Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2021, 2021:11
https://doi.org/10.28919/cmbn/5344
ISSN: 2052-2541

# A NOVEL CENTROID INITIALIZATION IN MISSING VALUE IMPUTATION TOWARDS MIXED DATASETS

TITIN SISWANTINING[1,*], TAUFIK ANWAR[2], DEVVI SARWINDA[1], HERLEY SHAORI AL-ASH[2]

[1]Department of Mathematics, Universitas Indonesia, Depok, Indonesia

[2]Universitas Indonesia, Depok, Indonesia

**Abstract.** Currently, many databases contain missing values, especially in medical data. Statistical and data mining approaches often require complete data conditions, where these two approaches will not provide adequate performance if the data contains missing values. Several techniques have been made to overcome missing values, one of which is by deleting data containing missing values. However, this approach will omit a lot of information if the data found includes many missing values. This study used an imputation approach (filling in the missing attributes) with a clustering approach. One of the most common clustering approaches is K-Means Clustering. In K-means clustering, the value of the centroid gets from the closest observed value. In this study, we propose updating the centroid value based on the harmonic average of the distance across all observations per centroid. This method is known as K-Harmonic Means Clustering (KHM). We proposed a new program approach for a mixed dataset on three scenarios for missing values of 10%, 20%, and 30%. From the experiments conducted on experimental data sets containing missing values, we get a small proportion of missing values (10%) with a small number of clusters or K, which gives a smaller RMSE value compared to other scenarios.

**Keywords:** clustering; harmonic series; imputation; K-means; mixed dataset.

**2010 AMS Subject Classification:** 68P01.

[*]Corresponding author

E-mail address: titin@sci.ui.ac.id

Received December 21, 2020

## 1. Introduction

Data mining is the process of discovering interesting patterns that consist of but not limited to interesting anomalous situations, trends, patterns, and sequences within the given dataset [1]. Intelligent data analysis techniques are useful for better exploring real-world data sets. However, the real-world data sets almost always suffer from missing data value which defined as a condition in which there is no value for an observation that can result in loss of information and statistical power, this makes the general method for data analysis inappropriate or difficult to apply, and can lead to biased results in estimates derived from statistical models that also becomes a major threat affecting data processing quality [2, 3, 4]. Research conducted by [5] mentioned that in order to ensure that data mining results useful and valuable. It is mandatory to ensure the quality of the collected data because no quality data means no quality mining results.

The missing value data set problem occurs in a wide range of datasets, such as microarray and gene expression data [6], mobile phone data [7], and software project data [8]. Therefore, the presence of missing values in the dataset needs to be addressed before processing data [9, 10, 11].

One of the solutions used to solve the imputation of missing values is divided into two fundamental solutions:

- Removes observations that contain missing values. Deleting observations that contain missing values is one possible solution to overcome the missing value problem but note that removing observations allows the dataset to suffer from observational shortages or loss of information.
- Change the value of the observation attribute that contains missing values. One of the step that can be taken to overcome the problem of missing values is to replace missing values with specific values, such as the mean of the attribute [12]. Using mean value as the imputation method has several drawbacks; mean value may reduce the variance of numerical data and means imputation may distort the relationship between variables because of mean value sensitive to extreme values [13].

Based on the weaknesses of the two approaches that have been mentioned, in this study, we propose an approach to impute missing values. Our proposed method that instead of deleting each observation, we replace the observed value with a method that eschews the extreme. This research's main contribution is the proposed k-harmonic means method for missing value imputation using mixed data (both numerical and categorical). The K-Harmonic Mean algorithm allows us to fill in the missing value with the centroid value of one cluster because the centroid value has a characteristic value that is similar to all observations (members) of a cluster.

## 2. Missing Values Imputation

### 2.1. Missing Values.

Missing values is a condition where the observation value of specific attributes in a dataset is not available. Missing values not only result in loss of information and the power of statistical analysis, causing general data analysis methods to be inappropriate or difficult to apply but can also lead to biased results in estimates derived from statistical models [2].

Based on the probability of the existence of missing values, there are three types of missing values, namely missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [14]. MCAR is a condition where the existence of missing value does not depend on the values of other variables and also does not depend on the existence of the missing value of other variables. Unlike MCAR, the existence of missing value in MAR depends on the value observed in other variables but does not depend on the presence of missing value from other variables. Missing value depends on the value observed in other variables and also depends on the presence of missing values of other variables included in the MNAR category. The handling of missing values in the three mechanisms is influenced by other variables outside the given dataset [15]. In this study, the MCAR mechanism is used in the process of data simulation.

### 2.2. Imputation.

Imputation is an estimation process to replace or estimate missing values with a value that can be estimated using different algorithms or techniques [16]. The imputation method that can be used can be divided into two types, namely the single imputation method and the multiple imputation method [17]. Single imputation is a simple imputation method where missing values are replaced by logical estimates (one estimate per missing values) before applying specific methods to filled data (datasets without missing values). Imputation using mean mode and imputation approach using clustering is an example of a single imputation method. Mean imputation can be considered the simplest approach; missing values are replaced by the average value of each variable in each observation for which the value does not exist as an estimator. Mean imputation is generally used in social science as a fast alternative to data deletion; Imputation with the clustering approach is a simple, intuitive method to accommodate incomplete data [18].

Imputation using mean and mode values can be called rough imputation and is a fast and straightforward imputation approach because it directly uses the mean or mode value of the entire data on the variable to be imputed. The mean is used for the imputation of numerical variables, while the mode is used for imputation for categorical variables. This rough imputation not only does not consider the value of variance but also does not pay attention to the relationship between variables and can produce estimates that are not appropriate. The rough imputation method can only be used if there are only a few missing values and are not intended for general use [19]. Imputation with the clustering approach can be used for more general use. This study uses a single imputation method.

Imputation with the clustering approach is a method that replaces missing values with the values contained in the cluster, through grouping a dataset and grouping observations that have missing values to the data group, and replacing missing values with the average of the same cluster observations. Observations that contain missing values are grouped into clusters, so the values obtained for imputation are values that have characteristics similar to actual data. This study uses a clustering approach with the implementation of the K-Harmonic Means algorithm as a missing value imputation method.

## 3. UNSUPERVISED LEARNING

Unsupervised learning includes a clustering algorithm, in which the input dataset that is still unknown label or target, then partitioned into clusters that meet specific criteria. Unsupervised learning can also be considered as supervised learning with unknown (class) outcomes. This introduces the difficulty of designing an objective function given a particular dataset [20]. The clustering method is a mechanism that can be used in unsupervised learning. Two types of clustering consist of hard clustering and soft clustering. In hard clustering, one data point can only be assigned to one cluster, but in soft clustering, one data point may be a member of more than one cluster [21]. The K-Harmonic means method used in this study is a hard clustering method.

### 3.1. K-Means Clustering.

K-Means clustering, which is one example of hard clustering, has a cost function or function that must be minimized [22] as follows:

$$\zeta = \Sigma_{i=1}^{n}\{min||d_i - C_j||^2|j = 1,...,k\} \tag{1}$$

$$C_{(j,z)} = \frac{1}{n_j}\Sigma_{i=1}^{n_j} d_i \tag{2}$$

According to 1 $d_i$ denotes $observation_i$, $C_j$ is the centroid at $cluster_j$, $i$ serving as observation index $(1,...,n)$, and $||d_i - C_j||^2$ is the Euclidean distance between $d_i$ and $C_j$ which produce smallest possible value.

K-Means clustering is then performed using the following steps [23]:

- define the number of cluster (k).
- insert the observation into cluster k based on the closest centroid and perform new centroid computation using 2.
- repeat steps two and three until the difference of two consecutive iterations on cost function value less than the threshold value.

The calculation of the cost function in K-Means clustering only considers the distance with the smallest value, while the cost function on K-Harmonic mean clustering uses the harmonic mean value from the distance of each observation to each centroid in each cluster. The value of the cost function that must be minimized is in line with the nature of the harmonic mean when compared to the two other mean values as given on 6.

$$A_n(x) = \frac{1}{n}\Sigma_{i=1}^{n} x_i \tag{3}$$

$$G_n(x) = \sqrt{\Pi_{i=1}^{n} x_i} \tag{4}$$

$$H_n(x) = \frac{n}{\Sigma_{i=1}^{n} \frac{1}{x_i}} \tag{5}$$

$$A_n(x) \geq G_n(x) \geq H_n(x) \tag{6}$$

For positive numerical value $x = (x_1, x_2, ..., x_n)$, the value of arithmetic, geometry, and harmonic are given on 3,4,5, respectively. The inequality on 6 shows that the harmonic mean value always generates the smallest value compare to arithmetic and geometry mean value.

## 4. Harmonic Means Clustering

### 4.1. Normalization.

Normalization is a technique to change the scale at which a new range can be obtained from the range of existing data. Normalization is usually done so that the scale of each variable is the same and has no unit value. In this study, normalization is done because there is a distance calculation that requires numerical data values to be between 0 and 1. This study uses the z-score normalization [24]. The results of the min-max normalization have values between 0 to 1, while the Z-score normalization has negative results. The normalization method used in this study is the min-max normalization. Min-max normalization can be done with the following equation:

$$d_{zi} = \frac{x_{zi} - x_{z,min}}{x_{z,max} - x_{z,min}} \tag{7}$$

According to 7 $d_{zi}$ is the normalization result on variable $z$. $x_{zi}$ min is the minimum value or the smallest value on the variable $z$ and $xz_max$ is the maximum value or the most significant value on the $z$ variable.

**4.2. K-Harmonic Means Clustering.** K-Harmonic Means (KHM) and KM are the center-based clustering algorithm, but the KHM algorithm is the result of further studies of the KM algorithm. The main difference between the two algorithms lies in the calculation of the cost function and centroid update calculation; KM only considers the closest observation to the centroid while the KHM uses the harmonic average of the distance of all observations per centroid [22].

$$\zeta = \Sigma_{i=1}^{n} \frac{K}{\Sigma_{j=1}^{K} \frac{1}{||d_i - C_j||^2}} \tag{8}$$

$$\alpha_i = \frac{1}{(\Sigma_{j=1}^{k} \frac{1}{||d_i - C_j||^2})^2} \tag{9}$$

$$q_{i,j} = \frac{\alpha_i}{||d_i - C_j||^4} \tag{10}$$

$$q_j = \Sigma_{i=1}^{i=1} q_{i,j} \tag{11}$$

$$p_{i,j} = \frac{q_{i,j}}{q_j} \tag{12}$$

$$C_{j,z} = \Sigma_{i=1}^{n} p_{i,j} d_{i,z} \tag{13}$$

The distance between observations and centroids commonly used in the KHM for numerical data i.e., the Euclidean distance that given in 7 with the following steps:

- Determine the number of cluster and perform random centroid value initialization.
- Compute cost function using 7, where $\zeta$ is the harmonic mean, $K$ is the number of cluster, $d_i$ is the $observation_i$, $C_j$ is the centroid on $cluster_j$ and $||d_i - C_j||^2$ is the Euclidean distance.

- Compute new centroid using 8, 9, 10, 11 and 12 where $\alpha_i$ is the observation distance function at index $i$, $q_{i,j}$ is the membership function of $observation_i$ on $cluster_j$, $p_{i,j}$ is the $observation_i$ weight forming centroid on $cluster_j$, $C_{j,z}$ is the centroid on $cluster_j$, $variable_z$ and $d_{i,z}$ are the $observation_i$ and $variable_z$.
- Enter the observation into the cluster with the closest centroid distance.
- Repeat step two until step four until the difference of the cost function is less than predefined threshold value or exceed the maximum allowed iteration.

## 5. K-Harmonic Clustering Implementation Methodology

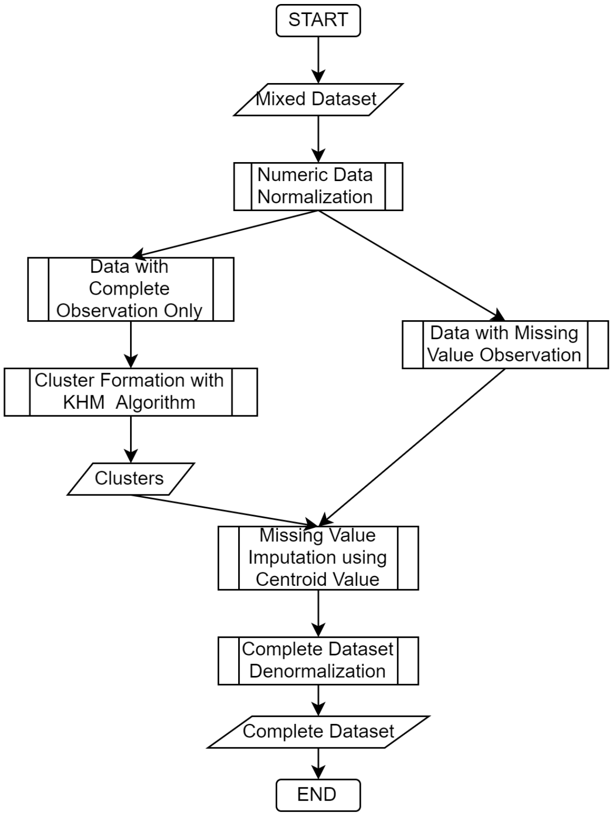
START
Mixed Dataset
Numeric Data Normalization
Data with Complete Observation Only
Data with Missing Value Observation
Cluster Formation with KHM Algorithm
Clusters
Missing Value Imputation using Centroid Value
Complete Dataset Denormalization
Complete Dataset
END

FIGURE 1. Imputation Method Flowchart using K-Harmonic Mean Algorithm

Figure 1 contains the research methodology written in this paper. The KHM approach to the imputation of mixed data (both numerical and categorical data) is represented in 1. Based on 1, the KHM imputation stage starts with normalizing min-max followed by the separation of mixed datasets that still contain missing values into two datasets consisting of complete observations (without missing values) and observations that contain missing values. A complete observation dataset is used to form clusters.

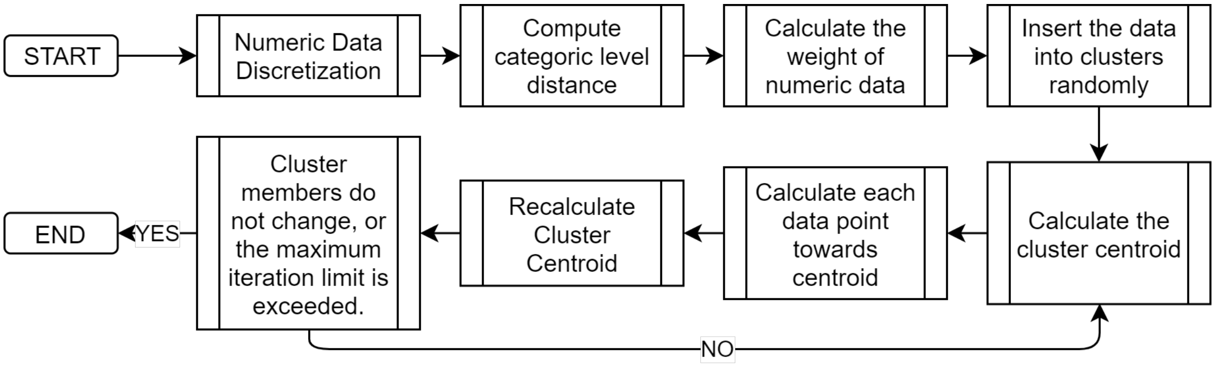
START
Numeric Data Discretization
Compute categoric level distance
Calculate the weight of numeric data
Insert the data into clusters randomly
Calculate the cluster centroid
Calculate each data point towards centroid
Recalculate Cluster Centroid
Cluster members do not change, or the maximum iteration limit is exceeded.
YES
END
NO

FIGURE 2. Imputation Method Flowchart using K-Harmonic Mean Algorithm

A detailed explanation of cluster formation using the KHM algorithm [25] is given in 2. The formation of clusters by the KHM clustering method begins with the discretization of numerical data that has been normalized so that a categorical form is obtained from numerical data using the Equal Width Discretization (EWD) method [26]. The distance between categorical levels can be calculated if all numeric datums have been converted to categorical datums. The distance between categorical levels computes the occurrence with a categorical level variable with other variables as given on 14.

$$\delta^{zl}(x,y) = P(v|x) + P(\neg v|y) - 1 \tag{14}$$

$$\delta(x,y) = \frac{1}{m-1} \sum_{l=1\ldots m, z\neq 1} \delta^{zl}(x,y) \tag{15}$$

$$\delta(x,y) = \frac{1}{m-1} \sum_{l=1\ldots m, z\neq 1} \delta^{zl}(x,y) \tag{16}$$

$$w_z = \sum_{r=1}^{S-1} \sum_{s>r}^{S} \frac{\delta(u[r],u[s])}{\frac{S(S-1)}{2}} \tag{17}$$

$$\vartheta(d_i, C_j) = \sum_{z=1}^{m_r} (w_z(d^r_{iz}) - C^r_{jz})^2 + \sum_{z=1}^{m_c} (\Omega(d^c_{iz}, C^c_{jz}))^2 \tag{18}$$

$$\begin{aligned} q_j &= \sum_{i=1}^{n} q_{i,j} \\ q_{i,j} &= \frac{\alpha_i}{\vartheta(d_i, C_j)^4} \\ \alpha_i &= \frac{1}{(\sum_{j=1}^{m} \frac{1}{\vartheta(d_i, C_j)^2})^2} \end{aligned} \tag{19}$$

$$C_{j,z} = \sum_{i=1}^{n} p_{i,j} d_{i,z} \tag{20}$$

$$\Theta_{z,t,j} = \sum_{i=1}^{n} \eta_i(x_{t,z}, C_j) \tag{21}$$

According to 14 $\delta^{zl}(x,y)$ states the categoric level $x$ and $y$ on variable $z$ considering categoric level co-occurrence variable $l$, $z$ denotes $variable_z$, $l$ denotes $variable_l$, $x$ and $y$ are categoric level on $variable_z$, $v$ denotes categoric level subset on $variable_v$, $\neg v$ denotes complement subset of $v$, $P(v|x)$ is the conditional probability where $x$ value on $variable_z$ occurred together with $v$ value on $variable_l$. The categorical level that is part of $v$ is chosen to maximize the value of $\delta^{zl}(x,y)$.

## 6. Experiments

**6.1. Data.** The data we use are primary data from dr. Cipto Mangunkusumo Hospital. This primary data has received ethical approval issued by the medical faculty of Universitas Indonesia [27]. The data used in this study are atrial fibrillation data without missing values. For simulation purposes, some values in the dataset are intentionally omitted. The dataset used consisted of 15 variables, and 145 observations are given in Table 1. Variables ”Name,” ”W” and ”H” are not included in data processing because the variables ”W” and ”H” are used to calculate ”BMI,” so the variable ”BMI” is included in the data processing.

ethical approval no: 0377/UN2.F1/ETIK/2018

TABLE 1. Dataset Attributes

| No. | Feature | Description |
|---|---|---|
| 1. | Name | Patient Name |
| 2. | Age | Year |
| 3. | Sex | $0 = M$, $1 = F$ |
| 4. | Weight (W) | Kilogram |
| 5. | Height (H) | Meter |
| 6. | Body Mass Index (BMI) | $\frac{W}{H}$ |
| 7. | Waist Size (WS) | Centimeter |
| 8. | Neck Size (NS) | Centimeter |
| 9. | Hypertension | 2 = HS 2, 1 = HS 1, 0 = NBP |
| 10. | Smoking | 0=No, 1=Yes |
| 11. | Alcohol | 0=No, 1=Yes |
| 12. | CHF | 0=No, 1=Yes |
| 13. | CHD | 0=No, 1=Yes |
| 14. | Stroke | 0=No, 1=Yes |
| 15. | Atrial Fibrillation (AF) | 0=No, 1=Yes |

There are 4 numerical variables included in data processing, namely ”Age”, ”BMI”, ”WS”, and ”NS” and 8 categorical variables, namely ”Sex”, ”Hypertension”, ”Smoking”, ”Alcohol”, ”CHF” (Congestive Heart Failure), ”Stroke”, ”CHD” (Coronary Heart Disease), and ”AF”. Note that the word ”HS” in description no. 9 in Table 1 means ”Hypertension Stage” where NBP means ”Normal Blood Pressure”. Descriptive statistics of the data consisting of the average, minimum, and maximum values of each numerical variable in Table 1 are given in Table 2.

### 6.2. Observation Imputation.

The variable ”Alcohol” has 137 data with a value of 0; however, only 8 data has a value of 1, and the variable ”Stroke” there are 138 data with a value of 0 while only 7 data have a value of 1. The next calculation phase uses the new column names (we rename all columns), as given in Table 3.

TABLE 2. Descriptive Statistic Numeric Variables

| No. | Measure | Age | BMI | WS | NS |
|---|---|---|---|---|---|
| 1. | Average | 47.22 | 26.02 | 89.88 | 37.96 |
| 2. | Minimum | 17 | 11.21 | 62 | 27 |
| 3. | Maximum | 83 | 64.12 | 167 | 55 |
| 4. | Range | 66 | 52.91 | 99 | 28 |

TABLE 3. Features Name

| No. | Feature | Variable |
|---|---|---|
| 1. | Age | $X_1$ |
| 2. | Sex | $X_2$ |
| 3. | Body Mass Index | $X_3$ |
| 4. | Waist Size | $X_4$ |
| 5. | Neck Size | $X_5$ |
| 6. | Hypertension | $X_6$ |
| 7. | Smoking | $X_7$ |
| 8. | Alcohol | $X_8$ |
| 9. | Congestive Heart Failure | $X_9$ |
| 10. | Stroke | $X_{10}$ |
| 11. | Coronary Heart Disease | $X_{11}$ |
| 12. | Atrial Fibrilation | $X_{12}$ |

In this imputation process, a process of manipulating missing values (creating missing values from a complete dataset without missing values) with the MCAR mechanism is to obtain random missing values. The proportions of observations made to eliminate this value are 10%, 15%, and 20% of observations. There are 2 to 6 missing values for each observation in the proposition. Data entered into the proportions of 10%, 15%, and 20%, as well as the number of missing values in each observation, were randomly determined.

The K-Harmonic Means (KHM) clustering method is implemented as a method of the imputation of missing values on data that already contains missing values using the MCAR mechanism. Twenty-one times imputation is done with a number of different clusters, from 2 clusters to 8 clusters with the proportion of missing values of 10%, 15%, and 20%. Observations that already have missing values from data that contain missing values with a proportion of 10% are shown in Table 5, where missing values are indicated by na. Imputation is done after getting the optimal cluster after several iterations in forming the cluster. Iteration is carried out in each cluster formation on complete data that has been separated from data containing the proportion of missing values of 10%, 15%, and 20% in each cluster (*K*) selected. Table 5 shows data containing missing values, with a proportion of 10%.

Based on Table 4, the number of iterations in the proportion of missing values (MV) of 10% and 15% tends to rise quite high when *K* is equal to 6, but the opposite occurs in the proportion of missing values of 20% where the number of iterations when K equals 6 has the smallest value. This phenomenon can occur because the number of iterations is influenced by the initial centroid formed from the initialization of randomly selected cluster members. After the optimal cluster is obtained, the centroid in the cluster is the centroid used in the imputation process. Table 6 is an illustration of the centroid value of the optimal cluster formed in the clustering process using complete data that has been separated from data containing missing values with a proportion of missing values of 20% and using 2 clusters.

TABLE 4. Total Iteration On Each Cluster

| Variables | | | | | | | MV Proportion |
|---|---|---|---|---|---|---|---|
| 2 | 3 | 4 | 5 | 6 | 7 | 8 | |
| 4 | 9 | 8 | 8 | 15 | 27 | 13 | 10% |
| 4 | 6 | 7 | 9 | 23 | 17 | 12 | 20% |
| 10 | 9 | 10 | 10 | 6 | 11 | 20 | 30% |

The centroid value of the numerical variable obtained is still in the normal form so that it can be used as the value of the imputation of missing values on the numeric variable, that is the centroid value that has been returned to its original form before normalization, and the

TABLE 5. The Proportion Of 10% Observation Contains Missing Values

| Datum | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 9 | 69 | 1 | 32 | 99 | 40 | 2 | na | na | 1 | 0 | na | na |
| 14 | 77 | 0 | na | na | 32 | 2 | na | 0 | 0 | 1 | 1 | na |
| 26 | na | 1 | 22 | 75 | 35 | 0 | 0 | 0 | na | 1 | na | na |
| 52 | 65 | 0 | 26 | 80 | na | na | 1 | na | na | 0 | 1 | 0 |
| 56 | na | 0 | na | na | na | 1 | 1 | 0 | na | 0 | na | 1 |
| 57 | 68 | na | na | 122 | na | 1 | na | 0 | na | 0 | na | 1 |
| 64 | 42 | 0 | 18 | 75 | na | 2 | 1 | na | 0 | na | 0 | 1 |
| 70 | 48 | 0 | 21 | 72 | 35 | na | 1 | 0 | na | 0 | 0 | 1 |
| 79 | na | na | na | 81 | 35 | 0 | 0 | 0 | na | 0 | 0 | na |
| 111 | 70 | 0 | 22 | 100 | 38 | 0 | 1 | na | 1 | na | na | na |
| 112 | na | na | 23 | 92 | 38 | na | na | na | 0 | na | 0 | 0 |
| 126 | 27 | 1 | na | 72 | 37 | 2 | 1 | 0 | na | 0 | na | na |
| 128 | 35 | 0 | na | 125 | 45 | na | 1 | na | 1 | na | 1 | na |
| 132 | 54 | na | 17 | na | na | 0 | na | 0 | 1 | 0 | na | 0 |

value used for imputation on the categorical variable is the categorical level that has the greatest centroid value. After observations containing missing values have determined the membership of the cluster, then the value becomes the value of imputation. The results of imputation carried out on data containing missing values with a proportion of 10% using 2 clusters are given in Table 7.

The results of the imputation value are shown in bold as given in Table 7. Imputation is carried out to obtain 21 datasets of imputation results from a process that uses several different clusters, namely as many as 2 clusters to 8 clusters and the proportion of different missing values, namely 10%, 15%, and 20%. The dataset of imputation results is then evaluated by looking at the RMSE on numeric variables and the level of accuracy of imputation values on categorical variables.

The results of the imputation of the KHM method on data containing missing values are evaluated by looking at the Root Mean Square Error (RMSE) on numeric variables and the level of accuracy of the imputation values on categorical variables. RMSE is calculated by looking at the results of imputation on each numerical variable compared to the initial value before the mechanism of the missing value is then performed an average RMSE calculation on each variable.

TABLE 6. Centroids In The Proportion Of Missing Values 10% Using $K = 2$

| Variable | Cluster 1 Centroid | Cluster 2 Centroid |
|---|---|---|
| $X_1$ | 0.28018 | 0.65098 |
| $X_2$ | $< 0.60000, 0.33800 >$ | $< 0.66153, 0.41952 >$ |
| $X_3$ | 0.23825 | 0.35535 |
| $X_4$ | 0.19543 | 0.38043 |
| $X_5$ | 0.34153 | 0.46918 |
| $X_6$ | $< 0.40000, 0.21538, 0.20888 >$ | $< 0.33846, 0.21538, 0.72748 >$ |
| $X_7$ | $< 0.50769, 0.47589 >$ | $< 0.49230, 0.52382 >$ |
| $X_8$ | $< 0.95384, 0.05925 >$ | $< 0.92307, 0.05393 >$ |
| $X_9$ | $< 0.66153, 0.16683 >$ | $< 0.60000, 0.64206 >$ |
| $X_{10}$ | $< 0.98461, 0.02469 >$ | $< 0.95384, 0.02409 >$ |
| $X_{11}$ | $< 0.73846, 0.11499 >$ | $< 0.73846, 0.42761 >$ |
| $X_{12}$ | $< 0.73846, 0.20199 >$ | $< 0.66153, 0.38189 >$ |

**6.3. Algorithm Simulation.** The data sample taken in Table 8 is atrial fibrillation data. $x_1$ is gender, $x_2$ is smoking history, and $x_3$ is height. Note that $x_1$ and $x_2$ are categorical variables, while $x_3$ are numeric variables. Each categoric variable has two categorical levels. Categorical levels for gender are male (M) and female (F). Categorical levels for smoking history were yes (Y) and no (N).

TABLE 7. Imputation Result With 10% Missing Value Proportion

| Datum | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 9 | 69 | 1 | 32.42 | 99 | 40 | 2 | **0** | **0** | 1 | 0 | **0** | **0** |
| 14 | 77 | 0 | **23.82** | **82.52** | 32 | 2 | **0** | 0 | 0 | 1 | 1 | **0** |
| 26 | **59.96** | 1 | 22.03 | 75 | 35 | 0 | 0 | 0 | **1** | 1 | **0** | **0** |
| 52 | 65 | 0 | 26.81 | 80 | **40.13** | **2** | 1 | **0** | **1** | 0 | 1 | 0 |
| 56 | **59.96** | 0 | **30.01** | **101.94** | **40.13** | 1 | 1 | 0 | **1** | 0 | **0** | 1 |
| 57 | 68 | 0 | **30.01** | 122 | **40.13** | 1 | **1** | 0 | **1** | 0 | **0** | 1 |
| 64 | 42 | 0 | 18.73 | 75 | **40.13** | 2 | 1 | **0** | 0 | **0** | 0 | 1 |
| 70 | 48 | 0 | 21.48 | 72 | 35 | **2** | 1 | 0 | **1** | 0 | 0 | 1 |
| 79 | **35.49** | 0 | **23.82** | 81 | 35 | 0 | 0 | 0 | **0** | 0 | 0 | **0** |
| 105 | 83 | 0 | 24.44 | 72 | 32 | 0 | **0** | 0 | 0 | 0 | 1 | 1 |
| 111 | 70 | 0 | 22.60 | 100 | 38 | 0 | 1 | **0** | 1 | **0** | **0** | **0** |
| 112 | **59.96** | 0 | 23.04 | 92 | 38 | **2** | **1** | **0** | 0 | **0** | 0 | 0 |
| 126 | 27 | 1 | **30.01** | 72 | 37 | **2** | 1 | **0** | **1** | 0 | **0** | **0** |
| 128 | 35 | 0 | **30.01** | 125 | 45 | **2** | 1 | **0** | 1 | **0** | 1 | **0** |
| 132 | 54 | 0 | 17.66 | **82.52** | **36.56** | 0 | **0** | 0 | 1 | 0 | **0** | 0 |

We deliberately bring up missing values in Table 9. In Table 9, missing values occur in the last three observations. We create missing values by deleting one observed value for one variable for each observation. The NA (not available) values in Table 9 are the missing values.

Then we normalized the numerical data using min-max normalization, with the results, as shown in Table 10. We used ten baseline data that did not contain missing values with $x_3$ due to normalization to form clusters. Then, we do the $x_3$ discretization using the equal width discretization method by changing the data with a value less than equal 0.5 to ”a” and data that is more than 0.5 to ”b.” Calculate the distances a and b to find the weights, starting by finding the conditional probabilities in the following calculations:

TABLE 8. Sample Dataset Without Missing Values

| **No.** | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| 1 | F | N | 1.58 |
| 2 | M | Y | 1.6 |
| 3 | M | Y | 1.7 |
| 4 | F | N | 1.6 |
| 5 | F | N | 1.6 |
| 6 | F | N | 1.62 |
| 7 | F | N | 1.6 |
| 8 | M | Y | 1.69 |
| 9 | M | Y | 1.62 |
| 10 | M | Y | 1.5 |
| 11 | F | N | 1.61 |
| 12 | F | N | 1.51 |
| 13 | M | Y | 1.7 |

$$P(P|a) = \frac{4}{6}$$
$$P(L|a) = \frac{2}{6}$$
$$P(P|b) = \frac{1}{4}$$
$$P(L|b) = \frac{3}{4}$$
$$P(T|a) = \frac{4}{6}$$
$$P(Y|a) = \frac{2}{6}$$
$$P(T|b) = \frac{1}{4}$$
$$P(Y|b) = \frac{3}{4}$$

The calculation of the distance concerning other variables uses equation 14. Then, the calculation of the distance between ”a” and ”b” employs equation 15.

TABLE 9. Sample Dataset With Missing Values

| **No.** | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| 1 | F | N | 1.58 |
| 2 | M | Y | 1.6 |
| 3 | M | Y | 1.7 |
| 4 | F | N | 1.6 |
| 5 | F | N | 1.6 |
| 6 | F | N | 1.62 |
| 7 | F | N | 1.6 |
| 8 | M | Y | 1.69 |
| 9 | M | Y | 1.62 |
| 10 | M | Y | 1.5 |
| 11 | F | N | NA |
| 12 | F | NA | 1.51 |
| 13 | NA | Y | 1.7 |

$$\delta^{3,1}(a,b) = 0.41667$$

$$\delta^{3,2}(a,b) = 0.41667$$

$$\delta(a,b) = 0.41667$$

Equation 16 is used to obtain the weight of $x_3$:

$$w_3 = 0.41667$$

The next step is to calculate the distance F to M and N to Y with equation 14 with the following calculations:

TABLE 10. Sample Dataset with $x_3$ Normalization

| **No.** | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| 1 | F | N | 0.4 |
| 2 | M | Y | 0.5 |
| 3 | M | Y | 1 |
| 4 | F | N | 0.5 |
| 5 | F | N | 0.5 |
| 6 | F | N | 0.6 |
| 7 | F | N | 0.5 |
| 8 | M | Y | 0.95 |
| 9 | M | Y | 0.6 |
| 10 | M | Y | 0 |
| 11 | F | N | NA |
| 12 | F | NA | 0.05 |
| 13 | NA | Y | 1 |

$$\delta^{1,2}(F,M) = 1$$

$$\delta^{1,3}(F,M) = 0.4$$

$$\delta^{2,1}(N,Y) = 1$$

$$\delta^{2,3}(N,Y) = 0.4$$

Equation 14 is used to get delta (F, M) = 0.7 and delta (N, Y) = 0.7. Then we randomly split the complete dataset into two clusters. The clusters are given in Table 3.6 for cluster 1 and Table 3.7 for cluster 2.

Next, we determine the initial centroid based on the cluster formed randomly in Table 12 and Table 13, representing the value of the centroid in the categorical variable representing the value ¡F,M¿ for centroid $x_1$ and ¡N,Y¿ for the centroid of $x_2$ according to the order of the categorical levels that appear on data. The categoric variable centroid's value is the proportion of each categoric level in each cluster, while the centroid in the numeric variable is the average of all

TABLE 11. Sample Dataset after Discretization

| **No.** | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| 1 | F | N | a |
| 2 | M | Y | a |
| 3 | M | Y | b |
| 4 | F | N | a |
| 5 | F | N | a |
| 6 | F | N | b |
| 7 | F | N | a |
| 8 | M | Y | b |
| 9 | M | Y | b |
| 10 | M | Y | a |
| 11 | F | N | NA |
| 12 | F | NA | 0.05 |
| 13 | NA | Y | 1 |

TABLE 12. Randomly Cluster 1

| **No.** | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| 1 | F | N | 0.4 |
| 2 | M | Y | 0.5 |
| 5 | F | N | 0.5 |
| 8 | M | Y | 0.95 |
| 9 | M | Y | 0.6 |

observations in that variable in each cluster. The initial centroids are given in Table 14. The next step is to update the initial centroid starting by calculating each observation's distance to each centroid in Table 14, which is calculated by equation 18, the distance is shown in Table 15.

The distance calculation results obtained in Table 15 are used to calculate the $p_{i,j}$ using equation 19. The obtained $p_{i,j}$ is given in Table 16. Then calculate the new centroid for numeric

TABLE 13. Randomly Cluster 2

| **No.** | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| 3 | M | Y | 1 |
| 4 | F | N | 0.5 |
| 6 | F | N | 0.6 |
| 7 | F | N | 0.5 |
| 10 | M | Y | 0 |

TABLE 14. The Initial Centroid

| **Centroid** | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| $C_{0,1}$ | <0.4 , 0.6 > | <0.4 , 0.6 > | 0.59 |
| $C_{0,2}$ | <0.6 , 0.4 > | <0.6 , 0.4 > | 0.52 |

TABLE 15. Distance between Observation and Initial Centroid

| $i$ | $\vartheta(d_i, C_{0,1})$ | $\vartheta(d_i, C_{0,2})$ |
|---|---|---|
| 1. | 0.359067361 | 0.1593 |
| 2. | 0.15820625 | 0.352869444 |
| 3. | 0.185984028 | 0.3928 |
| 4. | 0.35420625 | 0.156869444 |
| 5. | 0.35420625 | 0.156869444 |
| 6. | 0.352817361 | 0.157911111 |
| 7. | 0.35420625 | 0.156869444 |
| 8. | 0.1793 | 0.384900694 |
| 9. | 0.156817361 | 0.353911111 |
| 10. | 0.217234028 | 0.399744444 |

variables with equation 20 and new centroids for categorical variables with a shape like the one in equation 21. The new centroids obtained are given in Table 17. After getting the new centroid, recalculate each observation's distance to the centroid in Table 17 with equation 18, which can be seen from the calculation results in Table 18. Then each observation becomes a

TABLE 16. Value of $p_{i,j}$

| $i$ | $p_{i,1}$ | $p_{i,2}$ |
|---|---|---|
| 1. | 0.007803152 | 0.190681017 |
| 2. | 0.200022136 | 0.007650911 |
| 3. | 0.192520530 | 0.009159895 |
| 4. | 0.007757771 | 0.190899667 |
| 5. | 0.007757771 | 0.190899667 |
| 6. | 0.008035785 | 0.189571961 |
| 7. | 0.007757771 | 0.190899667 |
| 8. | 0.194800211 | 0.008683855 |
| 9. | 0.201588375 | 0.007356344 |
| 10. | 0.171956498 | 0.014197015 |

TABLE 17. Centroid on $Iteration_1$

| **Centroid** | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| $C_{1,1}$ | <0.03911225, 0.96088775> | <0.03911225, 0.96088775> | 0.618124212 |
| $C_{1,2}$ | <0.95295198, 0.04704802> | <0.95295198, 0.04704802> | 0.502013904 |

cluster member that has a distance from the smallest centroid or can be interpreted as the closest cluster so that a cluster is formed with cluster members as in Table 19.

The next step is to recalculate the centroid. The results of the new centroids are shown in Table 20. Recalculate the distance of each observation to the new centroid until the cluster membership is obtained in $iteration_2$. The results of the distance calculation can be seen in Table 21. Based on Table 19 and Table 20, the membership in $iteration_1$ and $iteration_2$ has not changed, so the iteration stops, and the centroid in the iteration is used in the imputation process of missing values. Data containing missing values are shown in Table 22. Calculate the distance of data containing missing values to the final centroid (centroid in $iteration_2$). The results of these calculations are given in Table 23. Observations that are members of the cluster with the closest centroid are shown in Table 24.

The imputation of missing values is then carried out as follows:

TABLE 18. The Distance between Observation and $Iteration_1$

| $i$ | $\vartheta(d_i, C_{1,1})$ | $\vartheta(d_i, C_{1,2})$ |
|---|---|---|
| 1 | 0.913099262 | 0.003975988 |
| 2 | 0.003921626 | 0.889955831 |
| 3 | 0.026816728 | 0.933008973 |
| 4 | 0.907261615 | 0.00216995 |
| 5 | 0.907261615 | 0.00216995 |
| 6 | 0.904896191 | 0.003836134 |
| 7 | 0.907261615 | 0.00216995 |
| 8 | 0.020620968 | 0.924797409 |
| 9 | 0.001556202 | 0.891622015 |
| 10 | 0.067832079 | 0.933708245 |

TABLE 19. Member of $Iteration_1$

| $Cluster_1$ | $Cluster_2$ |
|---|---|
| $Observation_2$ | $Observation_1$ |
| $Observation_3$ | $Observation_4$ |
| $Observation_8$ | $Observation_5$ |
| $Observation_9$ | $Observation_6$ |
| $Observation_{10}$ | $Observation_7$ |

TABLE 20. Centroid on $Iteration_2$

| **Centroid** | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| $C_{2,1}$ | ¡$1.5654 \times 10^{-10}$, 1¿ | ¡$1.5654 \times 10^{-10}$, 1¿ | 0.611085085 |
| $C_{2,2}$ | ¡0.999994302, $5.69839 \times 10^{-06}$¿ | ¡0.999994302, $5.69839 \times 10^{-06}$¿ | 0.499997374 |

- $observation_{11}$ contains missing values in the numeric variable $x_3$, so the imputation uses the centroid value from $cluster_2$, 0.499997374.
- $observation_{12}$ imputed using the categorical level with the largest $\Theta$ value on variable $x_2$ within cluster 2, which is $N$.

TABLE 21. Member of $Iteration_2$

| $Cluster_1$ | $Cluster_2$ |
|---|---|
| $Observasi_2$ | $Observasi_1$ |
| $Observasi_3$ | $Observasi_4$ |
| $Observasi_8$ | $Observasi_5$ |
| $Observasi_9$ | $Observasi_6$ |
| $Observasi_{10}$ | $Observasi_7$ |

TABLE 22. Data with Missing Values

| **No.** | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| 11 | F | N | ? |
| 12 | F | ? | 0.05 |
| 13 | ? | Y | 1 |

TABLE 23. The Distance between Observation and $Iteration_2$

| **i** | $\vartheta(d_i, C_{2,1})$ | $\vartheta(d_i, C_{2,2})$ |
|---|---|---|
| 11 | 0.904839162 | 0.002169246 |
| 12 | 0.508455192 | 0.03655625 |
| 13 | 0.026067142 | 0.488031409 |

TABLE 24. Membership Observation containing Missing Values

| $Cluster_1$ | $Cluster_2$ |
|---|---|
| $Observation_{13}$ | $Observation_{11}$ |
| | $Observation_{12}$ |

TABLE 25. Imputation Results

| **No.** | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| 11 | F | N | 0.499997374 |
| 12 | F | N | 0.05 |
| 13 | M | Y | 1 |

- $observation_{13}$ imputed using categorical level that has the largest $\Theta$ value on variable $x_1$ within cluster 1, which is $M$.

The results of imputation on the last three observations are given in Table 25. Return the normalization value carried out by min-max normalization, so that the imputation value of $observation_{11}$ is equal to 1.599999475 where the actual observation is 1.61. The imputation results on numerical variables were evaluated using the root mean square error (RMSE) to see the distance between the imputation results and the true value (actual observed value). In this example, there is only one value that is taken into account, which is $observation_{11}$, and the RMSE value is obtained by equation (2.10), which yield

$$RMSE = \sqrt{(1.61-1.599999475)^2} = 0.010000525$$

Imputation on categoric variables is evaluated by looking at the imputation's accuracy compared to the actual data. The imputation of $observation_{12}$ and $observation_{13}$ is in accordance with the original observation so that the imputation accuracy rate is 2/2 or 100% because there are only two categorical values that are imputed, and both are correct.

## 7. Evaluation and Analysis

The level of accuracy of the imputation value on categorical variables can be seen from how many imputation results that correctly fill the actual value of the data on all categorical variables. Note that the smaller the RMSE value, the imputation results get closer to the actual data value, while the higher the level of accuracy, the imputation value actually shows the imputation results are more in line with the actual data value. The number of missing values imputed on each variable for each proportion has the sum shown in Table 26. Evaluation is carried out on the results of imputation using the KHM method with K selected i.e., 2 to 8 clusters, and the proportion of missing values of 10%, 15%, and 20%.

### 7.1. 10 Percent Missing Value Proportion.

Numerical data is evaluated by calculating RMSE which is calculated from the difference in data derived from complete data and data from imputation. Table 27 shows the pair between data derived from the actual value or the initial

TABLE 26. The Number Of Imputation Value On Each Variable

| Variables | | | | | | | | | | | | Missing Value Proportion |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | |
| 4 | 5 | 6 | 3 | 3 | 5 | 5 | 6 | 7 | 4 | 7 | 7 | 10% |
| 3 | 6 | 6 | 6 | 6 | 7 | 6 | 6 | 8 | 7 | 8 | 6 | 20% |
| 6 | 7 | 9 | 7 | 6 | 10 | 7 | 7 | 13 | 13 | 13 | 12 | 30% |

observation (Obs) and imputation data (Imp) with the proportion of missing values 10% where $K = 2$ on numerical variables.

Table 27 shows that there is a difference between the actual values of observations with the imputation results and the RMSE of these values as given in Table 28 line $K = 2$. RMSE values on data containing missing values of 10% for each $K$ selected in numerical variables given in Table 28. We argue that the resulting RMSE value is not too large when considering the range of values for each variable. Figure 3 is given to represent the RMSE average value graph for data containing a 10% missing value. Based on the graph in Figure 3 the average value of RMSE in data that has a 10% missing values proportion is that the smallest RMSE value occurs when $K$ is 3, which is 6.415525 and when $K$ is 6, the value of RMSE tends to increase, reaching 8.586522.

Table 30 shows in pairs between data from complete observations (O) and imputation data (I) with the proportion of missing values of 10% where $K = 2$ on categorical variables. As can be seen in Table 30 that there is a difference between the actual value of the observation and the value of the imputation results, and the number of imputations that exactly match the actual value of the observations can be seen in line $K = 2$ in Table 29. The number of imputation results that exactly match the true value of the observations at each selected K is given in Table 29. Figure 4 illustrates the level of imputation accuracy for each variable and the $K$ value (number of clusters) that chosen.

Based on the exact number of imputations in Table 29 for the number of missing values imputed in Table 26, the level of accuracy is equal to $\frac{1}{5}$ where $K = 2$. Figure 4 represents the level of accuracy 0.2 on the $X_6$ variable not only at $K = 2$ but for each chosen K. Variable $X_8$

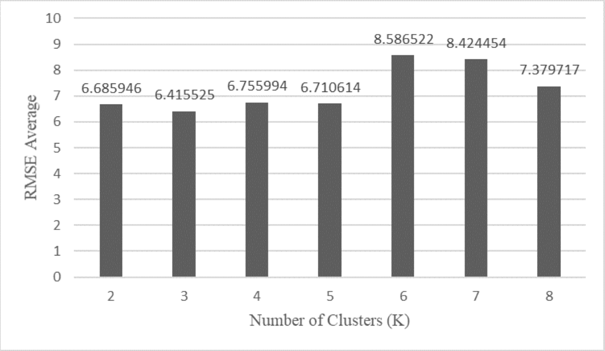

FIGURE 3. RMSE Average using 10% Proportion Of Missing Values

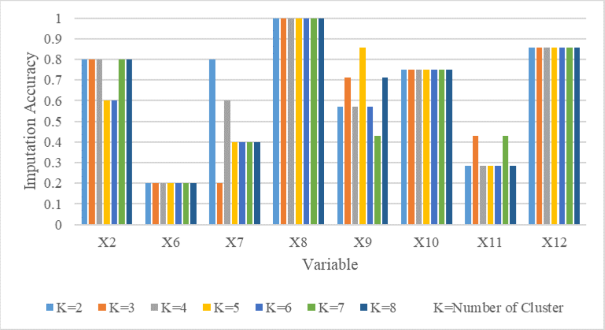

FIGURE 4. Accuracy of Missing Values Imputation with 10% Proportion of Missing Values

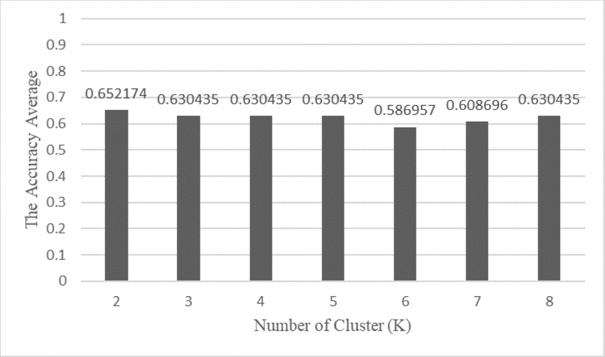

FIGURE 5. The Average Imputation Accuracy on 10% Missing Values

has the level of imputation accuracy reaching a value equal to 1 for each selected $K$. The level of accuracy for the results of imputation in the data of 10% missing values for each $K$ is seen from the average level of accuracy in each variable as given in Figure 5.

According to Figure 5, the level of imputation accuracy in data with the proportion of missing values of 10% can be interpreted that the level of imputation accuracy is approximately 0.6 with

TABLE 27. Actual Value Of Imputation Proportion Towards 10% $K = 2$ Numerical Variables

| $X_1$ | | $X_3$ | | $X_4$ | | $X_5$ | |
|---|---|---|---|---|---|---|---|
| Observed | Imputed | Observed | Imputed | Observed | Imputed | Observed | Imputed |
| 67 | 59.96 | 19.81 | 23.82 | 72 | 82.52 | 40 | 40.13 |
| 48 | 59.96 | 25.51 | 30.01 | 94 | 101.94 | 40 | 40.13 |
| 27 | 35.49 | 31.22 | 30.01 | 90 | 82.52 | 44 | 40.13 |
| 61 | 59.96 | 21.04 | 23.82 | | | 35 | 40.13 |
| | | 20.02 | 30.01 | | | 35 | 36.56 |
| | | 38.51 | 30.01 | | | | |

TABLE 28. RMSE Using 10% Proportion Of Missing Values

| | | Variable | | Clusters |
|---|---|---|---|---|
| $X_1$ | $X_3$ | $X_4$ | $X_5$ | |
| 8.152414 | 6.020911 | 8.751425999999999 | 3.819035 | 2 |
| 7.083157000000001 | 6.795877000000001 | 8.642387 | 3.140678 | 3 |
| 8.474581 | 7.86795 | 7.816814999999999 | 2.864632 | 4 |
| 8.654191 | 5.924933 | 7.7450470000000005 | 4.518288 | 5 |
| 9.947853 | 9.723507000000001 | 11.90986 | 2.764871 | 6 |
| 9.474825 | 8.702898 | 11.88722 | 3.6328699999999996 | 7 |
| 8.116772000000001 | 7.897187 | 10.94726 | 2.557649 | 8 |

the highest accuracy level of 0.652174 where $K$ value is equal to 2 and the smallest accuracy value is 0.586957 where $K$ is equal to 6.

**7.2. 15 Percent Missing Value Proportion.** RMSE values on data containing missing values of 15% in each variable for each $K$ selected in the numerical variable are shown in Table 31. The resulting RMSE is not too large when considering the range of values for each variable.

The average RMSE value for each $K$ chosen in the data containing 15% missing values can be seen from the graph in Figure 6. The average graph of RMSE as given in Figure 6, values in

TABLE 29. Total Imputation With 10% Missing Value Proportion, Which Exactly Fills In The Correct Value

| Variable | | | | | | | | Cluster ($K$) |
|---|---|---|---|---|---|---|---|---|
| $X_2$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | |
| 4 | 1 | 4 | 6 | 4 | 3 | 2 | 6 | 2 |
| 4 | 1 | 1 | 6 | 5 | 3 | 3 | 6 | 3 |
| 4 | 1 | 3 | 6 | 4 | 3 | 2 | 6 | 4 |
| 3 | 1 | 2 | 6 | 6 | 3 | 2 | 6 | 5 |
| 3 | 1 | 2 | 6 | 4 | 3 | 2 | 6 | 6 |
| 4 | 1 | 2 | 6 | 3 | 3 | 3 | 6 | 7 |
| 4 | 1 | 2 | 6 | 5 | 3 | 2 | 6 | 8 |

TABLE 30. Actual Value And Imputation Proportion Of 10% Where $K = 2$ Categorical Variables

| $X_2$ | | $X_6$ | | $X_7$ | | $X_8$ | | $X_9$ | | $X_{10}$ | | $X_{11}$ | | $X_{12}$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Obs | Imp | Obs | Imp | Obs | Imp | Obs | Imp | Obs | Imp | Obs | Imp | Obs | Imp | Obs | Imp |
| 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| 0 | 0 | 0 | 2 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| 1 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 1 | 1 | | | 1 | 0 | 0 | 0 |
| | | | | | | 0 | 0 | 0 | 0 | | | 0 | 0 | 0 | 0 |
| | | | | | | | | 0 | 1 | | | 0 | 0 | 0 | 0 |

data that have a proportion of missing values of 15%, the smallest RMSE value occurs when $K$ is 5, which is 8.15191 and has the highest RMSE value when $K$ is 6, 10.93973. The results of categorical data imputation that correctly fill in missing values are given in Table 32.

The number of imputations that correctly fills missing value with the actual values shown in Table 32 when compared with the number of values imputed in Table 26 yields the level

TABLE 31. RMSE Using 15% Proportion Of Missing Values

| Variable | | | | Cluster($K$) |
|---|---|---|---|---|
| $X_1$ | $X_3$ | $X_4$ | $X_5$ | |
| 6.04 | 5.85 | 15.41 | 5.76 | 2 |
| 6.31 | 3.49 | 16.52 | 5.71 | 3 |
| 5.47 | 5.77 | 17.08 | 5.9 | 4 |
| 4.23 | 4.23 | 13.40 | 6.01 | 5 |
| 4.74 | 8.18 | 17.99 | 6.52 | 6 |
| 6.79 | 4.92 | 12.44 | 7.68 | 7 |
| 7.22 | 5.44 | 12.90 | 6.39 | 8 |

TABLE 32. Total Imputation with 15% Missing Value Proportion, which Exactly Fills in The Correct Value

| Variable | | | | | | | | ($K$) |
|---|---|---|---|---|---|---|---|---|
| $X_2$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | |
| 6 | 0 | 4 | 5 | 5 | 7 | 5 | 5 | 2 |
| 6 | 1 | 2 | 5 | 5 | 7 | 5 | 5 | 3 |
| 6 | 1 | 1 | 5 | 3 | 7 | 5 | 5 | 4 |
| 6 | 0 | 4 | 5 | 4 | 7 | 5 | 5 | 5 |
| 6 | 1 | 3 | 5 | 3 | 7 | 5 | 4 | 6 |
| 6 | 1 | 2 | 5 | 4 | 7 | 5 | 5 | 7 |
| 6 | 0 | 2 | 5 | 4 | 7 | 6 | 5 | 8 |

of accuracy of the imputation results, as shown in Figure 7.Note that in table 32, column $K$ represents the cluster number. Figure 7 explains the highest level of accuracy in the $X_2$ and $X_{10}$ variables, which reach 100% and the lowest in the $X_6$ variable when the chosen $K$ are 2, 5, and 8, resulting in a precision level of 0%.

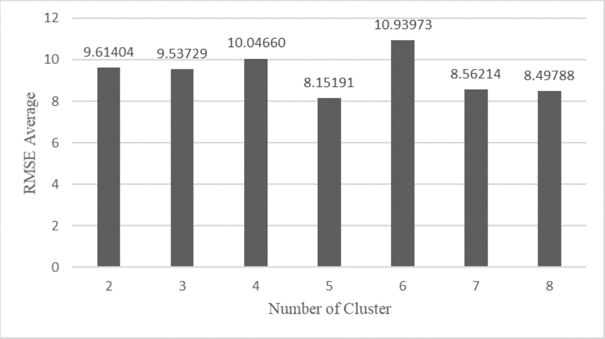

FIGURE 6. RMSE Average using 15% Proportion Of Missing Values

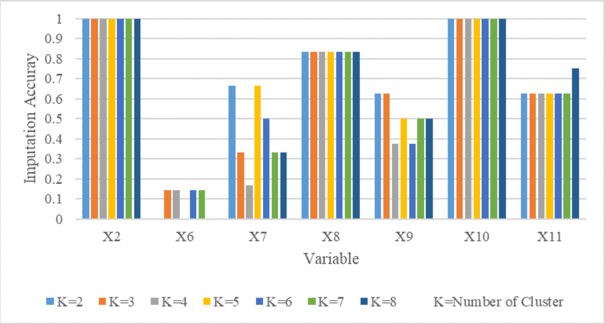

FIGURE 7. Accuracy of Missing Values Imputation with 15% Proportion of Missing Values

The accuracy of each variable in Figure 7 is calculated on average to see the accuracy in data imputation containing missing values with a proportion of 15% for each selected $K$ value given in Figure 6. From the graph, the accuracy level in the data with the proportion of missing values of 15% can be seen that the highest level of accuracy when $K$ equals two is 0.68519 and the smallest when $K$ equals four is 0.61111.

### 7.3. 20 Percent Missing Value Proportion.

The RMSE value in the given dataset containing missing values of 20% for each $K$ selected in the numerical variable is shown in Table 33. The resulting RMSE value is not too large when considering the range of values for each variable. The average RMSE value for each $K$ value (number of clusters) selected in the given dataset containing missing values of 20% is given in Figure 8.

Based on the graph Figure 10, the average RMSE value in the given data, it has a proportion of missing values of 20%, the smallest RMSE value occurs where $K$ equals 5 which reaches 11.86424, and the highest RMSE value occurs where $K$ equals 4 which reaches 14.66473.

TABLE 33. RMSE Using 20% Proportion Of Missing Values

| Variable | | | | Cluster ($K$) |
|---|---|---|---|---|
| $X_1$ | $X_3$ | $X_4$ | $X_5$ | |
| 19.84 | 7.42 | 20.78 | 3.13 | 2 |
| 19.66 | 6.94 | 17.15 | 4.14 | 3 |
| 27.36 | 7.36 | 19.67 | 4.25 | 4 |
| 18.83 | 8.10 | 16.11 | 4.40 | 5 |
| 18.96 | 9.30 | 15.24 | 4.76 | 6 |
| 19.09 | 10.21 | 15.74 | 5.87 | 7 |
| 19.02 | 8.24 | 18.44 | 5.01 | 8 |

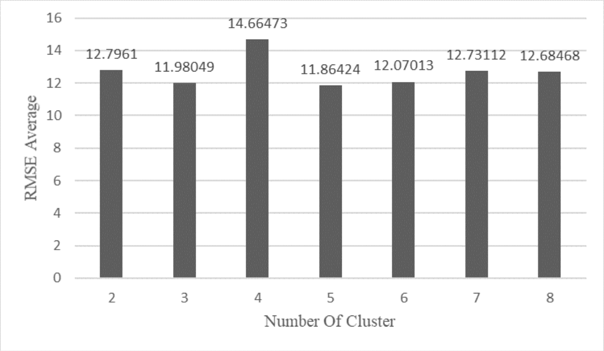

FIGURE 8. RMSE Average using 20% Proportion Of Missing Values

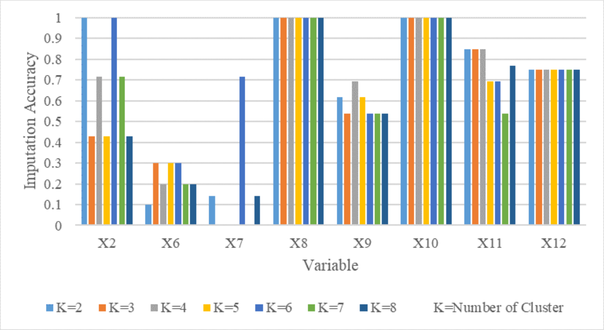

FIGURE 9. Accuracy of Missing Values Imputation with 20% Proportion of Missing Values

Table 8 shows the results of categorical data imputation based on the number of imputation results that correctly fill in the missing values. The exact number of imputations filling in the missing values matches the actual values shown in Table 8 when compared to the number of values imputed in Table IX, the accuracy of the imputation results is shown in Figure 9. According to Figure 9, it can be seen that the highest accuracy value on variables $X_8$ and $X_{10}$ which surprisingly reaches 100% and the lowest accuracy value on variables $X_6$ and $X_7$.

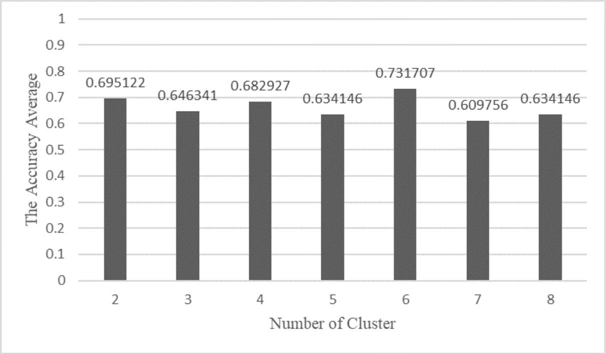

FIGURE 10. The Average Imputation Accuracy on 20% Missing Values

For variable $X_7$, when the chosen $K$ value is 3, 4, 5, and 7, it produces an accuracy of 0%. The average calculation is applied to the accuracy of each variable given in Figure 9 to see the accuracy of data imputation containing missing values with a proportion of 20% for each $K$ used as given in Figure 10. From the graph Figure 10, the level of accuracy in a given dataset with the proportion of missing values of 20% has the highest accuracy value where $K$ is 6, reaching 0.731707 and the smallest when $K$ is 7, which reaches 0.609756.

## 8. CONCLUSIONS

Based on the research that has been done, we concluded that the K-Harmonic Means Clustering method could be implemented as a method of the imputation of missing values in mixed data with a proportion of missing values of 10%, 15%, and 20%. The results of imputation experiments using the K-Harmonic Means method have the most optimal results on the dataset with the proportion of missing values of 10% where the selected $K$ (number of clusters) is equal to 3, RMSE produces 6,415,525, and an accuracy value of 0.630435. We suggest that future studies be able to use data with a more diverse categorical level and balanced frequency, apply the K-Harmonic Means imputation method for imputation of missing values in datasets that already contain missing values from the start and evaluate the results of imputation using other methods, and find methods for determining the number of clusters used so that the results of imputation obtained are optimal.

## ACKNOWLEDGMENTS

Universitas Indonesia funded this research with grant number NKB-1955/UN2.RST/HKP.05.00/2020.

## CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

## REFERENCES

[1] A. Viloria, J.R. López, D.M. García Leyva, C. Vargas-Mercado, H. Hernández-Palma, N.O. Llinas, M.A. David, J.V. Rodriguez, Data Mining Techniques and Multivariate Analysis to Discover Patterns in University Final Researches, Procedia Computer Sci. 155 (2019), 581–586.

[2] Q. Song, M. Shepperd, Missing Data Imputation Techniques, Int. J. Bus. Intell. Data Mining, 2 (2007), 261.

[3] Nurzaman, T. Siswantining, S.M. Soemartojo, D. Sarwinda, Application of Sequential Regression Multivariate Imputation Method on Multivariate Normal Missing Data, in: 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS), IEEE, Semarang, Indonesia, 2019: pp. 1–6.

[4] E.F. Akmam, T. Siswantining, S.M. Soemartojo, D. Sarwinda, Multiple Imputation with Predictive Mean Matching Method for Numerical Missing Data, in: 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS), IEEE, Semarang, Indonesia, 2019: pp. 1–6.

[5] C.-F. Tsai, M.-L. Li, W.-C. Lin, A class center based approach for missing value imputation, Knowl.-Based Syst. 151 (2018), 124–135.

[6] M.C. de Souto, P.A. Jaskowiak, I.G. Costa, Impact of missing data imputation methods on gene expression clustering and classification, BMC Bioinform. 16 (2015), 64.

[7] S. Liu, Selecting a destination image for a capital city rather than for a nation: A segmentation study, J. Destinat. Market. Manage. 3 (2014), 11–17.

[8] Q. Song, M. Shepperd, X. Chen, J. Liu, Can k-NN imputation improve the performance of C4.5 with small software project data sets? A comparative evaluation, J. Syst. Softw. 81 (2008), 2361–2370.

[9] Z. Liu, Q. Pan, J. Dezert, A. Martin, Adaptive imputation of missing values for incomplete pattern classification, Pattern Recognit. 52 (2016), 85–95.

[10] K. Aristiawati, T. Siswantining, D. Sarwinda, S.M. Soemartojo, Missing values imputation based on fuzzy C-Means algorithm for classification of chronic obstructive pulmonary disease (COPD), in: Yogyakarta, Indonesia, 2019: p. 060003.

[11] H.S. Al-Ash, M.F. Putri, P. Mursanto, A. Bustamam, Ensemble Learning Approach on Indonesian Fake News Classification, in: 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS), IEEE, Semarang, Indonesia, 2019: pp. 1–6.

[12] Y. Qin, S. Zhang, X. Zhu, J. Zhang, C. Zhang, POP algorithm: Kernel-based imputation to treat missing values in knowledge discovery from databases, Expert Syst. Appl. 36 (2009), 2794–2804.

[13] M. Zaki, W. Meira, Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge University Press, 2014.

[14] P.E. McKnight, K.M. McKnight, S. Sidani, A.J. Figueredo, Missing data: a gentle introduction. Guilford, New York, 2007.

[15] S. van Buuren, Flexible Imputation of Missing Data, Second Edition, Chapman & Hall, Boca Raton, 2018.

[16] T. Mahboob, A. Ijaz, A. Shahzad, M. Kalsoom, Handling Missing Values in Chronic Kidney Disease Datasets Using KNN, K-Means and K-Medoids Algorithms, in: 2018 12th International Conference on Open Source Systems and Technologies (ICOSST), IEEE, Lahore, Pakistan, 2018: pp. 76–81.

[17] C.K. Enders, Applied Missing Data Analysis. The Guilford Press, New York, NY, 2010.

[18] J.M. Jerez, I. Molina, P.J. García-Laencina, E. Alba, N. Ribelles, M. Martín, L. Franco, Missing data imputation using statistical and machine learning methods in a real breast cancer problem, Artif. Intell. Med. 50 (2010), 105–115.

[19] Z. Zhang, Missing data imputation: focusing on single imputation, Ann. Transl. Med. 4 (1) (2016), 9.

[20] T.J. Sejnowski, G.E. Hinton, (eds.): Unsupervised Learning: Foundations of Neural Computation. MIT Press, Cambridge, 1999.

[21] A.K. Bansal, J.I. Khan, S.K. Alam, Introduction to computational health informatics, CRC Press, Boca Raton, 2020.

[22] B. Zhang, M. Hsu, U. Dayal, K-harmonic means-a data clustering algorithm. Hewlett-Packard Labs Technical Report HPL-1999-124, 1999.

[23] O. Maimon, L. Rokach, (eds.): Data Mining and Knowledge Discovery Handbook, 2nd edn. Springer, Heidelberg, 2010.

[24] S.G.K. Patro, K.K. Sahu, Normalization: A Preprocessing Stage, ArXiv:1503.06462 [Cs]. (2015).

[25] T. Anwar, T. Siswantining, D. Sarwinda, S.M. Soemartojo, A. Bustamam, A study on missing values imputation using K-Harmonic means algorithm: Mixed datasets, in: Surakarta, Indonesia, 2019: p. 020038.

[26] R. Dash, R. Paramguru, R. Dash, Comparative Analysis of Supervised and Unsupervised Discretization Techniques, Int. J. Adv. Sci. Technol. 2 (2011), 29-37.

[27] D.N. Vitasari, T. Siswantining, T. Kamelia, Identification of factor affecting atrial fibrillation in a patient with risk of obstructive sleep apnea at Rumah Sakit dr.Cipto Mangunkusumo using decision tree method, J. Phys.: Conf. Ser. 1321 (2019), 022108.