A Comparative Study of K-Means and KNN Imputation for Handling Missing Data in Scholarship Applicant Datasets

Muhammad[1*], Tole Sutikno[2], Imam Riadi[3]

[1]*Department of Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia*
[1]*Department of Information System, STMIK PPKIA Tarakanita Rahmawati, Tarakan, Indonesia*
[2]*Department of Electrical Engineering, Universitas Ahmad Dahlan, Yogyakarta, Indonesia*
[3]*Department of Information System, Universitas Ahmad Dahlan, Yogyakarta, Indonesia*

*corr-author: 2437083001@webmail.uad.ac.id

Abstract - Handling missing values is a key issue in data processing, especially in financial records of prospective scholarship recipients where precision is vital for effective decision making. This research aims to analyze the effectiveness of two commonly used imputation methods, namely K-Nearest Neighbors (KNN) and K-Means, in filling missing values across key attributes such as Semester, Grade Point Average (GPA), number of dependents, number of credits, and parental income. Performance evaluation was conducted using Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). The results indicate that KNN generally provides more stable and accurate imputations, particularly for attributes with homogeneous distributions such as Semester and GPA, while K-Means demonstrates competitive performance on attributes with higher variability, provided that the number of clusters is optimally defined. Nonetheless, K-Means tends to be more sensitive to increasing proportions of missing data. These findings underscore the importance of selecting imputation methods that align with attribute distribution characteristics and the extent of missing data in order to develop reliable predictive models, as observed in scenarios with 15% and 25% missing data. The findings can also serve as a reference for developing more accurate scholarship selection processes in the presence of incomplete financial data.

Keywords: imputation, K-Means, KNN, missing values

I. INTRODUCTION

Missing values are a common challenge in data analysis across various domains, including scholarship recipient selection data. These missing values may arise due to input errors or omissions during data entry. If left unaddressed, they can compromise the quality of analysis and hinder optimal decision-making processes [1-3]. Thus, appropriate handling methods are essential to ensure accurate and reliable analytical outcomes. Handling missing values is particularly critical in decision support systems that leverage historical data for prediction or recommendation. Imputation techniques offer a solution by estimating missing values based on underlying data patterns [4-6]. This approach not only completes the dataset but also preserves the validity of the resulting information. However, choosing the proper imputation method is crucial to avoid introducing further inaccuracies.

Two machine learning techniques can be used to handle missing values: supervised and unsupervised learning [7-9]. Supervised learning relies on output labels to predict missing values through model-based approaches, such as K-Nearest Neighbors (K-NN), which imputes missing values using proximity between data points. On the other hand, unsupervised learning does not require labels and focuses more on uncovering hidden patterns within the data to group similar information before estimating the missing values. K-Means is one commonly used method in this category, where data is clustered based on similarity, and missing values are calculated based on the characteristics of the cluster [10-12].

In this research, K-Means and K-NN are chosen as the primary methods for handling missing values because they represent two different approaches in machine learning. As an unsupervised learning method, K-Means enables the estimation of missing values by utilizing patterns formed in the data without requiring target labels [13-14]. Conversely, K-NN, as a supervised learning method, performs imputation based on proximity between data points, using information from available samples to estimate the missing values [15-16].

The selection of these two methods is driven by the nature of the scholarship recipient dataset, which often lacks complete labeling, necessitating a flexible

approach. In addition to representing different learning paradigms, K-Means and KNN were selected for their ability to handle complex patterns without strong distributional assumptions. Compared to basic methods like mean or median imputation, they offer more adaptive estimations and are suitable for diverse data types in education-related datasets. By comparing K-Means and K-NN, this research aims to evaluate the effectiveness of supervised and unsupervised learning in managing missing data. Prior studies on missing data imputation have focused mainly on conventional statistical or standalone machine learning techniques, with limited comparative analysis between these two paradigms.

Several studies have been conducted related to the imputation process for missing values. Research by [17] used the Mean Imputation method to maintain classification performance on small datasets with missing values, such as Hepatitis and Chronic Kidney Disease data, without reducing the amount of data analyzed. Another study conducted by [18] applied the K-NN imputation method to handle missing values in user satisfaction data of university graduates. The study by [19] implemented K-Nearest Neighbors (KNN) imputation to address missing values in rain duration prediction data from BMKG. Research by [20] proposed the use of the K-NN Imputation method to handle missing values in corn production data. This method was applied to maintain data quality and support the classification process. Study by [10] proposed the use of K-Means for the imputation process on scholarship recipient data and demonstrated that K-Means can be applied to impute missing data for prospective scholarship recipients. However, most of the previous works focused on evaluating a single imputation method in isolation. To date, limited research has directly compared the performance of K-Means and KNN side by side, especially within the specific domain of financial data for scholarship selection. This research addresses that gap by offering a comparative perspective.

Studies comparing K-Means and K-Nearest Neighbors (K-NN) in the context of financial data or scholarship recipient selection remain limited. Consequently, no consensus exists regarding which method is more effective under specific conditions. In this research, both methods will be evaluated using Mean Absolute Percentage Error (MAPE) to assess the accuracy of the estimated missing values. MAPE is selected due to its intuitive interpretation, as it expresses prediction error as a percentage of the actual value, facilitating a straightforward comparison between methods [21-22]. Additionally, Root Mean Squared Error (RMSE) will be employed as a complementary metric to measure the extent to which the estimates generated by each method deviate from the original values [21]. By incorporating these two evaluation metrics, this research aims to provide a more comprehensive understanding of the effectiveness of K-Means and K-NN in addressing missing data within scholarship recipient datasets. The findings are expected to support identifying the more optimal method based on the specific characteristics of the available data.

II. METHOD

This research adopts a comparative experimental approach to evaluate the effectiveness of two imputation techniques in handling missing values. The research uses a dataset of prospective scholarship recipients, obtained from STMIK PPKIA Tarakanita Rahmawati. The dataset was intentionally modified to simulate missing values at two proportions: 15% and 25%. It consists of 120 entries and includes five key attributes: semester, GPA, number of dependents, parental income, and the number of completed credit hours (SKS). Although relatively limited in size, the dataset reflects the common structure and characteristics of applicant data typically used in scholarship selection processes, including both academic and financial dimensions.

Fig. 1 presents the research framework, which consists of five main stages. The first stage is data collection, where raw data were retrieved from the scholarship information system and cleaned through preprocessing. In the second stage, missing values were simulated using random masking, creating two data loss scenarios (15% and 25%). The third stage involves imputation, where K-Means and K-Nearest Neighbors (KNN) were applied to estimate the missing values. K-Means, as an unsupervised method, clusters data based on shared characteristics. In contrast, KNN, a supervised method, imputes values using the nearest valid instances in the dataset. The imputation process was conducted using the Python programming language in the Google Colaboratory (Colab) environment, which offers cloud-based computational resources.

A. K-Means Imputation

The K-Means Imputation method operates by clustering data into several groups based on the similarity of attribute values. Once the data is grouped into clusters, the missing values are filled using the average values from the cluster to which the data point belongs. This approach is particularly suitable for unlabeled data that exhibit patterns of similarity among attributes.

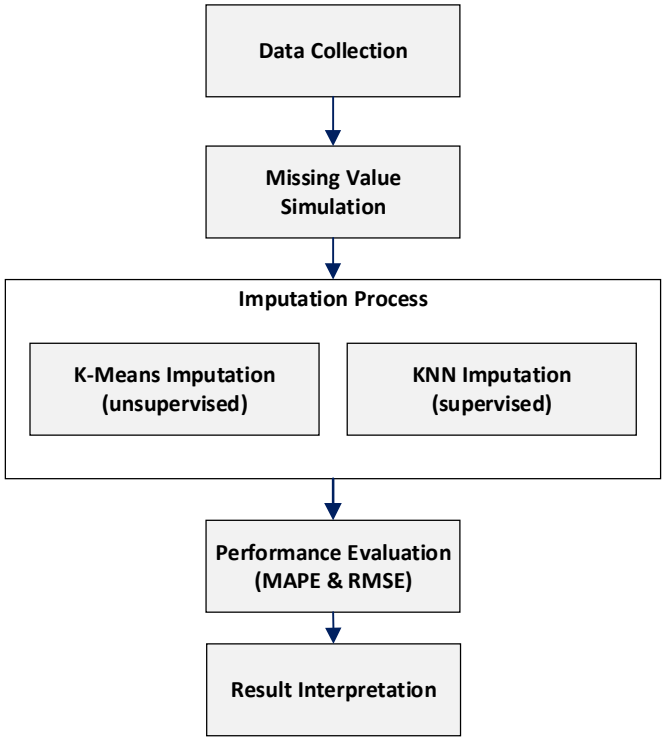

Fig. 1 Research flow

The imputation steps using K-Means in this research are based on procedures implemented in previous research. This method was selected again due to its proven stability and accuracy when applied to similar data types. This research will compare the K-Means approach to the K-Nearest Neighbors (K-NN) method to determine which technique is more effective in addressing missing data in scholarship applicant datasets. The imputation process using K-Means involves the following step [10]:

1) Initial Filling of Missing Values: The first step involves temporarily filling in the missing values in the dataset. This step is essential, as the K-Means algorithm cannot be executed if missing values are present.

2) Data Normalization: After initial filling, the data is normalized using the Min-Max Scaling method. This normalization aims to standardize the value ranges across attributes, ensuring that no single attribute dominates the clustering process.

3) Application of the K-Means Algorithm: The K-Means algorithm is applied to the normalized data. This process produces clusters (groups of data) and centroids (average values for each cluster).

4) Imputation Based on Centroids: The missing values are then filled using the centroid values of the corresponding attributes, depending on the cluster to which the data belongs. In other words, each missing value is replaced with a representative value from the cluster that exhibits similar data patterns.

Specifically, K-Means is used to group scholarship applicants based on similarities in characteristics such as GPA, parental income, number of dependents, and other attributes. Consequently, the missing values are imputed using general patterns identified within groups of similar records. This approach allows the imputation process to reflect the typical characteristics of similar data groups, thereby improving the accuracy of the imputation results.

B. K-NN Imputation

K-NN imputation operates by identifying the records that are most similar to those with missing values. The similarity is measured based on the distance between data points, which is calculated using Euclidean Distance in this research. Once the nearest neighbors are identified, the missing value is imputed using the values from these neighbors. The K-NN imputation process involves the following steps [5], [18]:

1) Identification of Missing Values: Determine which records contain missing values that require imputation.

2) Data Partitioning: The dataset is divided into two subsets: data testing and data training

3) Distance Calculation: Calculate the distance between the test data and all training data points using Euclidean Distance, as described in (1).

$$d_E\,(x,y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2} \qquad (1)$$

where d_E is the Euclidean distance computed between the testing data x and the training data y; I denotes the index of the attribute, and d represents the total number of attributes considered

4) Nearest Neighbor Selection: The training data are sorted based on the smallest distances, and the k closest data points are selected.

5) Missing Value Imputation: The missing value in the test data is imputed by calculating the average of the corresponding attribute values from the k nearest neighbors.

In this research, the k-NN method is particularly appropriate, as it enables the estimation of missing values in the scholarship applicant dataset based on similarity with other complete records. For instance, if a student has a missing GPA or several dependents, the k-NN algorithm can impute these values using data from other students with similar characteristics.

Performance evaluation was conducted using two primary metrics: Mean Absolute Percentage Error

(MAPE) and Root Mean Square Error (RMSE) to assess the effectiveness of the imputation methods. These metrics were used to measure the accuracy of the imputation results by comparing them to the original values in the dataset. Using both metrics enables a more comprehensive assessment of the performance of the applied imputation methods. The first formula represents the MAPE as described in (2), where *n* denotes the total number of observations, *xt* represents the actual value and *ft* represents the forecasted value. The formula calculates the average of the absolute percentage errors between the actual and forecasted values, and the result is expressed as a percentage.

$$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{x_t - f_t}{x_t}\right| \text{ x } 100\% \quad (2)$$

The second formula represents the RMSE as described in (3), where *n* again refers to the total number of observations, *At* denotes the actual value, and *Ft* denotes the forecasted value. The formula calculates the square root of the average squared differences between the actual and forecasted values.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(A_t - F_t)^2}{n}} \quad (3)$$

The final stage of the research process involves analyzing and interpreting the evaluation results. At this stage, the performance of each imputation technique is systematically compared to determine the most reliable method for handling missing data. The outcomes of this analysis are expected to offer valuable insights into each approach's strengths and limitations and serve as a basis for selecting the most appropriate imputation technique for similar data contexts in future applications.

III. RESULT AND DISCUSSION

This section contains a structured explanation of the research results, which includes the preparation of the missing values simulation, the imputation process using K-Means and K-NN, the evaluation of the performance of the algorithms used, and the analysis of the obtained research results.

A. Dataset and Missing Values

In this research, the proportion of missing data is set at 15% and 25%, as shown in Table I. This proportion was chosen to avoid directly replicating previous studies that used 10% and 20% proportions. Furthermore, the variation in the levels of missing data is expected to broaden the scope of the evaluation of the K-Means and K-NN imputation methods, thereby providing a more comprehensive overview of the effectiveness of each approach in the context of financial data for scholarship applicants.

To systematically simulate missing data conditions in this research, a random masking approach was applied to the dataset that had been previously prepared. Values in the complete dataset were randomly removed according to the established missing data proportions of 15% and 25%. This process was automatically applied to a set of relevant attributes without altering the original distribution characteristics of the data. This approach is widely used in imputation studies because it can simulate random missing data conditions, allowing the evaluation of imputation method performance in a real-world-like scenario while still being experimentally controlled. Using this approach, the study can objectively and measurably compare the effectiveness of the K-Means and K-NN methods in handling incomplete data within the context of financial datasets for scholarship applicants.

The imputation process on the dataset was performed using the Python programming language, leveraging libraries that are highly useful for data manipulation and the application of machine learning algorithms. In this research, the pandas library was used for manipulating and processing the dataset, and scikit-learn, a machine learning library that includes the K-Means and K-NN, was utilized for imputing missing data.

B. K-Means Imputation

The imputation process for missing values using the K-Means method was done in Python, utilizing the K-Means algorithm from the scikit-learn library and a clustering-based imputation technique. This approach groups the available data (without missing values) into clusters based on attribute similarity. Once the clusters are formed, missing values in each attribute are estimated using the centroid value of the relevant cluster. The centroid is calculated as the mean of the existing attribute values within each cluster. In this research, three variations of the number of clusters were used, namely k = 3, k = 4, and k = 5, along with two missing value proportion scenarios, specifically 15% and 25%. The imputation results were evaluated using RMSE and MAPE metrics. These metrics were used to compare the imputed values with the original data before masking to assess the quality of the missing value estimations. The evaluation results for the application of K-Means imputation at a 15% missing value are presented in Table II.

The evaluation results presented in Table II indicate that the imputation performance remains relatively stable for several attributes, even as the number of clusters

increases. The attributes SEMESTER and GPA, which tend to have a homogeneous value distribution, show RMSE and MAPE values that do not change significantly when the number of clusters is increased from 3 to 5. For the DEPENDENTS attribute, there is a slight increase in RMSE and MAPE as *k* increases, suggesting that too many clusters may lead to overfitting in attributes with moderate variation. Meanwhile, the INCOME attribute shows consistent results across all *k* values, indicating that the clustering structure does not adequately capture the wide variation in income data. For the CREDITS attribute, the best results were achieved at $k = 3$, with lower RMSE compared to $k = 4$ and 5, suggesting that there is an optimal number of clusters depending on the data distribution of each attribute.

In the scenario with a 25% missing value proportion, as shown in Table III, a degradation in imputation performance is observed compared to the 15% proportion, indicated by increased RMSE and MAPE values across almost all attributes. The SEMESTER and GPA attributes continue to exhibit a stable pattern, although the error rises along with the higher proportion of missing data. However, an anomaly is observed in the DEPENDENTS attribute, where the MAPE value spikes drastically to 1.25555E+13, indicating a severe issue in the imputation process—most likely due to the formation of non-representative clusters caused by the high volume of missing data. For the INCOME attribute, the imputation performance improves, with a lower RMSE observed at $k = 5$, suggesting that more clusters can better capture the high variability in the data. The CREDITS attribute shows RMSE and MAPE values fluctuations, with the best performance at $k = 3$. This indicates that sensitivity to the number of clusters increases as the proportion of missing values increases.

In general, the number of clusters used in the K-Means imputation process contextually affects the quality of the estimated results. Attributes with more stable and homogeneous data distributions, such as SMT and GPA, tend to be less significantly affected by changes in the number of clusters. In contrast, attributes with high data variability, such as DEPENDENT and INCOME, show improved results with increased clusters. However, there is a risk of significant errors, especially when the proportion of missing data is high. Therefore, selecting the optimal number of clusters highly depends on the characteristics of each attribute and the extent of missing data.

Based on the evaluation, it can be concluded that the K-Means method is more suitable for attributes with stable and homogeneous data distributions, such as SMT and GPA, which tend to exhibit relatively consistent imputation performance even when the number of clusters is increased. However, K-Means shows limitations for attributes with high and more complex data variability, such as DEPENDENTS and INCOME, especially when the proportion of missing values is high. This method tends to struggle to capture significant data variability, resulting in extreme imputation errors, as observed in the DEPENDENTS attribute. Therefore, K-Means is more effective for attributes with relatively consistent data distributions, while alternative methods may be more appropriate for accurately handling missing values for more variable attributes.

C. KNN Imputation

The imputation of missing values is carried out using the Python programming language by utilizing the KNNImputer function from the scikit-learn library. In addition, several other supporting libraries, such as pandas for data processing and numpy for numerical computation, are also used. Missing value imputation on the dataset is performed using the KNN algorithm with parameters $k = 3$ and $k = 5$. This not only allows for the imputation process but also helps to evaluate the impact of the number of neighbors on the quality of the missing value estimation. This process is applied to five numerical attributes, namely SEMESTER, GPA, DEPENDENTS, INCOME, and CREDITS. Two missing value proportion scenarios are simulated, namely, 15% and 25%, to observe the model's sensitivity to the level of data loss.

The selection of the k value in the KNN algorithm is based on standard practices in the literature, where small values such as $k = 3$ and $k = 5$ are often used because they are considered to balance bias and variance. The imputation process is performed by calculating the average of the attribute values of the k nearest neighbors based on Euclidean distance. The imputed results are then compared to the original data before masking, using two main evaluation metrics, namely RMSE and MAPE. The following Table IV presents the performance evaluation results of KNN for each combination of *k* values and missing data proportions.

The evaluation results of the KNN imputation presented in Table III show that the selection of the k value has a different impact on each attribute. For the SEMESTER and DEPENDENTS attributes, which have a more homogeneous data distribution, using $k = 3$ yields more optimal results, with lower RMSE and minimal MAPE values. This indicates that for attributes with relatively stable distributions, choosing a smaller *k* can improve the precision of the estimation. However, for attributes such as INCOME and GPA, which have more

varied data distributions, using k = 5 proves to be more effective. The Income attribute, which is heavily influenced by many external factors such as employment and location, shows more stable results with lower RMSE and MAPE when k = 5 compared to k = 3. This is due to the more diverse and complex nature of the data, which is better handled with more neighbors that can capture hidden patterns in the data.

Furthermore, although the difference in MAPE values between k = 3 and k = 5 is relatively minor for most attributes, a more significant difference is observed for INCOME and DEPENDANTS. The difference in the RMSE scale also provides a clearer picture of the suitability of the k value, with k = 3 being more effective for attributes with a more homogeneous distribution and k = 5 performing better for attributes with more significant data variations. It should be noted that Income has a much larger numerical scale compared to other attributes, so the absolute error produced is also higher. However, the MAPE value remains within an acceptable level. However, not all attributes can be well imputed using KNN. Some attributes may have data characteristics that are not suitable for KNN, such as highly variable distributions or significant outliers. Attributes like Income, which exhibit huge variations between individuals, demonstrate that using k = 5 is more effective in capturing hidden data patterns. Nevertheless, the RMSE and MAPE results for Income are still higher than those for other attributes, indicating that KNN may not be the best method for attributes with large data distributions.

On the other hand, attributes with more regular data distributions, such as GPA and DEPENDANTS, are better suited for imputation using k = 3, which results in lower RMSE values and very small MAPE. This indicates that KNN Imputation can provide exact results for attributes with more homogeneous distributions, where a smaller number of neighbors can yield more accurate estimates. Overall, KNN Imputation shows that the selection of the k value is highly dependent on the characteristics of each attribute. For attributes with more stable and homogeneous data distributions, k = 3 provides more optimal results. However, for attributes with more significant and more complex data variation, such as Income, k = 5 can provide more consistent and accurate results. Therefore, KNN Imputation should be applied with consideration of the nature of each attribute.

Based on the evaluation results using RMSE and MAPE metrics for the two imputation methods KNN and K-Means, several key findings emerge regarding the effectiveness of each approach in handling missing data among prospective scholarship recipients. KNN demonstrates more consistent and accurate performance than K-Means, particularly for attributes with homogeneous distributions such as Semester and GPA. This can be attributed to KNN's underlying principle of leveraging distance-based similarity, allowing for more contextual estimations based on surrounding similar instances. However, K-Means does not perform universally worse than KNN in this research. In certain conditions, K-Means yields better imputation results. For instance, in the case of the 'Income' attribute with 25% missing values, K-Means shows a significant reduction in RMSE when the number of clusters is increased to k = 5, indicating its capability to capture the complex variability of income data more effectively. Similarly, for the 'Credits' attribute, K-Means achieves optimal performance at k = 4 and k = 5, suggesting that it can work effectively on attributes with moderate variability, provided that the number of clusters is chosen appropriately.

TABLE I
DATASET

Dataset	Total Data	Missing Data (%)
1	120	18 (15%)
2	120	30 (25%)

TABLE II
EVALUATION OF K-MEANS IMPUTATION ON 15% MISSING VALUE

Atribut	RMSE			MAPE		
	K=3	K=4	K=5	K=3	K=4	K=5
Semester	0.000	0.000	0.000	0.00%	0.00%	0.00%
GPA	0.025	0.024	0.021	0.07%	0.06%	0.05%
Dependents	0.230	0.240	0.276	0.65%	0.80%	0.90%
Income	393.594	393.594	393.594	0.0416%	0.0416%	0.0416%
Credits	3.687	4.321	5.043	1.46%	1.50%	1.76%

TABLE III
EVALUATION OF K-MEANS IMPUTATION ON 25% MISSING VALUE

Atribut	RMSE			MAPE		
	K=3	K=4	K=5	K=3	K=4	K=5
Semester	0.045	0.048	0.052	0.12%	0.14%	0.15%
GPA	0.062	0.063	0.060	0.15%	0.17%	0.16%
Dependents	1.523	1.733	1.948	1.25555E+13	1.25555E+13	1.25555E+13
Income	415.674	387.134	376.118	0.0427%	0.0392%	0.0383%
Credits	6.023	6.437	6.501	1.70%	1.80%	1.85%

TABLE IV
EVALUATION OF KNN IMPUTATION

Atribut	RMSE (15%)		RMSE (25%)		MAPE (15%)		MAPE (25%)	
	K=3	K=5	K=3	K=5	K=3	K=5	K=3	K=5
Semester	0.000	0.103	0.183	0.245	0.00%	0.28%	0.83%	1.17%
GPA	0.025	0.021	0.048	0.064	0.07%	0.05%	0.35%	0.46%
Dependents	0.230	0.276	0.456	0.540	0.65%	0.50%	1.13%	1.20%
Income	376520.507	266268.937	88498.596	54212.331	3.61%	2.71%	0.78%	0.50%
Credits	3.687	5.043	6.265	4.199	1.46%	1.76%	2.52%	1.63%

This research also confirms that the performance of both methods deteriorates as the proportion of missing values increases from 15% to 25%, which is reflected in the general rise in RMSE and MAPE values. Fig. 2 illustrates a performance comparison between the two methods under the 15% missing value scenario, which highlights the stability of KNN across various attributes. At a missing value rate of 15% as shown in Fig. 2, the graph indicates that the KNN method tends to produce more stable and lower RMSE and MAPE values compared to K-Means, especially for attributes such as GPA and Semester. Although for the Income attribute, both methods yield relatively high RMSE and MAPE values, the difference between the methods is not very pronounced. This suggests that under moderate missing value conditions, KNN has the advantage of consistency and stability in estimates for attributes with low to medium variation.

However, KNN still shows a more controlled performance decline, while K-Means becomes more sensitive to the proportion of missing values, as seen with the DEPENDENTS attribute, which experiences a very extreme MAPE spike. This is clearly depicted in Fig. 3, where the MAPE for K-Means on the DEPENDENTS attribute spikes drastically, becoming an outlier in the overall performance comparison. In contrast, KNN maintains its stability even as the missing value proportion increases.

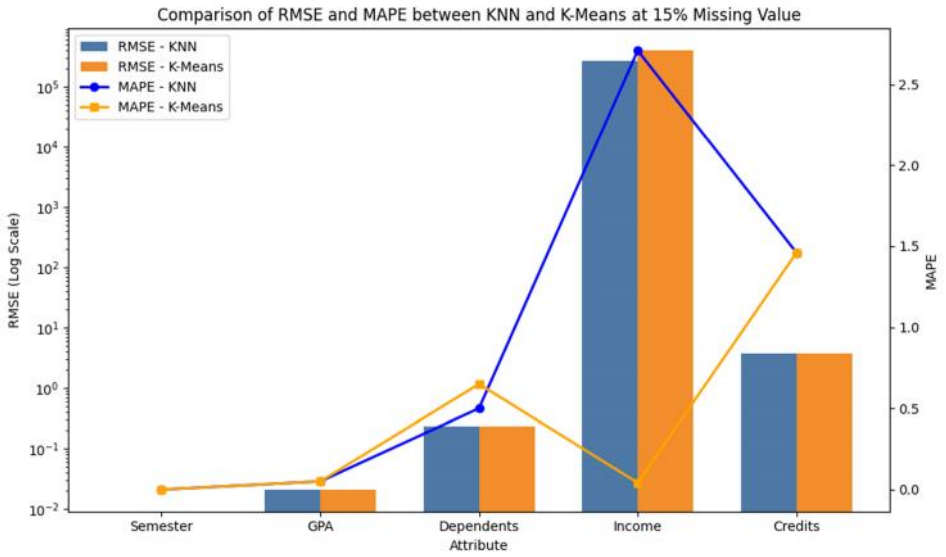

Fig. 2 Comparison of RMSE and MAPE at 15% Missing Value

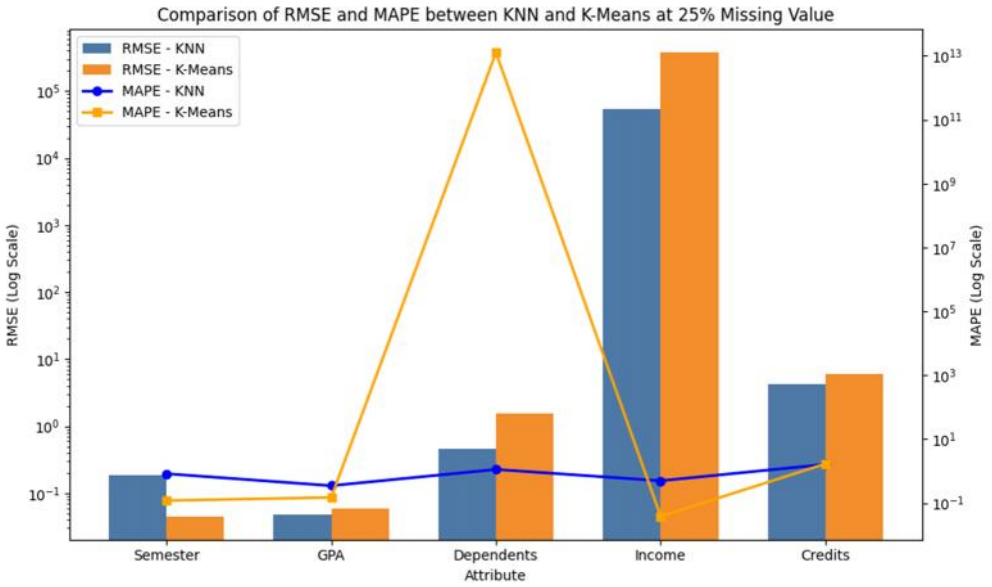

Fig. 3 Comparison of RMSE and MAPE at 25% Missing Value

In Fig. 3, when the missing value rate increases to 25%, K-Means' performance shows significant degradation in several attributes. The extreme spike in MAPE values for the Tanggungan (Dependents) attribute marks the presence of an outlier or considerable inconsistency in the imputation process using K-Means. This indicates that the K-Means method is susceptible to data imperfections, especially when the data distribution is uneven, and the amount of missing data is considerable. In contrast, KNN still shows relatively controlled performance despite a general increase in RMSE.

This suggests that the K-Means clustering process can be significantly disrupted when the data is incomplete, leading to the formation of centroids that are not representative. This weakness arises because the K-Means algorithm heavily relies on calculating each cluster's mean (centroid), which, in the case of incomplete data, becomes prone to bias. When the data used to form the centroids contains many missing values, the centroid position becomes inaccurate and no longer reflects the actual structure of the cluster. As a result, the imputation of values based on the cluster becomes irrelevant, causing overall performance to decline sharply.

On the other hand, KNN can maintain its performance because of its instance-based learning principle. This approach imputes missing values based on similarity to the nearest neighbours with complete data. Therefore, as long as enough valid comparison instances are available, KNN can continue to provide relatively stable results, even when the missing value proportion increases. Further comparisons of several aspects can be seen in Table V.

Based on Table V, the overall results indicate that the choice of imputation method cannot be standardized. However, the characteristics of the attributes and the proportion of missing values must be considered. KNN excels in stability and accuracy when the data exhibits linear and homogeneous patterns. At the same time, K-Means can be relied upon to handle attributes with high variability, provided the number of clusters is determined accurately. The findings in this research emphasize that an effective imputation strategy is highly dependent on the nature of the attributes and data conditions, making an adaptive approach key in managing missing values.

However, it is important to note several limitations. The small dataset size may limit the generalizability of the results and increase the risk of overfitting or biased estimation, particularly for attributes with skewed distributions. Additionally, the imbalance across attributes such as variations in range and scale could influence the sensitivity of each method differently. Future research is encouraged to employ larger and more diverse datasets, as well as to investigate hybrid or ensemble imputation techniques that could enhance the robustness, accuracy, and adaptability of the models in real world applications.

TABLE V
SUMMARY OF K-MEANS VS KNN IMPUTATION PERFORMANCE

Aspect	K-Means Imputation	KNN Imputation
Accuracy on homogeneous attributes	Very good, stable across all k	Good, but can improve with smaller k
Accuracy on complex attributes	Tends to be unstable, sensitive to k	More stable and accurate with larger k
Sensitivity to % of missing values	Higher sensitivity, performance degrades quickly	More resilient to increasing missing values
Occurrence of extreme anomalies	Present (e.g., Dependents)	No extreme anomalies observed
Overall suitability	Good for attributes with stable distribution	More flexible for all types of data distributions

IV. CONCLUSION

The evaluation results of the two imputation methods, namely K-Nearest Neighbors (KNN) and K-Means, indicate that KNN generally delivers more stable and accurate performance, particularly in handling attributes with homogeneous and linear distributions, such as Semester and GPA. KNN's robustness against increasing proportions of missing data is evident from the absence of extreme fluctuations in RMSE and MAPE values, even as the missing rate increases from 15% to 25%. This reflects KNN's advantage in maintaining consistent estimations based on instance similarity. Conversely, K-Means exhibits higher sensitivity to increasing missing value proportions, as illustrated in the 'Dependents' attribute, where MAPE values experienced a significant spike, emerging as a performance outlier. This limitation stems from K-Means' reliance on forming representative centroids, which can be disrupted when a substantial amount of data is missing. Nevertheless, K-Means shows competitive potential for attributes with high data variability, such as Income and Credits, particularly when the number of clusters (k) is optimally determined. Based on the comparative performance, it can be concluded that imputation approaches should be aligned with the attribute characteristics and data distribution. KNN is more favourable in contexts with stable distributions and higher missing proportions, whereas K-Means offers an efficient alternative for attributes with natural segmentation and lower missingness. However, this research is limited by the relatively small dataset size and its specificity to scholarship-related financial data, which may affect the generalizability of the findings. Future research should explore hybrid imputation models that integrate KNN's instance-based learning capabilities with the clustering efficiency of K-Means, and validate their effectiveness using larger and more diverse real-world datasets.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the Finance and Student Affairs Division of STMIK PPKIA Tarakanita Rahmawati for providing the data used in this research, and to the Doctoral Program in Informatics (S3DIFA), Universitas Ahmad Dahlan (UAD), for supporting this research.

REFERENCES

[1] M. Afkanpour, E. Hosseinzadeh, and H. Tabesh, "Identify the most appropriate imputation method for handling missing values in clinical structured datasets: a systematic review," *BMC Med Res Methodol*, vol. 24, no. 1, pp. 1–13, Dec. 2024, doi: 10.1186/s12874-024-02310-6.

[2] X. Miao, Y. Wu, L. Chen, Y. Gao, and J. Yin, "An Experimental Survey of Missing Data Imputation Algorithms," *IEEE Trans Knowl Data Eng*, vol. 35, no. 7, pp. 6630–6650, 2023, doi: 10.1109/TKDE.2022.3186498.

[3] P. Keerin and T. Boongoen, "Improved KNN imputation for missing values in gene expression data," *Computers, Materials and Continua*, vol. 70, no. 2, pp. 4009–4025, 2022, doi: 10.32604/cmc.2022.020261.

[4] K. Seu, M. S. Kang, and H. Lee, "An Intelligent Missing Data Imputation Techniques: A Review," *International Journal On Informatics Visualization*, vol. 6, no. 1, pp. 278–283, May 2022, doi: 10.30630/joiv.6.1-2.935.

[5] A. Fadlil, H. Herman, and D. M. Praseptian, "Single Imputation Using Statistics-Based and K Nearest Neighbor Methods for Numerical Datasets," *Ingenierie des Systemes d'Information*, vol. 28, no. 2, pp. 451–459, Apr. 2023, doi: 10.18280/isi.280221.

[6] S. I. Khan and A. S. M. L. Hoque, "SICE: an improved missing data imputation technique," *J Big Data*, vol. 7, no. 1, pp. 1–21, Dec. 2020, doi: 10.1186/s40537-020-00313-w.

[7] P. S. Raja and K. Thangavel, "Missing value imputation using unsupervised machine learning techniques," *Soft comput*, vol. 24, no. 6, pp. 4361–4392, Mar. 2020, doi: 10.1007/s00500-019-04199-6.

[8] N. Z. Abidin, A. R. Ismail, and N. A. Emran, "Performance Analysis of Machine Learning Algorithms for Missing Value Imputation," *IJACSA) International Journal of Advanced Computer Science*

and Applications, vol. 9, no. 6, pp. 442–447, 2018, doi: 10.14569/IJACSA.2018.090660.

[9] M. Alabadla, F. Sidi, I. Ishak, H. Ibrahim, and L. S. Affendey, "Systematic Review of Using Machine Learning in Imputing Missing Values," *IEEE Access*, vol. 10, pp. 44483–44502, Apr. 2022, doi: 10.1109/ACCESS.2022.3160841.

[10] M. Muhammad, T. Sutikno, and I. Riadi, "K-means clustering as an imputation strategy for missing values in scholarship candidate data," *Mantik Journal*, vol. 8, no. 4, pp. 2685–4236, 2025, doi: 10.35335/mantik.v8i4.5904.

[11] S. Wang, M. Li, N. Hu, E. Zhu, and J. Hu, "K-Means Clustering With Incomplete Data," *IEEE Access*, vol. 7, pp. 69162–69171, 2019, doi: 10.1109/ACCESS.2019.2910287.

[12] K. Hadi and E. Utami, "Analysis of K-NN with the Integration of Bag of Words, TF-IDF, and N-Grams for Hate Speech Classification on Twitter," *JUITA: Jurnal Informatika*, vol. 12, no. 2, pp. 289–298, 2024, doi: 10.30595/juita.v12i2.23829.

[13] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electronics (Basel)*, vol. 9, no. 8, pp. 1–12, Aug. 2020, doi: 10.3390/electronics9081295.

[14] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[15] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, IEEE, May 2019, pp. 1255–60. doi: 10.1109/ICCS45141.2019.9065747.

[16] D. M. P. Murti, A. P. Wibawa, M. I. Akbar, and U. Pujianto, "K-Nearest Neighbor (K-NN) based Missing Data Imputation," in *2019 5th International Conference on Science in Information Technology (ICSITech)*, IEEE, Oct. 2019, pp. 1–6. doi: 10.1109/ICSITech46713.2019.8987530.

[17] F. Y. Pamuji, A. R. Muslikh, R. M. Arief, and D. Muti, "Komparasi Metode Mean dan KNN Imputation Dalam Mengatasi Missing Value Pada Dataset Kecil," *JIP (Jurnal Informatika Polinema)* , vol. 10, no. 2, pp. 257–264, 2024, doi: 10.33795/jip.v10i2.5031.

[18] A. Fadlil, H. Herman, and D. M. Praseptian, "K Nearest Neighbor Imputation Performance on Missing Value Data Graduate User Satisfaction," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 4, pp. 570–576, Aug. 2022, doi: 10.29207/resti.v6i4.4173.

[19] I. D. Oktaviani and A. G. Putrada, "KNN imputation to missing values of regression-based rain duration prediction on BMKG data," *JURNAL INFOTEL*, vol. 14, no. 4, pp. 249–254, Nov. 2022, doi: 10.20895/infotel.v14i4.840.

[20] M. Lutfi and M. Hasyim, "Penanganan Data Missing Value Pada Kualitas Produksi Jagung Dengan Menggunakan Metode K-NN Imputation Pada Algoritma C4.5," *JURNAL RESISTOR*, vol. 2, no. 2, 2019, doi: 10.31598/jurnalresistor.v2i2.427.

[21] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput Sci*, vol. 7, pp. 1–24, 2021, doi: 10.7717/PEERJ-CS.623.

[22] U. Khair, H. Fahmi, S. Al Hakim, and R. Rahim, "Forecasting Error Calculation with Mean Absolute Deviation and Mean Absolute Percentage Error," in *Journal of Physics: Conference Series 930*, Institute of Physics Publishing, Dec. 2017, pp. 1–7. doi: 10.1088/1742-6596/930/1/012002.