Mantik Journal, 8 (4) (2025), ISSN 2685-4236 (Online)

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik

K-means clustering as an imputation strategy for missing values in scholarship candidate data

Muhammad[1], Tole Sutikno[2], Imam Riadi[3]
[1]Program Studi Informatika, Universitas Ahmad Dahlan, Indonesia
[2]Program Studi Teknik Elektro, Universitas Ahmad Dahlan, Indonesia
[3]Program Studi Sistem Informasi, Universitas Ahmad Dahlan, Indonesia

A R T I C L E I N F O

Article history:

Received Dec 14, 2024
Revised Dec 18, 2024
Accepted Jan 10, 2025

Keywords:

Imputation;
K-Means;
MAPE;
Missing Values;
Scholarship.

ABSTRACT

The issue of missing values in the scholarship selection process poses a challenge that can impact decision-making. This study aims to perform data imputation for scholarship candidate datasets using the K-Means method and evaluate its performance using the Mean Absolute Percentage Error (MAPE). K-Means was selected for its ability to group data based on pattern similarities, enabling it to estimate missing values in the scholarship candidate dataset. Two datasets were utilized in this study: one with 10% missing data and another with 20%. The results indicate that K-Means imputation can effectively apply to scholarship candidate data. Additionally, the findings reveal that the proportion of missing data influences the optimal number of clusters required. For the dataset with 10% missing data, the best configuration was achieved with 5 clusters, resulting in a MAPE of 13%. Conversely, for the dataset with 20% missing data, the optimal configuration required 2 clusters, yielding a MAPE of 14%.

This is an open access article under the CC BY-NC license.

Corresponding Author:

Muhammad,
Program Studi Informatika,
Universitas Ahmad Dahlan,
Jl. Ring Road Selatan, Bantul, D.I Yogyakarta, 55191, Indonesia.
Email: 2437083001@webmail.uad.ac.id

1. INTRODUCTION

In higher education, scholarship programs play a crucial role in providing opportunities for financially disadvantaged students to pursue their studies. Scholarship programs are essential in offering opportunities for students with financial limitations to continue their education (Dalla & Kewuel, 2023; Goa Wea & Adiwidjaja, 2018). A critical step in scholarship distribution is the selection process for potential recipients. This process typically involves analyzing various personal and economic data (Darlinda & Utamajaya, 2022; Ulandari, 2020). However, managing this data often faces the challenge of missing values, which can hinder the selection process.

Missing data is a common problem in data analysis, occurring when one or more values in a dataset are absent (Fadlil et al., 2023; Marcelino et al., 2022; Miao et al., 2023). The presence of missing data in the scholarship selection process poses significant challenges for administrators, as incomplete data can hinder an objective assessment of applicants' eligibility. This not only increases the time and resources needed but also risks compromising the fairness of the selection process. To ensure the selection process

is efficient and accurate, practical approaches are needed to handle missing data, ensuring that the data used for analysis is representative and genuinely reflects the actual conditions of scholarship candidates.

Missing data in the scholarship selection process can occur for various reasons, such as incomplete form submissions by applicants, system errors during data collection, or other administrative challenges. This data incompleteness significantly impacts the quality of analysis, introducing bias and reducing accuracy in evaluation processes (Bangun & Karim, 2024). When scholarship eligibility data is incomplete, the analysis may become inaccurate, leading to biased or unfair decisions. For example, missing information on family income or dependents can result in eligibility calculations that do not reflect the applicant's true circumstances, potentially causing inequitable outcomes. Thus, adopting effective methods to address missing data is crucial to ensure reliable and fair analysis results.

Simple techniques, such as mean imputation, are often used to handle missing data by replacing missing values with the variable's mean. While easy to implement, this method overlooks data variability and may reduce predictive accuracy, especially in complex datasets. Similarly, median and mode imputation are commonly applied but also fail to capture intricate data patterns, limiting their effectiveness in more complex scenarios. (Kabir et al., 2020; Lin & Tsai, 2020; Rangga Baihaqi et al., 2023; Yulian Pamuji et al., 2024). Consequently, this study proposes K-Means Clustering as an alternative method that is expected to be more effective in estimating missing data in the case examined in this research. Several studies have been conducted on missing data. For example, Baihaqi et al. implemented Mean Imputation and Single Center Imputation Chained Equation (SICE) with Linear Regression algorithms to address missing values in numerical data, enhancing the effectiveness of data mining processes (Rangga Baihaqi et al., 2023). Another study by Pamuji et al. aimed to maintain classification performance on small datasets with missing values, such as Hepatitis and Chronic Kidney datasets, by employing imputation methods like Mean Imputation to avoid reducing the dataset used in the classification process (Yulian Pamuji et al., 2024).

K-means clustering is a machine-learning method designed to group data into clusters based on similarity. This algorithm has been widely applied in various fields, ranging from customer segmentation in marketing, user behavior analysis, and clustering of educational institutions to analyzing cyberbullying patterns (Fatmawaty et al., 2024; Nasyuha et al., 2022; Praseptian M et al., 2022; Riadi & Prayudi, 2022). The K-Means algorithm has been shown to be effective for datasets with numerical characteristics and data distributions that allow for cluster formation, as shown by the study (Praseptian M et al., 2022), where K-Means was used for the Level of User Satisfaction of College Graduates. By identifying patterns within each cluster, the algorithm is not only capable of grouping data but is also expected to estimate missing values based on similar characteristics within the same cluster.

This study employs K-Means to estimate missing values in scholarship applicant datasets by utilizing the similarity between existing and missing data. After the imputation process, the accuracy is assessed using Mean Absolute Percentage Error (MAPE) (Khair et al., 2017; Liantoni & Agusti, 2020). This research focuses on developing a method for imputing missing data in scholarship applicant datasets using K-Means Clustering and evaluating the method's accuracy through MAPE. It is anticipated that this study will contribute as a reference for improving the management of scholarship selection data and as a guideline for imputation methods in other educational datasets. By addressing the missing data issue, the scholarship selection process is expected to become more transparent and provide more equitable opportunities for students in need.

2. RESEARCH METHOD

This research was conducted through several systematic stages designed to achieve the main objective, namely to overcome the problem of missing data in the selection data of

prospective scholarship recipients using the K-Means Clustering approach and measure its accuracy with MAPE. The stages in this study can be seen in Figure 1.

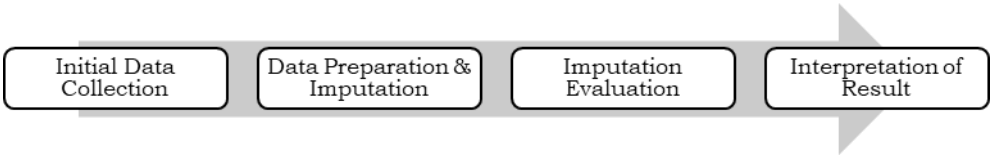


Figure 1. Research Stages

In Figure 1, the research stages begin with initial data collection. The dataset consists of 120 scholarship applicants with five attributes: semester, GPA, number of dependents, parents' income, and credits completed. The dataset attributes were chosen based on their relevance to the institution's scholarship selection process, reflecting core criteria for academic performance and financial need. These attributes have proven sufficient for decision-making in this context. Future studies may explore additional factors to enhance the selection process further. Two simulated datasets with missing data are used: Sample 1 (10% missing data, 12 values) and Sample 2 (20% missing data, 24 values), introduced randomly to reflect real-world scenarios. After identifying missing data, the next stage is data preparation. In this stage, simple imputation methods such as mean imputation are applied as a comparison, while the K-Means Clustering technique is used as the primary method to group similar data into several clusters. However, K-Means was chosen due to its compatibility with the dataset's numerical attributes and the pre-imputed data.

The next stage is the evaluation of the imputation model (imputation evaluation) using MAPE, which is a metric to assess the accuracy of the imputation results. This test is carried out by comparing the imputation results to the original data available. With MAPE, the lower the value, the more accurate the imputation or prediction method used. MAPE provides information on how far the imputation results deviate from the original data in percentage, which allows a precise evaluation of the effectiveness of the technique used. This analysis is essential to ensure that the K-Means approach can provide more representative estimates compared to conventional imputation techniques.

The final stage is interpreting the results and conclusions. Based on the testing and evaluation results, conclusions are made to answer the research questions and assess the effectiveness of the proposed approach in improving the accuracy of the scholarship candidate selection process. The applied method can be a valid solution and can be implemented in scholarship data management to create a fairer and more efficient selection process.

2.1 K-Means Clustering

K-means clustering is an unsupervised learning algorithm used to group data based on feature similarities. K-Means works by dividing data into several clusters, where each cluster has a center or average point (centroid). K-means is one of the simplest unsupervised learning algorithms and is often used to solve various data clustering problems. The general workflow of the K-Means algorithm can be seen in Figure 2. The steps of the K-Means algorithm are as follows (Hutagalung & Sonata, 2021; Privandhani & Sulastri, 2022; Rahmayani & Hidayati, 2022): (a) Determine the number of clusters, k, that you want to form and randomly select initial centroids. (b) Calculate the distance between each data point and each centroid using a distance formula, such as Euclidean distance, which can be seen in Equation (1), and assign the data points to the nearest centroid.

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

a) Update the centroid values of each cluster by calculating the mean of the data points within the cluster.
b) Repeat the distance calculation with the new centroids, as in step 2.

c) Repeat these steps until there are no changes in the cluster memberships.

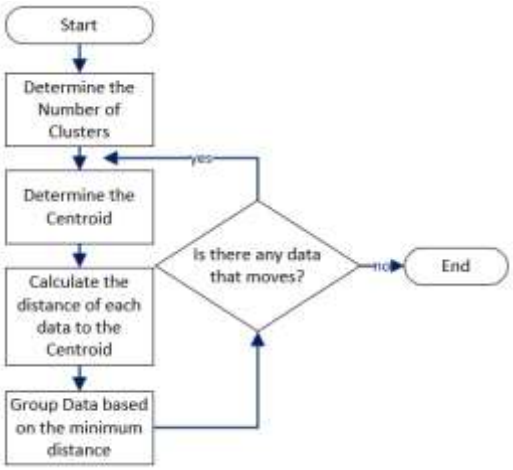


Figure 2. K-Means Algorithm (Rosmini et al., 2018; Rustam, 2018)

In this study, the K-Means algorithm was applied as an imputation method to address the issue of missing data in the scholarship applicants' dataset. This process works by grouping data with similar characteristics into clusters. Missing values are then imputed using the average (centroid) of the relevant cluster where the data belongs.

This approach is considered more accurate than simple imputation methods because it takes into account the relationships between attributes in the data. Generally, the imputation process using K-Means follows the steps of the K-Means algorithm (see Figure 2). The main difference lies in an additional final step performed when missing data exists in the dataset—filling in the missing values using the final centroid from the K-Means process in the cluster where the missing data is located (see Figure 3 for further details) (Chhabra et al., 2018).

Figure 3. Proposed Flow of K-Means Imputation

In Figure 3, the initial stage of the imputation process begins with filling in temporary values for the missing values in the dataset; this step is essential because the K-Means process cannot be run if there are empty values in the dataset. Next, the data normalization process is carried out using MinMax Scaling. This normalization aims to align the attribute scales that have different value ranges so that no attribute dominates the clustering process. The K-Means process is carried out after the data normalization process. The final results of the K-Means process, namely clusters and centroids, will be used in the imputation process for the existing missing values. The imputation process using centroids is carried out by replacing the missing value based on the attribute centroid that corresponds to the cluster where the data is located. Thus, the missing value is filled with a value that represents the data pattern from the related cluster.

2.2 Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error (MAPE) evaluates the accuracy of prediction models or data imputation by measuring the average absolute percentage error between predicted and actual values, expressed as a percentage for easier interpretation (Chicco

et al., 2021; de Myttenaere et al., 2016; Khair et al., 2017). In this study, MAPE assessed the accuracy of missing data imputation performed using K-Means. A lower MAPE signifies a more reliable and accurate imputation model, serving as a key indicator of the K-Means method's effectiveness in managing missing data in scholarship recipient datasets. The formula for MAPE is provided in Equation 2.

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

3. RESULTS AND DISCUSSIONS

3.1 Data Initialization

The data collected and used in this study are the scholarship applicants' data managed by STMIK PPKIA Tarakanita Rahmawati. The dataset consists of 120 scholarship applicants with five main attributes: semester, GPA, number of dependents, parents' income, and credits completed. A sample of the data can be seen in Table 1.

Tabel 1. Research Data

ID	SMT (A1)	GPA (A2)	Num. of Dependents (A3)	Parents Income (A4)	SKS (A5)
B1	6	3,84	0	1250000	118
B2	6	3,82	6	1000000	112
B3	4	3,75	3	750000	70
B4	2	3,85	5	2000000	23
B5	2	3,83	3	1500000	24
B6	2	3,60	0	1500000	23
B7	2	3,52	2	500000	23
B8	6	3,88	2	1500000	116
B9	6	3,55	3	1000000	113
B10	2	3,87	5	3000000	23

These attributes were chosen because they are considered necessary in the scholarship selection process, which aims to identify students with financial need and academic potential. However, in the data collection process, missing values were found in several attributes that could interfere with the accuracy of the selection. Therefore, this study will use the K-Means Clustering method as an imputation approach to estimate missing values so that the data can be used more accurately in the analysis and decision-making process of scholarship selection.

From a total of 120 scholarship candidate data available, some data will be randomly selected to be simulated as missing data. This study will use two types of missing data compositions, as can be seen in Table 2. Sample dataset 1 contains 10% missing data, so there are 12 missing data from the 120 existing data, and sample dataset 2 contains 20% missing data, so there are 24 missing data.

Tabel 2. Dataset Sample

Dataset	Percentage of Missing Data	Data with Missing Values	Data without Missing Values	Total Data
#1	10%	12	108	120
#2	20%	24	96	120

After identifying the missing data, the next step is to perform imputation using the K-Means Clustering method. This approach is expected to provide more accurate estimations of the missing values compared to simple imputation methods, by considering the data patterns formed within each cluster. A sample of the missing data composition can be seen in Table 3.

Tabel 3. Dataset With Missing Data

ID	SMT (A1)	GPA (A2)	Num. of Dependents (A3)	Parents Income (A4)	SKS (A5)
B4	2	-	5	2000000	23
B7	2	3,52	-	500000	23
B15	4	3,57	2	-	70
B18	-	3,50	1	1000000	104
B29	4	3,66	5	15000000	-
B42	2	3,50	-	3500000	23
B52	4	-	2	1000000	66
B63	4	3,35	3	-	64
B75	6	-	5	500000	118
B98	2	3,24	2	1665000	-

Note: - is missing data

3.2 K-Means Imputation

This study's imputation of missing values utilized the K-Means method, following the steps in Figure 3, primarily conducted using the RapidMiner tool. A significant challenge was that K-Means requires complete datasets to calculate distances, typically using Euclidean distance, which is disrupted by missing values.

To overcome this, an initial imputation was implemented, where missing values were temporarily replaced with the respective attribute's mean value. This ensured a complete dataset for K-Means to calculate distances and perform clustering effectively. For instance, as shown in Table 4, missing values in attribute A2 for data IDs B4, B52, and B75 were filled with the attribute's mean value, 3.4. This approach was applied across all datasets to enable clustering and imputation.

Tabel 4. Sample Dataset After Temporary Filling of Missing Data Using Mean Values

ID	SMT (A1)	GPA (A2)	Num. of Dependents (A3)	Parents Income (A4)	SKS (A5)
B4	2,0	3,4	5,0	2000000,0	23,0
B7	2,0	3,5	3,0	500000,0	23,0
B15	4,0	3,6	2,0	2155355,0	70,0
B18	4,0	3,5	1,0	1000000,0	104,0
B29	4,0	3,7	5,0	15000000,0	75,0
B42	2,0	3,5	3,0	3500000,0	23,0
B52	4,0	3,4	2,0	1000000,0	66,0
B63	4,0	3,4	3,0	2155355,0	64,0
B75	6,0	3,4	5,0	500000,0	118,0
B98	2,0	3,2	2,0	1665000,0	75,0

The next challenge encountered during the clustering process involves attributes with varying ranges of values. This discrepancy in scale can cause certain attributes to dominate the clustering outcomes, resulting in patterns that do not accurately represent the overall data distribution. To address this issue, data normalization was applied using the Min-Max Scaling method. This normalization adjusts the range of each attribute to a uniform scale between 0 and 1. Consequently, each attribute contributes equally during the clustering process, ensuring fairer and more accurate results. The results of min-max scaling can be seen in Table 5.

Table 5. Sample Results of Min-Max Scaling

ID	SMT (A1)	GPA (A2)	Num. of Dependents (A3)	Parents Income (A4)	SKS (A5)
B1	1,00	0,92	0,00	0,05	0,85
B2	1,00	0,91	0,60	0,03	0,79
B3	0,50	0,86	0,30	0,02	0,42
B4	0,00	0,64	0,50	0,10	0,00
B5	0,00	0,92	0,30	0,07	0,01

B6	0,00	0,76	0,00	0,07	0,00
B7	0,00	0,71	0,30	0,00	0,00
B8	1,00	0,95	0,20	0,07	0,83
B9	1,00	0,73	0,30	0,03	0,80
B10	0,00	0,94	0,50	0,17	0,00

In this study, K-Means was tested with three cluster variations: 3, 4, and 5 clusters. Each clustering generated centroids for dataset attributes, which were used as imputation values to replace missing data within the corresponding cluster. This ensures that missing values are replaced with representative values aligned with the cluster's characteristics. Testing different cluster numbers provided options to determine the optimal configuration for accurate imputation. The clustering results are detailed in Figure 4.

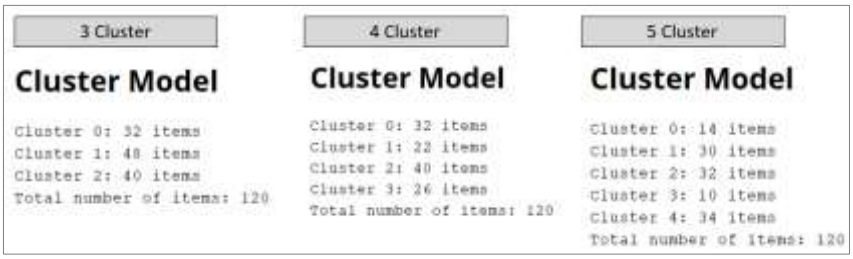
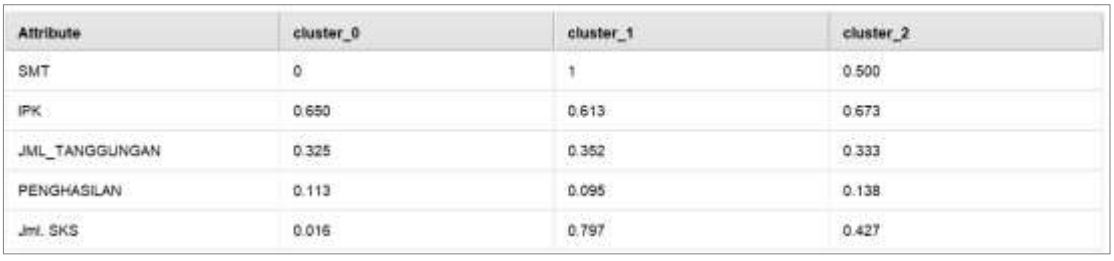


Figure 4. Clustering Results of Dataset 1

Figure 4 illustrates a sample result of clustering using three variations of cluster numbers on Dataset 1. For instance, the model with 3 clusters resulted in the distribution of 32 data points in Cluster 0, 48 data points in Cluster 1, and 40 data points in Cluster 2. Furthermore, Figure 5 shows an example of the centroid distribution for the 3-cluster model. Figure 5 shows the centroids for each attribute in each cluster, which serve as a reference for the data imputation process. Missing values in the dataset are replaced with the centroid value of the corresponding attribute based on the cluster to which the data belongs.



Attribute	cluster_0	cluster_1	cluster_2
SMT	0	1	0.500
IPK	0.650	0.613	0.673
JML_TANGGUNGAN	0.325	0.352	0.333
PENGHASILAN	0.113	0.095	0.138
Jml. SKS	0.016	0.797	0.427

Figure 5. Centroid Distribution in 3 Clusters

Table 6 presents the complete results of data imputation for Dataset 1, demonstrating variations in outcomes depending on the number of clusters used. This process ensures that missing data are filled with values relevant to the clustering patterns. For instance, for data with ID B15, which has a missing value in attribute A4, the imputed values are 0.138 when using 3 clusters, 0.138 with 4 clusters, and 0.118 with 5 clusters.

Table 6. K-Means Imputation on Dataset #1

ID	Atribut dengan Missing Data	Nilai Asli	3 Cluster	4 Cluster	5 Cluster
B4	A2	0,64	0,65	0,65	0,65

B7	A3	0,30	0,325	0,325	0,325
B15	A4	0,11	0,138	0,138	0,118
B18	A1	0,5	0,5	0,5	0,5
B29	A5	0,46	0,427	0,427	0,406
B42	A3	0,30	0,325	0,325	0,325
B52	A2	0,64	0,673	0,673	0,659
B63	A4	0,11	0,138	0,138	0,118
B75	A2	0,64	0,613	0,637	0,622
B98	A5	0,46	0,016	0,016	0,016
B104	A1	0,5	0,5	0,5	0,5
B120	A3	0,30	0,352	0,2	0,247

The data imputation process was also applied to Dataset #2 using centroids from K-Means clusters to replace missing values. The primary distinction between Dataset #1 and Dataset #2 lies in their missing value compositions, as detailed in Table 2. This difference influences the distribution of missing data and the resulting imputations, as cluster centroids are tailored to the unique data patterns of each dataset. The imputation of missing values using the K-Means method is evaluated on two datasets with 10% and 20% missing data. Each dataset is tested with different cluster numbers to assess how the cluster count impacts imputation accuracy. Accuracy is measured using the Mean Absolute Percentage Error (MAPE) in Figure 6.

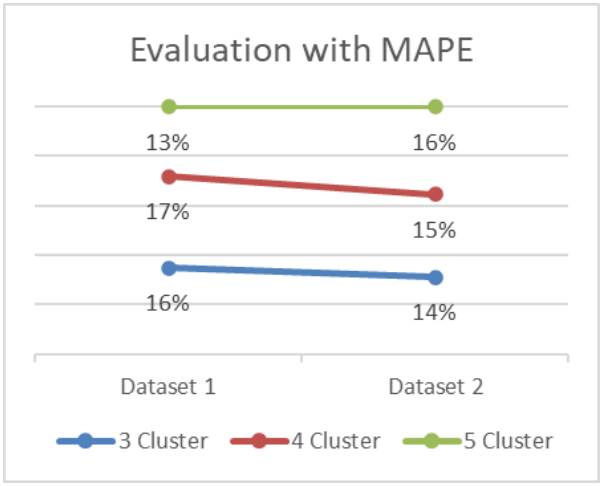


Figure 6. Graph of K-Means Imputation Evaluation Results with MAPE

In Figure 6, for Dataset 1, the best results were achieved using 5 clusters, which resulted in a Mean Absolute Percentage Error (MAPE) of 13%, indicating the lowest error rate. In contrast, configurations with 3 clusters and 4 clusters resulted in MAPE values of 16% and 17%, respectively, reflecting poorer performance compared to 5 clusters. For the dataset with a missing data rate of 20%, the best result was obtained with 2 clusters, which produced a MAPE of 14%. Increasing the number of clusters to 3 and 4 increased the MAPE to 15% and 16%, respectively, indicating that datasets with a higher missing data rate tend to require fewer clusters to generate effective centroids.

From these results, the optimal number of clusters for the missing data imputation process highly depends on the level of missing data. For datasets with a lower missing data rate, a higher number of clusters tends to result in more accurate imputations because they can represent data patterns more precisely. However, for datasets with a higher missing data rate, fewer clusters are more effective because they prevent the data from being divided into overly small groups, which would lead to less accurate imputations. Therefore, selecting the appropriate number of clusters is crucial for achieving optimal missing value imputation.

4. CONCLUSION

This study demonstrates that K-Means Clustering effectively imputes missing data in datasets with numerical attributes and moderate missing data rates (up to 20%). Performance evaluation using MAPE shows that the optimal number of clusters varies with the level of missing data. For a 10% missing data rate, five clusters yielded a MAPE of 13%, while two clusters were optimal for a 20% missing data rate, with a MAPE of 14%. These results suggest that more clusters are effective at capturing detailed patterns in datasets with lower missing data rates, whereas fewer clusters help prevent over-partitioning in datasets with higher missing rates. To enhance generalizability, future research should test this method on datasets with more complex attributes and explore hybrid approaches that combine K-Means with other imputation methods. Developing adaptive techniques that dynamically adjust the number of clusters based on missing data levels also holds potential for improving imputation accuracy.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to STMIK PPKIA Tarakanita Rahmawati and the Doctoral Program in Informatics at Universitas Ahmad Dahlan for their support in facilitating the completion of this research.

REFERENCES

Bangun, B., & Karim, A. K. (2024). Pengembalian Data Yang Hilang Pada Dataset Dengan Menggunakan Algoritma K-Nearest Neighbor Imputation Data Mining. *Jurnal Media Informatika Budidarma*, *8*(3), 1706. https://doi.org/10.30865/mib.v8i3.8014

Chhabra, G., Vashisht, V., & Ranjan, J. (2018). Missing Value Imputation using Hybrid K-Means and Association Rules. *2018 International Conference on Advances in Computing, Communication Control and Networking*, 1163.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, 1–24. https://doi.org/10.7717/PEERJ-CS.623

Dalla, D. P., & Kewuel, H. K. (2023). Ketimpangan Akses Beasiswa dan Pengaruhnya Terhadap Keberlangsungan Studi Mahasiswa. *Educare : Jurnal Penelitian Pendidikan Dan Pembelajaran*, *3*(2), 52–59. https://doi.org/10.56393/educare.v3i2.1702

Darlinda, D., & Utamajaya, J. N. (2022). Sistem Pendukung Keputusan Penerima Beasiswa Program Indonesia Pintar Menggunakan Metode Algoritma K-Means Clustering. *JURIKOM (Jurnal Riset Komputer)*, *9*(2), 167. https://doi.org/10.30865/jurikom.v9i2.3971

de Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean Absolute Percentage Error for regression models. *Neurocomputing*, *192*, 38–48. https://doi.org/10.1016/j.neucom.2015.12.114

Fadlil, A., Herman, & Dikky Praseptian, M. (2023). Single Imputation Using Statistics-Based and K Nearest Neighbor Methods for Numerical Datasets. *Ingenierie Des Systemes d'Information*, *28*(2), 451–459. https://doi.org/10.18280/isi.280221

Fatmawaty, V. S., Riadi, I., & Herman, H. (2024). Klasterisasi Perguruan Tinggi LLDIKTI V Berdasarkan Indikator Kinerja Utama dan PDDIKTI Menggunakan K-Means Clustering. *Jurnal Media Informatika Budidarma*, *8*(2), 878. https://doi.org/10.30865/mib.v8i2.7497

Goa Wea, A., & Adiwidjaja, I. (2018). Pengaruh Beasiswa Terhadap Motivasi Dan Prestasi Belajar Mahasiswa Universitas Tribhuwana Tunggadewi Malang. In *JISIP* (Vol. 7, Issue 1). www.publikasi.unitri.ac.id

Hutagalung, J., & Sonata, F. (2021). Penerapan Metode K-Means Untuk Menganalisis Minat Nasabah. *Jurnal Media Informatika Budidarma*, *5*(3), 1187. https://doi.org/10.30865/mib.v5i3.3113

Kabir, G., Tesfamariam, S., Hemsing, J., & Sadiq, R. (2020). Handling incomplete and missing data in water network database using imputation methods. *Sustainable and Resilient Infrastructure*, *5*(6), 365–377. https://doi.org/10.1080/23789689.2019.1600960

Khair, U., Fahmi, H., Hakim, S. Al, & Rahim, R. (2017). Forecasting Error Calculation with Mean Absolute Deviation and Mean Absolute Percentage Error. *Journal of Physics: Conference Series*, *930*(1). https://doi.org/10.1088/1742-6596/930/1/012002

Liantoni, F., & Agusti, A. (2020). Forecasting Bitcoin Using Double Exponential Smoothing Method Based on Mean Absolute Percentage Error. *JOIV : International Journal On Informatics Visualization* , *4*(2). www.cryptocompare.com.

Lin, W. C., & Tsai, C. F. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, *53*(2), 1487–1509. https://doi.org/10.1007/s10462-019-09709-4

Marcelino, C. G., Leite, G. M. C., Celes, P., & Pedreira, C. E. (2022). Missing Data Analysis in Regression. *Applied Artificial Intelligence*, *36*(1), 2032925. https://doi.org/10.1080/08839514.2022.2032925

Miao, X., Wu, Y., Chen, L., Gao, Y., & Yin, J. (2023). An Experimental Survey of Missing Data Imputation Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *35*(7), 6630–6650. https://doi.org/10.1109/TKDE.2022.3186498

Nasyuha, A. H., Zulham, & Rusydi, I. (2022). Implementation of K-means algorithm in data analysis. *Telkomnika (Telecommunication Computing Electronics and Control)*, *20*(2), 307–313. https://doi.org/10.12928/TELKOMNIKA.v20i2.21986

Praseptian M, D., Fadlil, A., & Herman, H. (2022). Penerapan Clustering K-Means untuk Pengelompokan Tingkat Kepuasan Pengguna Lulusan Perguruan Tinggi. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *6*(3), 1693. https://doi.org/10.30865/mib.v6i3.4191

Privandhani, N. A., & Sulastri, S. (2022). Clustering Pop Songs Based On Spotify Data Using K-Means And K Medoids Algorithm. *Jurnal Mantik*, *6*(2). https://doi.org/10.35335/mantik.v6i2.2517

Rahmayani, T. M. I., & Hidayati, N. (2022). Implemention K-Means Algorithm Determine the Recovery Rate Of Covid-19 Patients In Indonesia. *Jurnal Mantik*, *6*(1), 127–135. https://doi.org/10.35335/jurnalmantik.v6i1.2059

Rangga Baihaqi, M., Padilah, T. N., & Jajuli, M. (2023). Implementasi Metode Imputasi Mean dan Single Center Imputation Chained Equation (SICE) Terhadap Hasil Prediksi Linear Regression pada Data Numerik. *Jurnal Teknologi Informasi Dan Komunikasi)*, *7*(4), 2023. https://doi.org/10.35870/jti

Riadi, A., & Prayudi, I. (2022). Cyberbullying Analysis on Instagram Using K-Means Clustering. *JUITA : Jurnal Informatika*, *10*(2), 261–271.

Rosmini, R., Fadlil, A., & Sunardi, S. (2018). Implementasi Metode K-Means Dalam Pemetaan Kelompok Mahasiswa Melalui Data Aktivitas Kuliah. *IT Journal Research And Development*, *3*(1), 22–31. https://doi.org/10.25299/itjrd.2018.vol3(1).1773

Rustam, S. (2018). Analisa Clustering Phising Dengan K-Means Dalam Meningkatkan Keamanan Komputer. *ILKOM Jurnal Ilmiah*, *10*(2), 175–181.

Ulandari, N. W. A. (2020). Implementasi Metode MOORA pada Proses Seleksi Beasiswa Bidikmisi di Institut Teknologi dan Bisnis STIKOM Bali. *Jurnal Eksplora Informatika*, *10*(1), 53–58. https://doi.org/10.30864/eksplora.v10i1.379

Yulian Pamuji, F., Rofiqul Muslikh, A., Muhammad Arief, R., & Muti, D. (2024). Komparasi Metode Mean dan KNN Imputation Dalam Mengatasi Missing Value Pada Dataset Kecil. *JIP (Jurnal Informatika Polinema)* , *10*(2). https://archive.ics.uci.edu/datasets.