Performance Comparison of Hot-Deck Imputation, K-Nearest Neighbor Imputation, and Predictive Mean Matching in Missing Value Handling, Case Study: March 2019 SUSENAS Kor Dataset

T Rhaudatunnisa[1], **N Wilantika**[1]

[1]Politeknik Statistika STIS, Jakarta, Indonesia

*Corresponding author's e-mail: 221710035@stis.ac.id

Abstract. Missing value can cause bias and makes the dataset not represent the actual situation. The selection of methods for handling missing values is important because it will affect the estimated value generated. Therefore, this study aims to compare three imputation methods to handle missing values—Hot-Deck Imputation, K-Nearest Neighbor Imputation (KNNI), and Predictive Mean Matching (PMM). The difference in the way the three methods work causes the estimation results to be different. The criteria used to compare the three methods are the Root Mean Squared Error (RMSE), Unsupervised Classification Error (UCE), Supervised Classification Error (SCE), and the time used to run the algorithm. This study uses two pieces of analysis, comparison analysis, and scoring analysis. The comparative analysis applying a simulation that pays attention to the mechanism of missing value. The mechanism of the missing value used in the simulation is Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). Then, scoring analysis aims to narrow down the results of comparative analysis by giving a score on the results of the imputation of the three methods. The result suggests Hot-Deck Imputation is the most excellent in dealing with a missing value based on the score.

1. Introduction

In the data production process, a dataset needs to go through pre-processing first before it is carried out for further analysis. One process that is quite important in data pre-processing is data cleaning. The data cleaning process aims to reduce the impact of data that does not meet the data normality requirements, such as value inconsistencies, noise, and data incompleteness. One of the topics that are quite a lot raised in data cleaning is missing value [1]. As stated in [2], a missing value is a value that is not available for an object. Missing values are divided based on the pattern and the type of missing[3]. Based on its type, the missing value is divided into three:

1. Missing at Random (MAR)
   MAR is a mechanism for missing data distributed randomly for some units of observation. In other words, MAR means missing data is only related to the response/observation variable.
2. Missing Completely at Random (MCAR)
   MCAR is a mechanism for missing data that is randomly distributed for all observation units. In other words, MCAR means that missing data is not related to the values of all variables. Whether they are variables with missing values or observed variables, it means that missing data occur randomly.
3. Missing Not at Random (MNAR)

MNAR mechanism for missing data that is not randomly distributed. In other words, Missingness Is Non-Ignorable that the occurrence of missing data on a variable is related to the variable itself, so it cannot be predicted from other variables in a dataset.

There are so many causes of the missing value in a dataset, such as a difficulty to meet the respondents, data not recorded by officers, errors in the application or equipment used, and otherwise[4]. Missing value can appear as an outlier or inconsistent value from the previous value or an abnormal entry in the data[1]. Missing value basically will not be a problem for the accurate data, especially if the amount is only a little, for example, only 1% of all data. However, in reality, the missing value in the data has a relatively large percentage. The phenomenon of missing value is often found in surveys. Non-response is one of the biggest reasons for missing values in the dataset [5]. Several problems are related to missing values, ranging from loss of efficiency, complications in handling and analyzing data, to the problem of bias generated between data containing missing values and complete data[6]. The impact of missing values is that the data will be biased and not represent the actual situation.

To avoid problems caused by missing values, researchers usually use several methods to handle missing values, including Listwise Deletion, Pairwise Deletion, and Imputation[7]. Listwise Deletion is deleting cases (objects) that contain missing data as a whole[4]. An easy picture is if a variable with "A" missing value in the observation unit, "A" will be removed from the dataset. Pairwise Deletion is removing missing data so that only the available values are analyzed [4]. In contrast to Listwise Deletion, Pairwise deletion will only delete variables with missing values in one unit of observation. Only variables that have a value are left from that unit. Meanwhile, Imputation fills in the missing value with possible values based on the information available in the data. Of the three methods, the imputation method is the best method that can be used to overcome missing values compared to the other two methods[8]. The use of the listwise deletion method and the pairwise deletion method makes it possible to eliminate many unit variables. If because of the deletion, the observation unit does not meet the required number of samples, then resampling is needed so that the number of observation units is sufficient for the required number of samples. Of course, this will take a long time and cost a lot of money. Therefore, it is not recommended to use the deletion method, especially for surveys with many samples. The imputation method can overcome problems in both methods because it avoids the resampling while estimating the missing value by using an observation unit that does not experience a missing value[9].

The imputation method is divided into two types: the statistical-based imputation method and the machine learning-based imputation method[10]. *Statistical imputation technique* is an imputation technique using statistical rules in imputation. While the machine learning imputation technique is an imputation technique that utilizes training on data that will later be used to predict the value to be imputed. The imputation method is broadly divided into two types: the single imputation method and the multiple imputation method[4]. The difference between the two imputation methods lies in the use of training data to impute missing values. The selection of methods for handling missing values is important because it will affect the estimated value generated. In addition, users often have difficulty determining the proper missing value handling method for their data[3].

Therefore, several imputation methods will be compared in this study, including Hot-Deck Imputation and K-Nearest Neighbor Imputation (KNNI), and Predictive Mean Matching (PMM). The Hot-Deck Imputation method is used to overcome missing values by imputing values taken from complete observation units that are considered to have similarities with observation units that have missing values[11]. The hot-deck imputation method was chosen because this method is suitable for use on many types of data and can be used for imputing various types of data[5]. In contrast, the KNNI method provides a more robust and sensitive method[5]. The KNNI method itself is an imputation method that utilizes the K-Nearest Neighbor (KNN) algorithm. The KNN algorithm is a non-hierarchical data grouping algorithm. The number of groups to be formed shall be determined first [12]. In addition, this method does not use any assumptions, does not require the formation of a predictive model, and can overcome missing values in both numerical and categorical data. Hot-Deck Imputation was chosen because this method is the best in statistical-based imputation methods. This method is suitable for use on many types of data and can be used for imputing various types of data [5]. Meanwhile, the best machine learning-based imputation method is held by the KNNI method [12].

Although both methods are considered excellent in dealing with missing values, both methods have quite complicated algorithms which cause these two methods to be rarely used [13]. In a simple search on Google with the keyword “missing data”, it was found that the Predictive Mean Matching (PMM) method had more than 21,000 results from all of the current results [13]. This number is much higher compared to KNNI and Hot-Deck Imputation, which only had 13,000 results and 4,300 results. This shows that the PMM method is a method that is used quite often, even though this method shows lower performance compared to the other two methods. Based on this reason, this paper will compare the three imputation methods by applying them to some datasets. The three methods will be compared based on the scores obtained from several main criteria that have been determined. In particular, this paper will look at the differences in the three methods in dealing with missing values in the data, know the advantages and disadvantages of the Hot-Deck Imputation, KNNI, and PMM methods, and determine the best method based on the highest score in dealing with missing values.

2. Literature Review

2.1. Hot-Deck Imputation

The Hot-Deck Imputation method is used to overcome missing values by imputing values taken from complete observation units that are considered to have similarities with observation units that have missing values. The unit of observation whose value is taken to input the missing value is called the donor[14]. The unit of observation that receives the value contribution is named the recipient. The most significant possible error in this method is that one donor can be selected for several recipients. This method makes the opportunity for all missing values to be filled by the same donor. Therefore, to overcome this problem, one donor is only allowed for a maximum of three recipients in this study.

The phrase Hot-Deck Imputation generally refers to Sequential Hot-Deck Imputation, which means that the imputation of missing values is carried out sequentially from one observation unit to another[15]. The relatively fast and straightforward algorithm makes Hot-Deck Imputation quite popular[5]. In addition, Hot-Deck Imputation is an improvement from the previous method so that this method is suitable for various data types. The hot-Deck Imputation method is illustrated in Figure 1

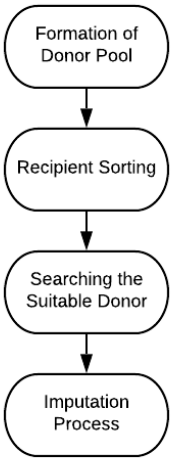

Figure 1. Hot-Deck Imputation Process

2.2. K-Nearest Neighbor Imputation (KNNI)

The KNNI method is an imputation method that utilizes the K-Nearest Neighbor (KNN) algorithm. The KNN algorithm is a non-hierarchical data grouping algorithm. The number of groups to be formed is known and determined[12]. One of the popular techniques for imputation is K-Nearest Neighbor. The value to be imputed is calculated by finding the value in the training set closest to it and the average of this nearest point to fill in the value[16]. KNNI can provide a more robust and sensitive method for

estimating the missing value[17]. In addition, KNNI can exceed the predictive ability of the commonly used K-Means method. The KNNI method algorithm is summarized in Figure 2

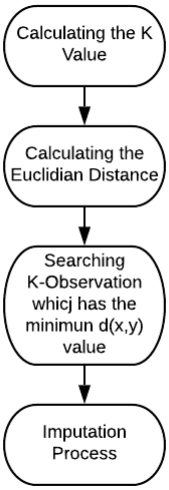

Figure 2. KNNI Process

The Euclidian distance used in this study is calculated using the following formula:

$$d(x,y) = \sqrt{\sum_{j=1}^{s} (x_j - y_j)^2} \quad (1)$$

where:
= target observation vector with s variables
= observation vector that does not contain missing values with s variables
($\boldsymbol{x}$,) = distance between and
j = value of the j-th variable in
j = value of the j-th variable in
j = 1,2, ..., s

In the KNNI method, the imputation process uses a weighted mean imputation procedure which is calculated using the formula:

$$\hat{x_j} = \frac{1}{W} \sum_{k}^{K} w_k y_{kj} \quad (2)$$

where:
j = imputed value
kj = value of the j-th variable on the k-th observation,
k = weight of k-th observation,

2.3. Predictive Mean Matching (PMM)

One approach to estimating missing data is taking a random sample from the observed data based on the minimum distance difference. The method in multiple imputations that utilizes this approach is the PMM method[18]. The PMM method has a similar way of working with the linear regression method. This is what causes this method to be used quite often for research methods because the algorithm is simple but produces a pretty good imputation value. This method needs to pay attention to several things: missing data patterns, missing data mechanisms, types of variables, and data distribution[19]. The stages carried out in this method are depicted in Figure 3.

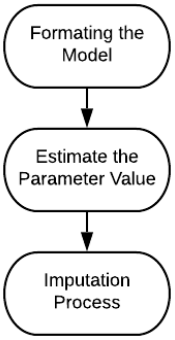

Figure 3. PMM Process

The linear regression model used is $Y_i \sim N(X_i\beta, \sigma^2)$, with the parameters of the model, are:

$$\widehat{\sigma_1}^2 = \frac{\sum_{obs}(Y_i - X_i\hat{\beta}_1)^2}{(n_1 - q)} \tag{3}$$

$$\hat{\beta}_1 = V[\sum_{obs} X^t{}_i Y_i] \tag{4}$$

$$V = \left[\sum_{obs} X^t{}_i Y_i\right]^{-1} \tag{5}$$

The steps taken in the imputation process in the PMM method are [20]:

1. Take one value of a random variable that spreads ${\chi^2}_{n_1-q}$ for example g and calculate:

$$\sigma_*{}^2 = \frac{\hat{\sigma}^2{}_1(n_1 - q)}{g} \tag{6}$$

2. Take q random variables that spread N(0,1) to create a q-component vector Z and calculate:

$$\beta_* = \hat{\beta}_1 + \sigma_*[V]^{1/2}Z \tag{7}$$

where $[V]^{1/2}$ is the upper triangular matrix in the Cholesky decomposition.

3. Calculate the value of *Ymis* by $Y_{i*} = X_i\beta_* \;\; i \; \epsilon \; mis$, for each $Y_{i*} \;\; i \; \epsilon \; mis$ find respondent Y_i whose value is closest to Y_{i*} and input that value for *Ymis*.

2.4. Root Mean Squared Error (RMSE)

The definition of Root Mean Square Error (RMSE) is a measurement method by measuring the difference in the value of the prediction of a model as an estimate of the observed value[21]. The Root Mean Square Error is the result of the square root of the Mean Square Error. The accuracy of the estimation method is characterized by a small RMSE value[21]. The estimation method that has a smaller Root Mean Square Error (RMSE) is said to be more accurate than the estimation method that has a more significant Root Mean Square Error (RMSE). The formula used to calculate the RMSE value is:

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(\hat{y}_i - y_I)^2} \tag{8}$$

where:
$\hat{y}_i$ = Predicted value of the i-th observation
y_I= The actual value of the i-th observation
M = Number of forecasts

2.5. Unsupervised Classification Error (UCE)

Unsupervised Classification Error (UCE) measures how well the imputed dataset is grouped with the complete dataset [0]. The approach used for UCE is Hierarchical Clustering with correlation d=1-Pearson as distance and Ward aggregation. The UCE calculation formula defines as[13]:

$$UCE = \%missclassification\ sample \tag{9}$$

2.6. Supervised Classification Error (SCE)

The Supervised Classification Error (SCE) assesses the imputation method's discriminatory power or predictive power by measuring the difference between the predicted subgroup after the imputation of missing data and the actual subgroup[13]. The approach used for SCE is Linear Discriminant Analysis (LDA) on a set of variables selected a priori in each reference data set without missing values. [0] defines supervised misclassification as:

$$SCE = 1 - AUC \tag{10}$$

with:
Area Under Curve (AUC) is the area under the Receiver Operating Characteristic (ROC) curve of the LDA predictive model.

3. Datasets

In this study, five datasets were used. It aims to determine the performance of each method from various datasets. The datasets used in this study are divided into two groups, namely large datasets and small datasets based on the number of observations.

3.1. SUSENAS KOR MARET 2019 Dataset

The SUSENAS KOR MARET 2019 dataset used in this study has 36 variables with a total row of 23,785 rows owned by this dataset. Most of the questions in this dataset ask about the economic situation in one household. In the SUSENAS KOR MARET 2019 dataset, before the sampling process for the simulation process, this dataset first underwent a manipulation process. The manipulations are related to the columns, which are the answers to the same question. The primary purpose of the manipulation process is to avoid anomalies regarding missing values in the SUSENAS KOR MARET 2019 dataset. In addition, the purpose of the data manipulation stage is to change columns that have no numeric values into numeric values. In the data manipulation stage in this study, the application used is the RStudio application.

Examples of column cases that must go through the data manipulation stage are attached in Table 1. This table contains questions in columns R701_A, R701_B, R701_C, R701_D, and R701_X. This column asks about "what activities were carried out during the past week?". In this column, respondent was given answers in the form of choices. Answer A is for work, answer B is for school, answer C is for taking care of the household, answer D is for doing activities other than personal activities, and answer X is for not doing any activities. So, in this process, the answers 0 and 1 were applied. Answer 0 is given for the line that has the answer X, which means that the respondent does not do any activities in a week. Answer 1 is given for rows that have answers A, B, C, or D. The answer codes 0 and 1 refer to column R408, the valid answers are 0 for no and 1 for yes. The results obtained are as shown in Table 2.

Table 1. Column sample before the manipulation process
Column R701_A, R701_B, R701_C, and R701_D

R701_A	R701_B	R701_C	R701_D	R701_X
			D	
A			D	
		C		
A				
A		C		
	B			

Table 2. Column sample after the manipulation process

R_701
1
1
1
1
1
1

After the data manipulation stage was carried out, the column in the 2019 SUSENAS KOR MARCH 2019 data was reduced from 63 columns to 42 columns. The columns that have been manipulated are columns R701, R807, and R1101.

For research purposes, in the March 2019 SUSENAS KOR dataset, two samples were taken. The first is sampling for a small subset of the dataset, which is 600 observations. Then the second is sampling for a large dataset subset, which is 9000 observations. This sampling followed the average number of SUSENAS household samples in 2017. In the district, the household sample was taken as many as 578 observations, while 8,743 observations were taken in the provinces. In the simulation process, samples that have been taken in the March 2019 SUSENAS KOR dataset were treated as a population.

3.2. Iris dataset

The iris dataset used in this study is the iris dataset introduced by Fisher in 1996 for discriminant analysis. This dataset has four variables with a total of 100 rows. In this study, this dataset is used as one of the small datasets.

3.3. E. Coli dataset

The E. Coli dataset used in this study has five variables with a total of 129 rows. Similar to the iris dataset, the E. Coli dataset is a dataset used for simulating small datasets.

3.4. Dataset Breast Cancer 1

For large datasets, this study uses the Breast Cancer 1 dataset and the Breast Cancer 2 dataset. The Breast Cancer 1 dataset is a dataset that has 65 variables with 80 rows. This dataset was first used in 2012 for taxonomic research on breast cancer.

3.5. Dataset Breast Cancer 2

The Breast Cancer 2 dataset has 60 variables with a total of 89 rows. This dataset was used in 2002 for a Breast Cancer Survival Prediction Study.

4. Experimental Evaluations

In analyzing the dataset, several methods were applied in this study to obtain analytical results following the research objectives. So, several analytical methods are used in this study.

4.1. Data Simulation

In this study, the data analysis phase begins with applying a simulation to the dataset. The initial dataset that is still complete was formed into a dataset that has a missing value. The datasets used in this study are:

- A small subset of the March 2019 SUSENAS KOR data. This dataset is a subset of the March 2019 SUSENAS KOR dataset, which has been sampled by 600 observation units.
- A large subset of the March 2019 SUSENAS KOR data. This dataset is a subset of the March 2019 SUSENAS KOR dataset, which has been sampled by 9000 observation units.
- Iris dataset. This dataset does not undergo a sampling process and is a small dataset.
- E.Coli dataset. This dataset does not undergo a sampling process and is a small dataset.
- Dataset Breast Cancer I. This dataset does not undergo a sampling process and is a large dataset.
- Breast Cancer II dataset. This dataset does not undergo a sampling process and is a large dataset.

The formation of missing values applied all missing value mechanisms, namely Missing Completely at Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR), at ten levels of missing value with an interval of 5% starting from the missing value as much as 5% until the missing value is 50%. After this dataset is formed, the next step is to impute this dataset with the three selected methods, namely Hot-Deck Imputation, KNNI, and PMM. After the imputation process is complete, each method at each level of missing value and each dataset calculated the value of the criteria in the comparative analysis. The criteria values obtained from this simulation process were recorded and inputted into the recording table.

4.2. Comparison Analysis

In this process, the average value of the four criteria used in the comparative analysis was calculated. The criteria used in the comparative analysis in this study are RMSE, UCE, and SCE. The explanation of the three methods is in the previous section. Similar studies, such as [5] and [13], also used this method of evaluation.

Besides RMSE, UCE, and SCE, we also evaluated the performance of each imputation method from the time used to execute the algorithm like conducted by [13]. The time used to execute the algorithm calculates the time used from the start of entering the dataset into the RStudio application until the algorithm is completed. The time used to run the algorithm was calculated manually using a stopwatch from the researcher's smartphone, which will then be recorded in the recording table. The criteria assessed from this aspect are the shorter the time used by a method in carrying out the imputation process. The method is considered to have a practical algorithm.

4.3. Scoring Analysis

The scoring analysis conducted in this study aims to determine the best method for handling missing values in the five datasets. Scoring analysis uses values from the comparison table that has been obtained from the comparative analysis. The score is done by giving a cutting point based on the quartile value of the criteria in each mechanism. The quartiles used in this study were the first quartile (Q1), the second quartile (Q2), and the third quartile (Q3). Each quartile is obtained through calculations carried out with the help of the RStudio application.

After calculating the quartile value, a cutting point was formed, divided into four groups. For the more accessible analysis, the score is denoted by an *. Thus, the more * signs obtained, the better the computational performance of a method. This sign applies to the opposite. The fewer * signs a method gets, the less good its computational performance is. Based on the quartiles, the maximum score obtained by a method is ****, which means that the method has an excellent criterion value. The minimum score

is *, which means that the method has a poor criterion value. The scoring for the attached imputation method is shown in Table 3 below.

Table 3. Score Table

Skor	Range
*	Q_3 – Maximum Score
**	Q_2 - Q_3
***	Q_1 - Q_2
****	0 - Q_1

The scores obtained by the three methods were totaled. The method with the highest score was considered the best method in dealing with missing values. Scoring analysis is also carried out to determine each method's advantages, disadvantages, and characteristics based on the scores obtained.

4.4. Implementation

The stage begins with processing the March 2019 SUSENAS KOR dataset. The SUSENAS dataset at first went through a manipulation process before entering the sampling process. After that, the SUSENAS dataset was sampled as many as 600 and 9000 units of observation. The final results are a small SUSENAS dataset with 600 observations and a large SUSENAS dataset with 9000 observations.

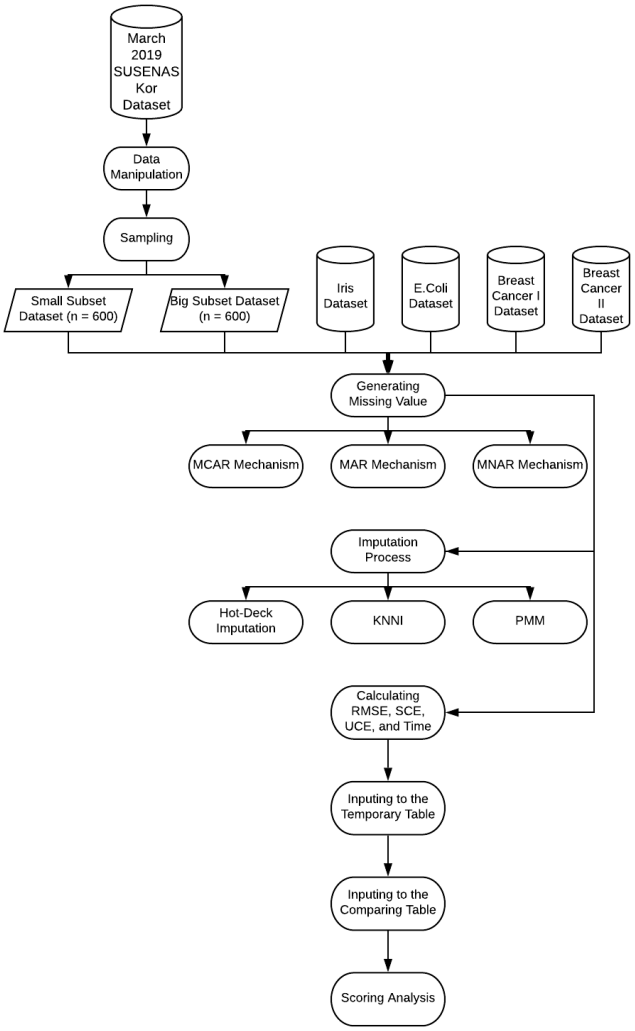

Figure 4. Analysis Process

Furthermore, as shown in Figure 4, forming a dataset with missing values is carried out. In this process, the six datasets are formed with a missing value of 5% until later they end up with a missing value of 50%. The formation of this missing value paid attention to the missing value mechanism. Thus, the results obtained from the missing value generalization process are datasets with missing values in the range of 5% - 50% with the MCAR, MNAR, and MNAR mechanisms.

The dataset that already has this missing value is imputed using three imputation methods, namely Hot-Deck Imputation, KNNI, and PMM. Furthermore, the performance of the three imputation methods was assessed using the RMSE, UCE, SCE, and time values used. The value of the criteria obtained was entered into the recording table and summarized based on the average using a comparison table. The use of comparison tables aims to make the analysis process more straightforward. After the average table of imputation, results are obtained. The last is the scoring process. In this process, the purpose of this research was answered, namely, to get the best method. Scoring is done based on the quantile value of the average table of imputation results.

5. Results

Comparative analysis was conducted to compare the Hot-Deck Imputation, KNNI, and PMM methods. In the analysis process, the values obtained from the simulations on each dataset were averaged and combined into a comparison table as in Table 4, Table 5, and Table 6. These tables contain the values obtained of the comparison criteria determined using the MCAR mechanism, the MAR mechanism, and the MNAR mechanism.

Table 4 is a table of the average results of imputation with the MCAR mechanism. The MCAR mechanism is a situation where missing values occur randomly. So, the user cannot predict in what variable or even in which unit of observation the missing value occurred.

Table 4. Imputation result with MCAR mechanism

Method	Dataset	RMSE	UCE	SCE	Time
Hot-Deck Imputation	1	33.098	2.710	2.222	0.016
	2	33.199	2.789	2.019	0.015
	3	1479681.356	36.246	20.344	0.270
	4	1456234.198	36.330	24.665	0.450
	5	32.249	3.535	2.361	0.019
	6	33.298	3.564	2.431	0.019
KNNI	1	15.233	3.812	2.430	0.033
	2	15.984	3.766	2.653	0.034
	3	1278961.190	27.634	21.444	0.222
	4	926501.660	21.093	23.987	9.360
	5	9.667	3.210	1.798	0.168
	6	9.699	3.246	2.356	0.210
PMM	1	45.120	4.015	3.221	0.008
	2	45.361	4.015	3.199	0.010
	3	1567256.146	29.345	25.667	0.156
	4	1432115.800	30.998	24.109	0.256
	5	39.723	3.850	2.567	0.018
	6	39.500	3.856	2.439	0.020

Next is the missing value mechanism that we often find in our daily life. The MAR mechanism usually occurs due to non-response to a survey. The results of comparing the Hot-Deck Imputation method, the KNNI method, and the PMM method in this mechanism are attached in Table 5 below.

Table 5. Imputation result with MAR mechanism

Method	Dataset	RMSE	UCE	SCE	Time
Hot-Deck Imputation	1	32.910	3.583	2.094	0.016
	2	33.017	3.577	2.142	0.017
	3	1476738.300	27.099	19.868	0.290
	4	1447584.899	27.492	23.531	0.490
	5	32.141	3.512	1.843	0.019
	6	32.290	3.528	2.574	0.019
KNNI	1	14.763	3.620	2.130	0.051
	2	14.709	3.637	2.901	0.054
	3	1257838.932	26.238	20.530	0.250
	4	924856.834	20.935	23.122	9.950
	5	8.132	3.200	1.209	0.199
	6	8.156	3.227	2.356	0.300
PMM	1	44.654	3.905	2.168	0.009
	2	44.78	3.933	2.382	0.014
	3	1562372.820	28.111	25.100	0.210
	4	1428484.378	30.149	23.789	0.277
	5	39.296	3.732	2.943	0.020
	6	39.381	3.755	3.146	0.034

The last mechanism used in the comparative analysis in this study is the MNAR mechanism. This mechanism occurs because there is a standard for a question in the survey. So that when the respondent does not meet the conditions specified for the question, this question was automatically be emptied or missing. For this reason, usually missing values with this mechanism have their patterns and are easy to guess where the data was empty. The imputation results of the Hot-Deck Imputation method, the KNNI method, and the PMM method in dealing with missing values with this mechanism are attached in Table 6 below.

Table 6. Imputation result with MNAR mechanism

Method	Dataset	RMSE	UCE	SCE	Time
Hot-Deck Imputation	1	31.654	3.439	1.846	0.017
	2	31.93	3.457	1.835	0.016
	3	1467939.508	26.260	18.674	0.300
	4	1467890.642	26.601	22.957	0.512
	5	30.980	3.470	1.700	0.021
	6	31.123	3.488	2.328	0.024
KNNI	1	14.783	3.553	1.860	0.065
	2	14.99	3.556	1.875	0.066

Method	Dataset	RMSE	UCE	SCE	Time
	3	1256383.930	24.910	19.983	0.467
	4	918348.493	20.103	22.742	13.792
	5	7.997	3.286	1.079	0.256
	6	8.012	3.289	2.245	0.333
PMM	1	43.880	3.868	2.701	0.010
	2	43.096	3.888	2.771	0.012
	3	1557934.309	27.827	24.742	0.273
	4	1427545.489	28.999	22.999	0.310
	5	38.991	3.694	1.951	0.036
	6	39.091	3.688	2.300	0.045

After obtaining the average table of the results of the imputation of the Hot-Deck Imputation method, the KNNI method, and the PMM method, the next step is to perform a scoring analysis. The scoring analysis was carried out to narrow the results of the comparison obtained from the comparative analysis so that conclusions were obtained regarding the best method in handling missing values. Scoring is done using quantile values from the average table of imputation results in Table 4, Table 5, and Table 6. After scoring, the total score obtained by each method is calculated based on the missing value mechanism. The scoring for each method is based on the assessment attached in Table 7 below.

Table 7. Scoring Table

Criterion	Mechanism	Range	Score
RMSE	MCAR	0 - 3.246	****
		3.247 - 3.961	***
		3.961 - 1190846.000	**
		1190846.001 - 1567256.146	*
	MAR	0 - 3.217	****
		3.218 - 3.933	***
		3.934 - 1174593.000	**
		1174593.001 - 1562372.280	*
	MNAR	0 - 3.101	****
		3.102 - 3.904	***
		3.905 - 1171875.000	**
		1171875.001 - 1557934.309	*
UCE	MCAR	0 - 3.542	****
		3.543 - 3.853	***
		3.854 - 25.998	**
		25.999 - 36.330	*
	MAR	0 - 3.578	****
		3.579 - 3.743	***
		3.744 - 24.912	**
		24.913 - 30.149	*
	MNAR	0 - 3.474	****

Criterion	Mechanism	Range	Score
		3.475 - 3.691	***
		3.692 - 23.708	**
		23.709 - 28.999	*
SCE	MCAR	0 - 2.378	****
		2.379 - 2.610	***
		2.611 - 21.169	**
		21.170 - 25.667	*
	MAR	0 - 2.148	****
		2.149 - 2.737	***
		2.738 - 20.364	**
		20.365 - 25.100	*
	MNAR	0 - 1.863	****
		1.864 - 2.314	***
		2.315 - 19.655	**
		19.655 - 24.742	*
Time	MCAR	0 - 0.018	****
		0.019 -0.033	***
		0.034 - 0.219	**
		0.219 - 9.360	*
	MAR	0 - 0.019	****
		0.020 - 0.052	***
		0.053 - 0.270	**
		0.271 - 9.950	*
	MNAR	0 - 0.021	****
		0.022 - 0.065	***
		0.066 - 0.307	**
		0.308 - 13.792	*

Based on the scoring table that has been formed in Table 7, the scoring results were obtained as shown in Table 8, Table 9, and Table 10 below.

Table 8. Scoring result in MNAR mechanism

Method	Dataset	RMSE	UCE	SCE	Time	Total	Total Score
Hot-Deck Imputation	1	***	****	****	****	15	61
	2	***	****	****	****	15	
	3	*	*	*	*	4	
	4	*	*	*	*	4	
	5	***	****	**	***	12	
	6	***	***	**	***	11	
KNNI	1	****	**	***	**	11	59
	2	****	***	**	**	11	

Method	Dataset	RMSE	UCE	SCE	Time	Total	Total Score
	3	*	*	*	***	7	
	4	**	*	*	*	5	
	5	*	****	****	****	13	
	6	*	****	****	***	12	
PMM	1	**	**	**	****	10	52
	2	**	**	**	****	10	
	3	*	*	*	**	5	
	4	*	*	*	*	4	
	5	***	**	***	****	12	
	6	***	**	***	***	11	

In the MCAR mechanism, the Hot-Deck Imputation method is the method that has the highest score among the other two methods, with a total score of 61. The hot-Deck Imputation method got 2 points higher than the KNNI method and 9 points higher than the PMM method. With this score, the Hot-Deck Imputation method has better imputation to handle missing values with the MCAR mechanism compared to the KNNI and PMM methods.

Table 9. Scoring result in MAR mechanism

Method	Dataset	RMSE	UCE	SCE	Time	Total	Total Score
Hot-Deck Imputation	1	***	***	****	****	14	65
	2	***	****	****	****	15	
	3	*	*	*	*	4	
	4	*	*	*	*	4	
	5	****	****	****	***	15	
	6	***	****	***	***	13	
KNNI	1	****	***	****	**	13	62
	2	****	***	**	**	11	
	3	*	*	*	*	4	
	4	**	*	*	*	5	
	5	****	****	****	***	15	
	6	****	****	***	***	14	
PMM	1	**	**	***	****	11	58
	2	**	**	***	****	11	
	3	****	*	*	**	8	
	4	****	*	*	***	9	
	5	***	***	**	***	11	
	6	**	**	**	**	8	

Based on Table 9, it can be concluded that the Hot-Deck method is the best in dealing with missing values with the MAR mechanism. The points obtained by the Hot-Deck Imputation method are above the KNNI method and the PMM method. The Hot-Deck Imputation's point is 3 points higher than the points obtained by KNNI and 7 points higher than PMM's points.

Table 10. Scoring result in MNAR mechanism

Method	Dataset	RMSE	UCE	SCE	Time	Total	Total Score
Hot-Deck Imputation	1	***	****	****	****	15	65
	2	***	****	****	****	15	
	3	*	*	*	**	5	
	4	*	*	*	*	4	
	5	****	****	****	****	16	
	6	**	***	**	***	10	
KNNI	1	****	***	****	**	11	52
	2	****	***	***	**	12	
	3	*	*	*	*	4	
	4	**	*	*	*	5	
	5	*	****	****	**	11	
	6	*	****	***	*	9	
PMM	1	**	**	**	****	10	52
	2	**	**	**	****	10	
	3	*	*	*	***	6	
	4	*	*	*	*	4	
	5	***	**	***	***	11	
	6	**	***	***	***	11	

With the MNAR mechanism, the Hot-Deck Imputation method is superior to the KNNI and PMM methods. This method gets 65 points which is the highest point in the MNAR mechanism. The Hot-Deck Imputation is far above the PMM and KNNI methods, which get the same 52 points.

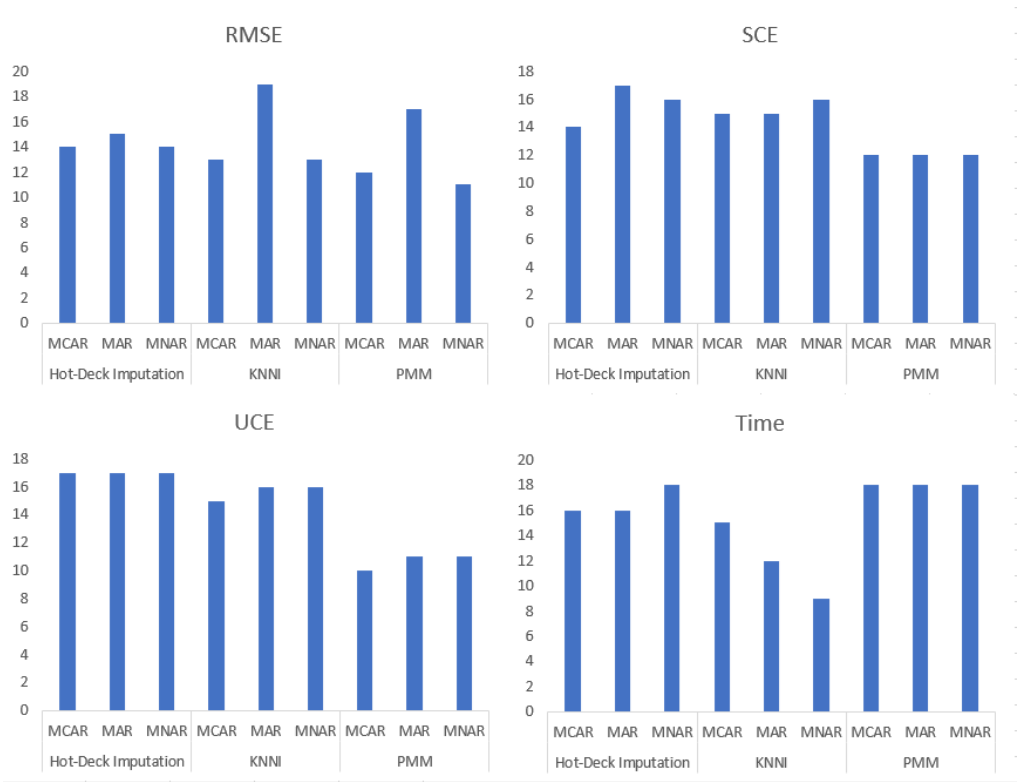

Figure 5. Score comparison for each criterion

Based on Figure 5, it can be concluded that the Hot-Deck Imputation method is the best in dealing with missing values with the MCAR, MAR, and MNAR mechanisms. Based on the scoring results, it can be concluded that Hot-Deck Imputation is the best method to handle missing values. Although overall, this method gets relatively high points, if it is examined again, it is relatively weak in dealing with missing values in the official statistics dataset. As shown in Table 8, Table 9, and Table 10, the Hot-Deck Imputation method gets very minimal points in dataset 3 and dataset 4, which are simulations for the official statistics dataset, namely the March 2019 SUSENAS KOR dataset. The imputation method would be better applied to normally distributed data [5]. While the raw data used in dataset 3 and dataset 4 do not meet this assumption. Therefore, there is a possibility that if the data is transformed into standard data points that will be obtained by the Hot-Deck Imputation method in the scoring analysis will be better. In addition, the number of variables also affects the performance of the Hot-Deck Imputation method[14]. Hot-Deck Imputation is more suitable when applied to datasets that have many variables. It can be proven in dataset 5 with the MNAR mechanism. The Hot-Deck Imputation method gets a perfect score of 16. In dataset 5, the Breast Cancer I dataset is used, which has 65 variables, which is the dataset with the most variables in this study. So, it can be concluded that the Hot-Deck Imputation method has better computational performance because the total points obtained in the scoring analysis are always at the top. In addition, the Hot-Deck Imputation method will be more effective in running on datasets with many variables. However, this method is relatively weak in the consistency of the estimator based on the RMSE value of this method which is relatively low compared to the other two methods.

Regarding the estimator's accuracy, the KNNI method has better accuracy than the Hot-Deck Imputation method and the PMM method based on the RMSE value of the KNNI method, which is quite good. The RMSE value of the KNNI method can be even higher if there is a data standardization process in the six datasets. One of the main factors that must be considered in the KNNI method is the data unit. So the use of raw data in this study will undoubtedly affect the performance of the KNNI method[12].

In contrast to the Hot-Deck Imputation method, which is quite good at handling datasets with large numbers and many variables, the KNNI method is not well applied in big datasets, especially dataset with many variables. KNNI method has decreased computational performance, based on the results of scoring in dataset 4, dataset 5, and dataset 6. In addition, the time required to run the algorithm on the KNNI method is quite time-consuming compared to the Hot-Deck Imputation method and the PMM method. As can be seen from the results of the KNNI's average time scoring is only in the range of 1 to 3. The use of quite a long time in the KNNI method is driven by the algorithm owned by the KNNI method is quite long, so the time required to execute the algorithm method The KNNI has also become longer. So that this KNNI method would be better if used for data with a small unit of observation and few variables.

The problem of using time in the KNNI method does not apply to the PMM method. The PMM method has a good performance in the time criterion. The scoring results show that this method has a reasonably short time to execute the algorithm influenced by the algorithm owned by the PMM method is quite simple, making the PMM method one of the most frequently used imputation methods. The PMM method works better on non-normally distributed data compared to datasets with normal distribution[19]. The scoring analysis shows that the PMM method has higher points than the Hot-Deck Imputation method and the KNNI method in dataset 3 and dataset 4. From the points obtained by the PMM method. In that case, the PMM method works better when handling the March 2019 SUSENAS KOR dataset, which is official statistical data that does not meet the normality assumption. So the PMM method can be the best choice compared to the Hot-Deck Imputation and method.

6. Discussion and Conclusion

Based on the results and discussion in the previous chapter, it can be concluded from this study that, in general, the Hot-Deck Imputation method is the best in dealing with missing values. This method is very good at handling missing values in datasets with many variables but less able to handle missing values in small datasets. The shortcomings of the Hot-Deck Imputation method do not apply to the KNNI method, which is very good at handling small datasets. However, the KNNI method is not suitable in terms of time. The algorithm is quite complicated, making this method take quite a lot of time in its

execution. The time problem experienced by the KNNI method does not apply to the PMM method, which has a short time in imputing missing values in the data.

The three methods are equally not good at handling missing values in SUSENAS data. The decrease in performance was experienced by the three methods in the third and fourth simulations, which are simulations using SUSENAS data. The decrease in performance can be caused by the SUSENAS data used in the simulation is raw data so that the data is not normally distributed. It is possible that if normalization is carried out, the imputation results can be better.

Based on the results and conclusions obtained from this research, some suggestions that researchers can give for further research are to pay attention to the pattern of missing values and develop a better scoring method.

References

[1] J. Han, M. Kamber, dan J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[2] R. E. A. Joseph F. Hair Jr, William C. Black, Barry J.Babin, "Multivariate Data Analysis.pdf." hal. 758, 2010.

[3] T. Hendrawati, "Kajian Metode Imputasi Dalam Menangani Missing Data," *Pros. Semin. Nas. Stat. | Dapertemen Stat. FMIPA Univ. Padjadjaran*, hal. 637–642, 2015.

[4] S. Van Buuren, *Flexible Imputation of Missing Data*, Second Edi. London, New York: CRC Press, 2018.

[5] I. J. Fadillah dan S. Muchlisoh, "Perbandingan Metode Hot-Deck Imputation Dan Metode Knni Dalam Mengatasi Missing Values," *Semin. Nas. Off. Stat.*, vol. 2019, no. 1, hal. 275–285, 2020, doi: 10.34123/semnasoffstat.v2019i1.101.

[6] J. Luengo, S. García, dan F. Herrera, "A study on the use of imputation methods for experimentation with Radial Basis Function Network classifiers handling missing attribute values: The good synergy between RBFNs and EventCovering method," *Neural Networks*, vol. 23, no. 3, hal. 406–418, 2010, doi: 10.1016/j.neunet.2009.11.014.

[7] C. Chatfield, R. J. A. Little, dan D. B. Rubin, *Statistical Analysis with Missing Data.*, vol. 151, no. 2. 1988.

[8] H. Migdady dan M. Al-Talib, "An enhanced fuzzy K-means clustering with application to missing data imputation," *Electron. J. Appl. Stat. Anal.*, vol. 11, no. 2, hal. 674–686, 2018, doi: 10.1285/i20705948v11n2p674.

[9] M. R. Stavseth, T. Clausen, dan J. Røislien, "How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data," *SAGE Open Med.*, vol. 7, hal. 205031211882291, 2019, doi: 10.1177/2050312118822912.

[10] J. M. Jerez et al., "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artif. Intell. Med.*, vol. 50, no. 2, hal. 105–115, 2010, doi: 10.1016/j.artmed.2010.05.002.

[11] R. R. Andridge dan R. J. A. Little, "A review of hot deck imputation for survey non-response," *Int. Stat. Rev.*, vol. 78, no. 1, hal. 40–64, 2010, doi: 10.1111/j.1751-5823.2010.00103.x.

[12] M. Mukarromah, S. Martha, dan I. Ilhamsyah, "Perbandingan Imputasi Missing Data Menggunakan Metode Mean Dan Metode Algoritma K-Means," *Bimaster*, vol. 4, no. 3, hal. 305–312, 2015, [Daring]. Tersedia pada: http://jurnal.untan.ac.id/index.php/jbmstr/article/view/12425/.

[13] S. P. Mandel J, "A Comparison of Six Methods for Missing Data Imputation," *J. Biom. Biostat.*, vol. 06, no. 01, hal. 1–6, 2015, doi: 10.4172/2155-6180.1000224.

[14] D. W. Joenssen dan U. Bankhofer, "Hot deck methods for imputing missing data: The effects of limiting donor usage," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics*), vol. 7376 LNAI, hal. 63–75, 2012, doi: 10.1007/978-3-642-31537-4_6.

[15] A. Kowarik dan M. Templ, "Imputation with the R package VIM," *J. Stat. Softw.*, vol. 74, no. 7, 2016, doi: 10.18637/jss.v074.i07.

[16] M. Kuhn dan K. Johnson, *Applied Predictive Modeling*. Springer New York, 2013.

[17] O. Troyanskaya et al., "Missing value estimation methods for DNA microarrays," *Bioinformatics*,

vol. 17, no. 6, hal. 520–525, 2001, doi: 10.1093/bioinformatics/17.6.520.
[18] W. Wilsen, W. Rahayu, dan V. M. Santi, "Penerapan Imputasi Ganda dengan Metode Predictive Mean Matching (PMM) untuk Pendugaan Data Hilang pada Model Regresi Linear," *J. Stat. dan Apl.*, vol. 2, no. 1, hal. 12–20, 2018, doi: 10.21009/jsa.02102.
[19] H. Anisa, Ladpoje, "Imputasi Ganda Dengan Metode Predictive Mean Matching Untuk Estimasi Data Hilang pada Item Nonrespon," *Digilib.Unhas.Ac.Id*, no. Matematika, hal. 68–70, 2015.
[20] N. Malahayati, "PERBANDINGAN METODE IMPUTASI GANDA: METODE REGRESI LINEAR VERSUS METODE PREDICTIVE MEAN MATCHING UNTUK MENGATASI DATA HILANG PADA DATA SURVEI," Institut Pertanian Bogor, 2008.
[21] A. von Eye dan C. Schuster, *Outlier Analysis*. 1998.