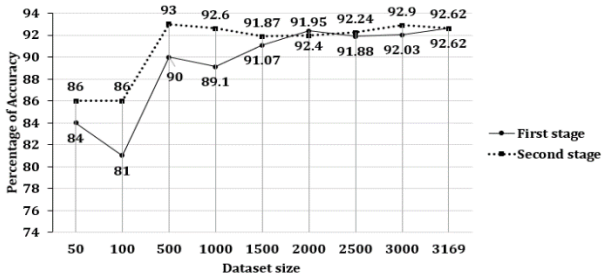

Fig. 2. The accuracy of the results for the 4th scenario.

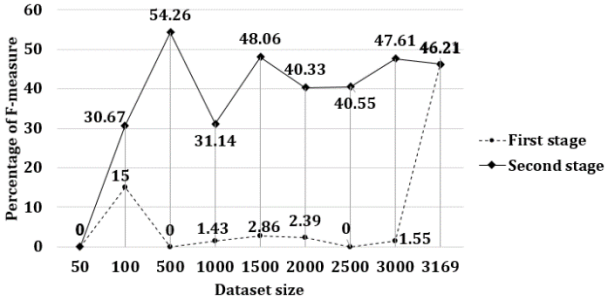

Fig. 3. The F-measure of the results for the 4th scenario.

The tree model generated from datasets of different sizes is presented in Fig. 4, starting from the smallest in Fig. 4(a) to the largest in Fig. 4(i). It was observed from each of the trees produced that the predictor variable played a crucial role in the prediction model and the size of the trees formed was different. This means the model can perform effectively when the training data represent the heterogeneity in the test data. It was also discovered that the training data with a large size but lacks the ability to accommodate heterogeneity in the test data could produce poor performance. An increment in the size of the training data was expected to increase the heterogeneity in order to accommodate the test data but this could not always be achieved. Therefore, the problem can be anticipated through other aspects such as the selection of a suitable predictor variable.

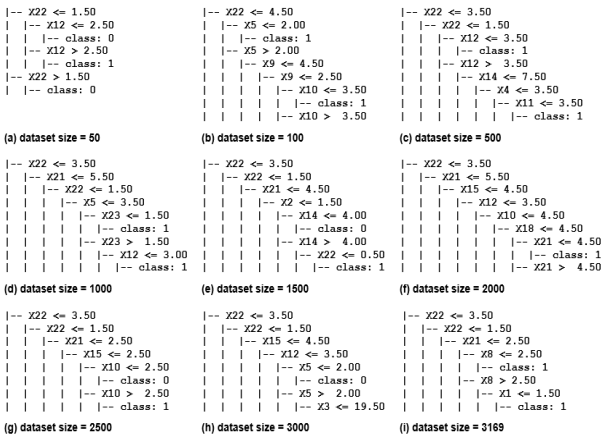

Fig. 4. Trees generated from the 4th scenario.

IV. CONCLUSION

This research was conducted to determine the effect of using different training data on the prediction model of students' academic performance in some courses. The dataset

used in this research was limited to a private higher education institution in East Java, Indonesia which was taken during the Covid 19 pandemic. The prediction model was developed in two stages to be used as a tool to monitor students' academic progress. The trials were conducted based on four different scenarios, including the effect of using sub-datasets from certain knowledge area, admission years, courses, and different dataset sizes. The results showed that the performance of the prediction model in the second stage was mostly better than the first, even though the average accuracy of the first was higher than 80%. These findings showed that the addition of appropriate predictor variables can improve model performance thereby increasing confidence in the results provided by the model to monitor academic progress.

The results showed that the performance of the model was greatly influenced by the heterogeneity of both predictor and target variables, and this was necessary to accommodate all possible outcomes. Therefore, the use of datasets with specific characteristics or sizes can only improve the prediction model's performance when the heterogeneity of the dataset is representative of the larger population. This means, there is a need to ensure the data's heterogeneity is considered to achieve satisfactory performance measures. Research that will be considered further is to develop a synthetic data oversampling strategy to increase the heterogeneity of the dataset so that the performance of the predictive model can be improved.

ACKNOWLEDGMENT

This work was supported by a doctoral dissertation research grant from Indonesian Ministry of Education, Culture, Research and Technology.

REFERENCES

[1] E. Tjandra, S. S. Kusumawardani, and R. Ferdiana, "Student Performance Prediction in Higher Education : A Comprehensive Review," in 3rd International Conference on Informatics, Technology, and Engineering (InCITE), 2021, p.

[2] R. G. Santosa, Y. Lukito, and A. R. Chrismanto, "Classification and Prediction of Students' GPA Using K-Means Clustering Algorithm to Assist Student Admission Process," J. Inf. Syst. Eng. Bus. Intell., vol. 7, no. 1, p. 1, 2021, doi: 10.20473/jisebi.7.1.1-10.

[3] 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.

[4] A. Elbadrawy and G. Karypis, "Domain-aware grade prediction and top-n course recommendation," RecSys 2016 - Proc. 10th ACM Conf. Recomm. Syst., pp. 183–190, 2016, doi: 10.1145/2959100.2959133.

[5] L. M. Abu Zohair, "Prediction of Student's performance by modelling small dataset size," Int. J. Educ. Technol. High. Educ., vol. 16, no. 1, 2019, doi: 10.1186/s41239-019-0160-3.

[6] S. Limanto, J. L. Buliali, and A. Saikhu, "A Two-Stage Early Prediction Model to Monitor the Students' Academic Progress," in 2022 10th International Conference on Information and Communication Technology (ICoICT), 2022, pp. 82–87, doi: 10.1109/ICoICT55009.2022.9914882.

[7] P. C. Chang, C. H. Lin, and M. H. Chen, "A hybrid course recommendation system by integrating collaborative filtering and artificial immune systems," Algorithms, vol. 9, no. 3, 2016, doi: 10.3390/a9030047.

[8] M. R. Ayyagari, "Classification of Imbalanced Datasets using One-Class SVM, k-Nearest Neighbors and CART Algorithm," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 11, pp. 1–5, 2020, doi: 10.14569/IJACSA.2020.0111101.

[9] H. Zeineddine, U. Braendle, and A. Farah, "Enhancing prediction of student success: Automated machine learning approach," Comput. Electr. Eng., vol. 89, no. November 2020, pp. 1–10, 2021, doi: 10.1016/j.compeleceng.2020.106903.

[10] M. A. Yehuala, "Application Of Data Mining Techniques For Student Success And Failure Prediction The Case Of DebreMarkos University," Int. J. Sci. Technol. Res., vol. 4, no. 4, pp. 91–94, 2015.

[11] A. K. Hamoud and A. M. Humadi, "Student's Success Prediction Model Based on Artificial Neural Networks (ANN) and A Combination of Feature Selection Methods," J. Southwest Jiaotong Univ., vol. 54, no. 3, 2019.

[12] N. T. Hendy and M. D. Biderman, "Using bifactor model of personality to predict academic performance and dishonesty," Int. J. Manag. Educ., vol. 17, no. 2, pp. 294–303, 2019, doi: 10.1016/j.ijme.2019.05.003.

[13] M. Komarraju, S. J. Karau, and R. R. Schmeck, "Role of the Big Five personality traits in predicting college students' academic motivation and achievement," Learn. Individ. Differ., vol. 19, no. 1, pp. 47–52, 2009, doi: 10.1016/j.lindif.2008.07.001.

[14] S. V. Paunonen and M. C. Ashton, "On the prediction of academic performance with personality traits: A replication study," J. Res. Pers., vol. 47, no. 6, pp. 778–781, 2013, doi: 10.1016/j.jrp.2013.08.003.

[15] A. Vedel and A. Poropat, "Encyclopedia of Personality and Individual Differences," in Encyclopedia of Personality and Individual Differences, no. Januari, 2017.

[16] G. Akçapınar, A. Altun, and P. Aşkar, "Using learning analytics to develop early-warning system for at-risk students," Int. J. Educ. Technol. High. Educ., vol. 16, no. 1, 2019, doi: 10.1186/s41239-019-0172-z.

[17] N. Hutagaol and Suharjito, "Predictive modelling of student dropout using ensemble classifier method in higher education," Adv. Sci. Technol. Eng. Syst., vol. 4, no. 4, pp. 206–211, 2019, doi: 10.25046/aj040425.

[18] T. Fahrudin, J. L. Buliali, and C. Fatichah, "Predictive modeling of the first year evaluation based on demographics data: Case study students of Telkom University, Indonesia," Proc. 2016 Int. Conf. Data Softw. Eng. ICoDSE 2016, pp. 0–5, 2017, doi: 10.1109/ICODSE.2016.7936158.

[19] M. K. Uçar, M. Nour, H. Sindi, and K. Polat, "The Effect of Training and Testing Process on Machine Learning in Biomedical Datasets," Math. Probl. Eng., vol. 2020, pp. 1–17, 2020, doi: 10.1155/2020/2836236.

[20] J. Lin, A. Zhang, M. Lecuyer, J. Li, A. Panda, and S. Sen, "Measuring the Effect of Training Data on Deep Learning Predictions via Randomized Experiments," in Proceedings of Machine Learning Research, 2022, vol. 162, pp. 13468–13504, doi: https://doi.org/10.48550/arXiv.2206.10013.

[21] J. Mange, "Effect of training data order for machine learning," in Proceedings - 6th Annual Conference on Computational Science and Computational Intelligence, CSCI 2019, 2019, pp. 406–407, doi: 10.1109/CSCI49370.2019.00078.

[22] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: literature review and best practices," Int. J. Educ. Technol. High. Educ., vol. 17, no. 1, 2020, doi: 10.1186/s41239-020-0177-7.

[23] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," Procedia Comput. Sci., vol. 72, pp. 414–422, 2015, doi: 10.1016/j.procs.2015.12.157.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[25] P. N. Tan, M. Steinbach, A. Katpatne, and V. Kumar, Introduction to Data Mining, 2nd ed. the United States of America: Pearson Education, Inc, 2019.

[26] ACM and IEEE, CS2013: Computer Science Curricula 2013, vol. 48, no. 3. ACM and IEEE, 2015.