Enhancing the Readability of Academic Data for Machine Learning through Preprocessing Techniques

Anna Mayyah Soraya
Electrical Engineering Department
Sebelas Maret University
Surakarta, Indonesia
mayyahsor@student.uns.ac.id

Faisal Rahutomo *
Electrical Engineering Department
Sebelas Maret University
Surakarta, Indonesia
faisal_r@staff.uns.ac.id

Miftahul Anwar
Electrical Engineering Department
Sebelas Maret University
Surakarta, Indonesia
miftahwar@staff.uns.ac.id

*Corresponding author: faisal_r@staff.uns.ac.id

Received: 2025-05-29; **Accepted:** 2026-01-20

Abstract—**Academic data plays a central role in supporting decision-making in educational institutions. However, the successful implementation of machine learning to analyze and make predictions based on academic data highly depends on the quality and readability of the data. To fully harness the potential of machine learning, careful preprocessing of educational data is essential. This research aims to design and implement preprocessing techniques, including imputation, winsorizing, and data dropping, on academic datasets. To handle missing values, the multivariate imputation by chained equations method is employed, utilizing three different algorithms: linear regression, random forest, and KNN. The accuracy of these algorithms in predicting missing values is then compared. Additionally, the winsorizing method is applied to outliers, and duplicate data is addressed by removing duplicate entries. Based on the testing results through evaluation metrics, these preprocessing techniques can improve model accuracy by 0.037 for MAE, 0.11 for RMSE, and 0.006 for MSE. The processed data allows the model to function more optimally and produce more reliable results.**

Keywords— Evaluation Metrics; Machine Learning; MICE; Preprocessing; Winsorizing

I. Introduction

Data is a valuable asset in higher education that must be maintained even after students have graduated for years to come. The university maintains a vast amount of condition data in the Higher Education Database (HEDB). The HEDB is a data storage system overseen by the Center for Data and Information (Pusat Data dan Informasi, or Pusdatin in Indonesian) under the Ministry of Technology, Research, and Higher Education. (Kementerian Riset Teknologi dan Pendidikan Tinggi/ Kemenristekdikti in Indonesian) [1]. The number of UNS students in the 2023 intake is 10,409 students. If calculated from this figure, it can be seen that there will be approximately 40,000 data points for each of the four intakes. Of course, collecting, analyzing, and reporting student academic data from a large student population can be a very complex and time-consuming task if done manually. To overcome this, various types of data can be collected using techniques such as data warehouses, big data, and machine learning. [2], [3], [4].

In the era of information technology and artificial intelligence development, academic data has a central role in supporting decision-making in educational institutions. However, the success of implementing machine learning to analyze and make predictions based on scholarly data is highly dependent on the quality and readability of the data. [5]. Primarily, when data contains redundancy, repetition, noise, and other anomalies, it makes it difficult for the algorithm to understand and process information, thereby hindering improved performance. Optimal data readability is crucial for efficient machine learning and accurate solutions. [6]. To fully utilize the potential of machine learning, a careful preprocessing process is necessary for this academic data. Preprocessing techniques play a crucial role in enhancing data readability, which in turn strengthens the quality of analysis and predictions produced by machine learning. [7]. This study aims to explore and apply various effective preprocessing techniques to improve the readability of academic data in machine learning.

II. Research Method

A. Machine Learning

Machine learning is a computer programming process to optimize performance criteria using examples of data or past experiences. [8]. Machine learning may not be able to identify the entire process that exists, but it can make accurate and functional estimates. The estimate may not accurately represent or approximate all the original values, but it can still explain certain aspects of the data. Through the identification and analysis of data, machine learning can detect specific patterns and regularities to produce estimates. Machine learning has a model that is determined from several parameters. This model can be predictive to make predictions in the future, or it can also be descriptive to gain knowledge from data, and/or both. [8].

B. Preprocessing Data

Data preprocessing is a crucial stage in the data analysis process, aiming to clean, transform, and prepare the data for further analysis. The primary goal is to ensure that the data used in a study or model are of high quality, accurate, and ready for use. Various data preprocessing techniques play an essential role in preparing data for analysis. Some data preprocessing techniques are shown in Fig. 1.

Data cleaning involves removing noise and correcting data inconsistencies to ensure accuracy. Data integration is another approach that combines information from multiple sources into a unified data store, often a data warehouse.

DOI: https://dx.doi.org/10.20961/jeeict.8.1.103104

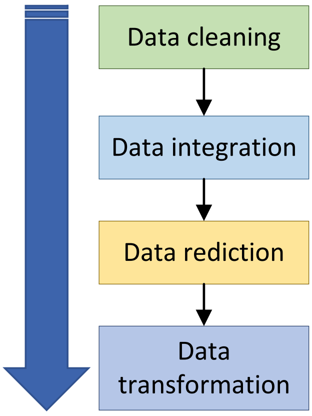


Fig. 1. Preprocessing techniques

Data reduction techniques focus on reducing the size of data, using methods such as aggregation, removing redundant features, or clustering. This helps simplify the data set while retaining its essential characteristics. These data preprocessing techniques are not mutually exclusive, meaning they can be used in combination to optimize the quality and usability of the data. The combination of these techniques ensures that the data is refined and ready for practical analysis, thereby contributing to more accurate and insightful results in data mining and analytics [9].

C. *Multiple Imputation by Chained Equation*

Multiple imputation by chained equations (MICE), also known as wholly conditional specification, has become a prominent method in statistical literature for addressing missing data. This approach imputes missing values using the observed data, assuming that the observed variables are incorporated into the imputation model. MICE is highly versatile and applicable in various contexts. Since multiple imputations generate several predictions for each missing value, the analysis of these imputations accounts for the uncertainty in the imputation process, resulting in accurate standard errors. [10].

The imputed values are essentially predictions that approximate the original data and can be utilized in statistical analysis to estimate parameters and create models. To account for uncertainty, a random component is included in these estimates. A series of parameter estimates is generated and then evaluated independently but consistently. Most statistical software performs multiple imputations based on the missing at random (MAR) mechanism to produce unbiased and accurate estimates of associations from the available data. This method not only effectively estimates the parameters for variables with missing data but also provides a complete forecast for all other variables without missing data [11].

D. *Methodology*

In conducting the research, 6 stages were carried out to create an air quality prediction model, as follows.

1) Problem identification: This process involves identifying and understanding the root cause of the problem, which is the accumulation of student academic data at Universitas Sebelas Maret that is not ready to be processed by available machine learning.

2) Literature research: Literature research is conducted by searching for information involving the collection, analysis, and synthesis of information from relevant literature sources related to methods used to solve data preprocessing, such as missing values, outliers, and data duplication.

3) Collecting data: The data used in this research is academic data of 16,785 UNS students, with details including semester GPAs from 1 to 6, place of origin, high school origin, gender, student ID, cohort, and significant.

4) Preprocessing data: Data cleaning is the primary step in data preprocessing used to handle missing values, smooth data noise, address outliers, and correct inconsistencies. Each dataset has its unique characteristics, so the data cleaning process must be tailored to the specific needs of that dataset. Unclean data can disrupt the data analysis process and produce unreliable results. Therefore, it is crucial to involve several data cleaning procedures to ensure data integrity and reliability before proceeding to further analysis stages [7].

5) Testing: The testing phase is conducted by determining the accuracy level through evaluation metrics. Evaluation metrics are tools used in the data analysis and machine learning process to measure the quality and performance of a model in predicting outcomes [9]. In this paper, we evaluate the model using mean absolute error (MAE), mean squared error (MSE), mean absolute percentage error (MAPE), and root mean squared error (RMSE) [12], [13]. The MAE, MSE, MAPE, and RMSE formulas are shown in (1), (2), (3), and (4), respectively.

$$MAE = \frac{\sum_{t=1}^{n}|X_t - F_t|}{n} \tag{1}$$

$$MSE = \frac{\sum_{t=1}^{n}(X_t - F_t)^2}{n} \tag{2}$$

$$MAPE = \frac{\left(\sum_{t=1}^{n}\left|\frac{X_t - F_t}{X_t}\right|\right)}{n} \tag{3}$$

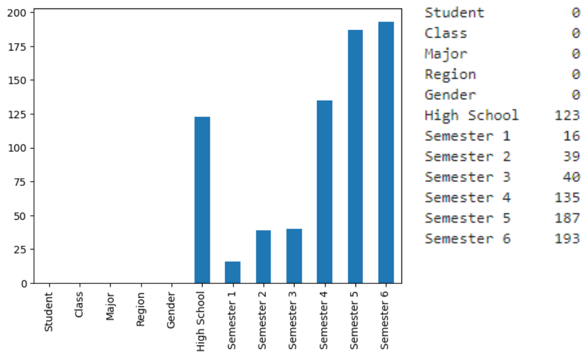
$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{t=1}^{n}(X_t - F_t)^2}{n}} \tag{4}$$

6) Evaluation: Evaluation is conducted after observing the accuracy results from the testing phase to compare which algorithms and techniques are better than others.

III. Experimental Results

A. *Data Exploration*

As one of the frequently used technical analysis methods, time series analysis comes with many variations, but the purpose remains the same [14]. This research uses UNS student academic data from the 2017 to 2019 cohorts. Data were obtained from UPT TIK, which had previously undergone independent data processing, including renaming and converting NIMs to student IDs, to maintain student privacy. In this paper, each data point is assigned a weighted factor for every time series dataset, such as region, major, and high school. This study focuses on handling data disruptions, so data recognition and identifying the amount and location of missing data are



```
Student          0
Class            0
Major            0
Region           0
Gender           0
High School    123
Semester 1      16
Semester 2      39
Semester 3      40
Semester 4     135
Semester 5     187
Semester 6     193
```

Fig. 2. Missing value graph on dataset

necessary. From 16,785 data points with 12 attributes, the amount of missing data is illustrated in the graph in Fig. 2.

Based on Fig. 2, there is a total of 733 missing data points, or 4,4% missing data, with the most missing data in the 6th semester, amounting to 193. In the variables of student ID, class, major, district, and gender, all datasets are complete, with no missing data found. There are 123 missing data points in the high school variable, 16 missing data points in the 1st semester, 39 missing data points in the 2nd semester, 40 missing data points in the 3rd semester, 135 missing data points in the 4th semester, 187 missing data points in the 5th semester, and 193 missing data points in the 6th semester.

B. *Data Imputation*

During data exploration, it was found that there are two types of missing data: high school data with enumerated data types and student GPA data with float data types. The missing data were then separated to determine the most appropriate and suitable imputation method. For high school data, the imputation method used is entering a constant value, which is the mode value of the missing data. For GPA data, three imputation tests for missing data were conducted using three different algorithms: linear regression, random forest, and KNN using the Multivariate Imputation by Chained Equations (MICE) method.

TABLE I. METRICS EVALUATION

Algorithm	**MAE**	**RMSE**	**MAPE**
Linear Regression	**0.024**	**0.092**	**0.82 %**
Random Forest	0.026	0.101	0.89%
KNN	0.033	0.131	1.15 %

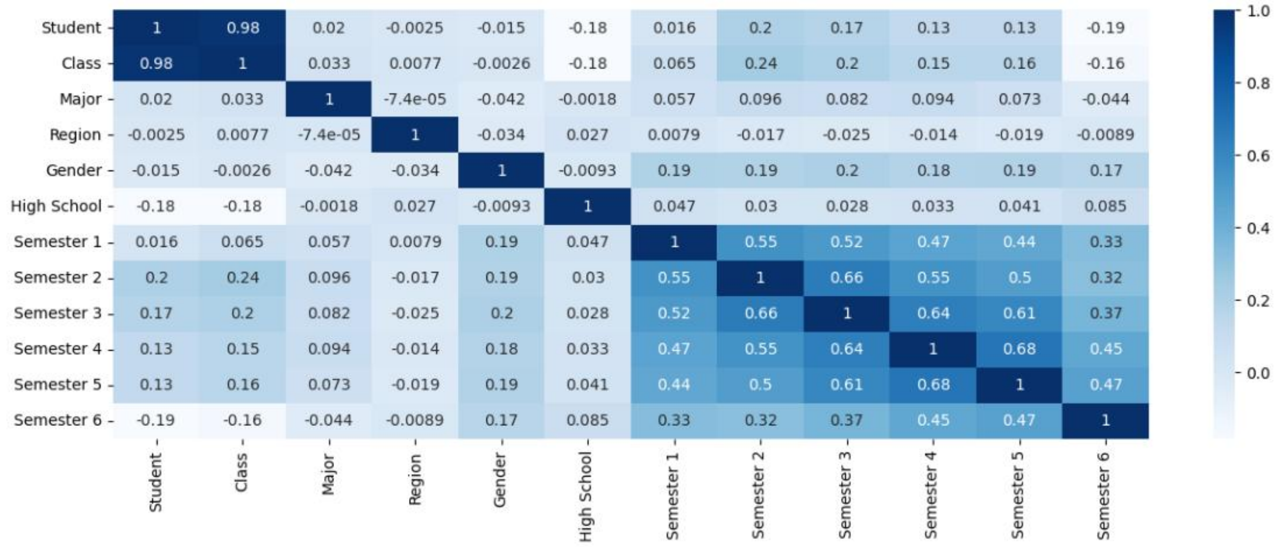
The first step taken was to examine the correlation or linear relationship between the variables. This is necessary to identify dependencies and provide more accurate prediction indications. This correlation measures the degree of association between two variables.

Fig. 3 shows that the strongest correlation between variables is the GPA from one semester to the next. The next step is imputing using three different algorithms: linear regression, random forest, and KNN. After imputation using these three algorithms, accuracy comparisons were made to evaluate the performance results of the algorithms and the new data generated. The performance evaluation results of the three algorithms used are shown in Table I.

Based on Table I, the algorithm with the smallest MAE value is linear regression, with an MAE value of 0.024. Based on Table I, the algorithm with the smallest RMSE value is linear regression, with an RMSE value of 0.092. Based on Table I, the algorithm with the smallest MAPE value is linear regression, with a MAPE value of 0.82%. Therefore, the best performance evaluation among the three algorithms is linear regression; thus, the author applies the linear regression algorithm to all missing data on the Semester GPA.

C. *Outlier Data*

In this study, the winsorizing method was used by determining its lower and upper bounds. For the semester GPA, the lower bound is 0.1, and the upper bound is 4.00. To identify outliers, the author employs a visual method. The first visualization is a boxplot of the semester GPA dataset, as shown in Fig. 4. The distribution pattern of the data in the boxplot indicates the presence of points outside the interquartile range. This suggests the presence of



	Student	Class	Major	Region	Gender	High School	Semester 1	Semester 2	Semester 3	Semester 4	Semester 5	Semester 6
Student	1	0.98	0.02	-0.0025	-0.015	-0.18	0.016	0.2	0.17	0.13	0.13	-0.19
Class	0.98	1	0.033	0.0077	-0.0026	-0.18	0.065	0.24	0.2	0.15	0.16	-0.16
Major	0.02	0.033	1	-7.4e-05	-0.042	-0.0018	0.057	0.096	0.082	0.094	0.073	-0.044
Region	-0.0025	0.0077	-7.4e-05	1	-0.034	0.027	0.0079	-0.017	-0.025	-0.014	-0.019	-0.0089
Gender	-0.015	-0.0026	-0.042	-0.034	1	-0.0093	0.19	0.19	0.2	0.18	0.19	0.17
High School	-0.18	-0.18	-0.0018	0.027	-0.0093	1	0.047	0.03	0.028	0.033	0.041	0.085
Semester 1	0.016	0.065	0.057	0.0079	0.19	0.047	1	0.55	0.52	0.47	0.44	0.33
Semester 2	0.2	0.24	0.096	-0.017	0.19	0.03	0.55	1	0.66	0.55	0.5	0.32
Semester 3	0.17	0.2	0.082	-0.025	0.2	0.028	0.52	0.66	1	0.64	0.61	0.37
Semester 4	0.13	0.15	0.094	-0.014	0.18	0.033	0.47	0.55	0.64	1	0.68	0.45
Semester 5	0.13	0.16	0.073	-0.019	0.19	0.041	0.44	0.5	0.61	0.68	1	0.47
Semester 6	-0.19	-0.16	-0.044	-0.0089	0.17	0.085	0.33	0.32	0.37	0.45	0.47	1

Fig. 3. Correlation table

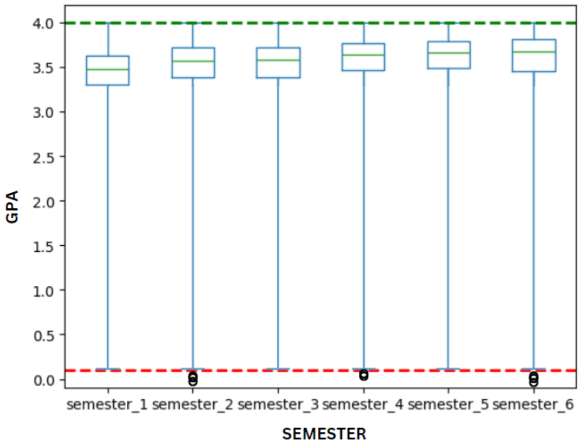


Fig. 4. Boxplot before winsorizing

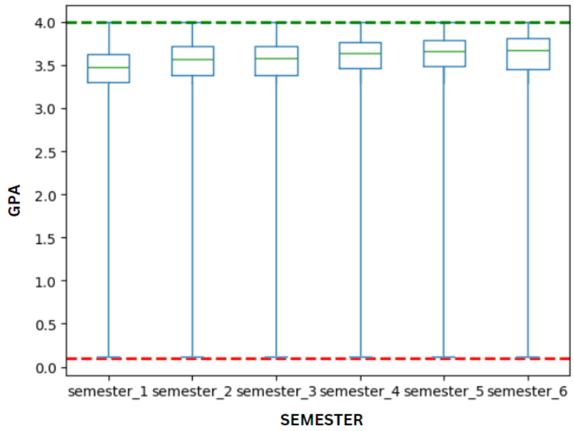


Fig. 5. Boxplot after winsorizing

outliers in the data. The green dashed line indicates the upper bound with a value of 4.00, and the red dashed line indicates the lower bound with a value of 0.1. Since data points are still found outside the red and green lines, further handling is required for the semester GPA data using winsorizing.

Fig. 5 is a boxplot graph of the dataset after winsorizing. There is a noticeable difference in the lower points, which no longer exceed the lower bound. This indicates that there are no more outliers in the data. This makes the data range narrower. Besides the graph, the reliability of data winsorizing can also be seen from its standard deviation. The comparison of standard deviations before and after data preprocessing is shown in Table II.

From Table II, it can be observed that the standard deviation values in the data before and after preprocessing differ. Before preprocessing, the standard deviation value is greater than after preprocessing. This means that the deviation of the data from the mean tends to decrease. A smaller standard deviation also indicates that the data is more consistent and stable. This means that the outlier handling process using winsorizing makes the data more representative, allowing algorithms to work with more stable and less variable data, thereby preventing distortion in the modeling analysis results.

TABLE II METRICS EVALUATION

Data	MAE	RMSE	MSE
Before preprocessing	0.171	0.327	0.107
After imputation	0.138	0.234	0.054
After winsorizing	0.134	0.217	0.047
Different	0.037	0.11	0.006
Percentage decrease	21.63%	33.64%	56.07%

D. *Evaluation*

In this study, three performance evaluation metrics were used for comparison: mean absolute error (MAE), root mean squared error (RMSE), and mean squared error (MSE). Table III shows the experimental results. In the data preprocessing process, it is evident that imputation is the most effective step. This is evident from the significant decrease in the values of MAE, MSE, and RMSE after imputation. The reduction in these values indicates that the model becomes more accurate and its overall performance improves. In comparison, other preprocessing processes, such as winsorizing, did not result in a minor decrease in the values of MAE, MSE, and RMSE. Therefore, imputation is the most effective method in improving data quality and model accuracy in this analysis.

TABLE III. DEVIATION STANDARD

Semester	1	2	3	4	5	6
Before pre-processing	0.31	0.33	0.34	0.33	0.34	0.40
After pre-processing	0.28	0.30	0.31	0.30	0.30	0.37
Different	0.03	0.03	0.03	0.03	0.04	0.03

IV. CONCLUSION

Data preprocessing is an essential step before conducting data analysis or applying machine learning. This process involves several stages to ensure that the data is clean and ready for use, including data collection, data imputation, and data winsorizing. Missing values in a dataset can hinder algorithms from understanding data patterns. Data imputation can ensure the data is complete and ready for use by the algorithms. Outliers can significantly affect the algorithm's results. Handling outliers makes the data more representative, allowing the algorithm to work with more stable and less varied data. This can also prevent distortion in the modeling analysis results. Based on testing results through metric evaluations, these preprocessing techniques can increase the model's accuracy by 0.037 for MAE, 0.11 for RMSE, and 0.006 for MSE. The processed data allows the model to function more optimally and produce more reliable results.

The case study of academic data presented in this research is limited to forecasting GPA. It is recommended for future research to investigate other academic data case studies. Future research should explore various academic datasets, such as enrollment patterns, student dropout rates, or student performance by faculty. Expanding the scope of this research will contribute to a more comprehensive understanding of the effectiveness and application of forecasting techniques in the academic environment.

Acknowledgment

This research is supported by funding from the research and community service institution of the applied research scheme of Sebelas Maret University, with contract number 228/UN27.22/PT.01.03/2023 "Development of academic and student data analytics applications".

References

[1] M. B. Musthafa, N. Ngatmari, C. Rahmad, R. A. Asmara, and F. Rahutomo, “Evaluation of university accreditation prediction system,” IOP Conf Ser Mater Sci Eng, vol. 732, no. 1, p. 12041, Jan. 2020, doi: 10.1088/1757-899x/732/1/012041.

[2] R. Sharda, D. Delen, and E. Turban, Business Intelligence: A Managerial Perspective on Analytics (3rd Edition), 3rd ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2013.

[3] V. L. Sauter, Decision Support Systems for Business Intelligence, 2nd ed. John Wiley & Sons, 2014.

[4] E. Turban, J. E. Aronson, T.-P. Liang, and R. Sharda, Decision Support and Business Intelligence Systems (8th Edition). Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2006.

[5] M. Yağcı, “Educational data mining: prediction of students’ academic performance using machine learning algorithms,” Smart Learning Environments, vol. 9, no. 1, p. 11, 2022, doi: 10.1186/s40561-022-00192-z.

[6] J. Xie, L. Sun, and Y. F. Zhao, “On the Data Quality and Imbalance in Machine Learning-based Design and Manufacturing—A Systematic Review,” Engineering, vol. 45, pp. 105–131, 2025, doi: 10.1016/j.eng.2024.04.024.

[7] J. Li, H. Fu, K. Hu, and W. Chen, “Data Preprocessing and Machine Learning Modeling for Rockburst Assessment,” Sustainability, vol. 15, no. 18, 2023, doi: 10.3390/su151813282.

[8] E. Alpaydın, Introduction to Machine Learning, Fourth edition. The MIT Press, 2020.

[9] J. Han, Data Mining: Concepts and Techniques. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005.

[10] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, “Multiple imputation by chained equations: what is it and how does it work?” Int J Methods Psychiatr Res, vol. 20, no. 1, pp. 40–49, 2011, doi: 10.1002/mpr.329.

[11] A. Z. Alruhaymi and C. J. Kim, “Why Can Multiple Imputations and How (MICE) Algorithm Work?” Open J Stat, vol. 11, no. 05, pp. 759–777, 2021, doi: 10.4236/ojs.2021.115045.

[12] R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice. OTexts, 2018.

[13] F. Petropoulos et al., “Forecasting: theory and practice,” Int J Forecast, vol. 38, no. 3, pp. 705–871, 2022, doi: 10.1016/j.ijforecast.2021.11.001.

[14] F. Rahutomo, M. M. Huda, R. A. Asmara, A. Setiawan, and A. A. Septarina, “The experiment of text–number combination forecasting,” J Phys Conf Ser, vol. 1402, no. 6, p. 66037, Dec. 2019, doi: 10.1088/1742-6596/1402/6/066037.