Ingénierie des Systèmes d'Information
Vol. 28, No. 2, April, 2023, pp. 451-459
Journal homepage: http://iieta.org/journals/isi

Single Imputation Using Statistics-Based and K Nearest Neighbor Methods for Numerical Datasets

Abdul Fadlil[1], Herman[2], Dikky Praseptian M[2*]

[1] Department of Electrical Engineering, Industrial Technology, Universitas Ahmad Dahlan, Yogyakarta 55166, Indonesia
[2] Master Program of Informatics, Industrial Technology, Universitas Ahmad Dahlan, Yogyakarta 55166, Indonesia

Corresponding Author Email: dikky2107048008@webmail.uad.ac.id

https://doi.org/10.18280/isi.280221

Received: 6 December 2022
Accepted: 26 March 2023

Keywords:
imputation, kNNI, missing value, numerical dataset, statistic-based

ABSTRACT

Handling missing values is often an unavoidable problem. Imputation is a preferred option in handling missing values compared to removing all row records which will reduce the number of datasets and can lead to poor research results if the size of the remaining data is too small. The problem that often occurs is that there are often wrong conclusions due to some records that have missing values, therefore this study will test several simple imputation methods, namely statistical-based imputation and kNNI. The results of testing the error value with RMSE and MAPE show that kNNI imputation results are much better than statistical-based imputation. Based on the standard used in the MAPE test, the kNNI test results (error values) are almost entirely very good because the error value is <10% except for three test results in dataset 1 at k=10, k=15 and k=20, while the statistical-based imputation results are only good because the error value is between 10% and 20%, even one of the results exceeds 20% Although kNNI is better than statistical-based imputation, it is necessary to choose the right k value to get the best imputation results.

1. INTRODUCTION

Missing data or information is inevitable due to various reasons such as damage to storage equipment, failed pixels, limited data capacity, acquisition equipment, missed questions in surveys and so on [1]. Data quality is a major concern for data scientists and researchers working in the field of data analysis science [2]. Most statistical algorithms and machine learning are not powerful enough to handle missing values. The missing data will cause an element of ambiguity when analyzing the data and that can affect the nature of the statistical estimator and result in loss of strength and misdirection in the conclusion so that the data is unreliable [3, 4]. Three types of problems associated with missing values in data mining loss efficiency, complications in handling and analyzing data, and bias resulting from the difference between missing data and complete data [5, 6]. Handling missing values appropriately is an important and challenging task because it requires careful examination of all training data to understand and be able to identify patterns of missing values in the data as well as a clear understanding of different imputation techniques. This study uses the single imputation missing value pattern meaning that there is only one attribute from one record that has a missing value. The most common and widely used method, namely statistical-based imputation, namely mean, median imputation, mode imputation and machine learning model imputation with the k Nearest Neighbor imputation method because it is considered the simplest and fastest.

Handling missing value is often an unavoidable problem. Imputation is the preferred choice in handling missing values as opposed to eliminating the entire row of records which will reduce the number of datasets and can lead to poor research results if the remaining data size is too small [7, 8]. The graduate user satisfaction level survey is a way conducted by universities in Indonesia to assess the quality of universities in terms of aspects of graduate user satisfaction. The results of the survey are used as material for evaluating and improving the quality of the universities surveyed. The survey is carried out to the institution/agency/company where graduates from the college work. Filling out the survey is carried out by direct superiors in the field of work or division where graduates work, this is done so that the assessment of graduates is more objective. This research used user satisfaction survey data for graduates of STMIK PPKIA Tarakanita Rahmawati, an informatics university in Indonesia. The data collected is the level of satisfaction for 100 graduates from 2017 to 2021 which is hereinafter referred to as the research dataset. Each entity of the dataset which is hereinafter referred to as a data point represents one graduate who has seven attributes. Each attribute or criterion represents aspects that are assessed to show the level of user satisfaction with graduates, namely C1=Ethics, C2=Main Competencies, C3=Foreign Language Ability, C4=Use of Information Technology, C5=Communication Ability, A6=Cooperation, and C7=Self-Development. The value for each attribute uses a likert scale with a scale 1-5 [9]. The details are 1=Unsatisfied, 2=Unsatisfied, 3=Moderately Satisfied, 4=Satisfied, and 5=Very Satisfied. Table 1 shows a snippet of graduate user satisfaction level datasets.

The level of user satisfaction for a graduate will be obtained by equalizing the value of all attributes of the graduate. Unfortunately, it often happens that there is a missing attribute value that is referred to as a missing value. This missing value can occur for several reasons but most often cannot judge it because the aspect in question is not used in the field of

graduate work. Missing value can lead to inferences of inaccurate grouping results, so there is a need for a way to overcome this. The method that the author chooses is to predict the missing value instead of eliminating the record row with the missing value which will result in a lack of data in the dataset. This study will test simple imputation predictions performed with statistical approaches (mean, median, mode) and k-Nearest Neighbor Imputation. This study is also limited to cases where there is only one missing value at one data point (single imputation). Previous research has also shown a lack of variation from datasets in testing missing value treatment methods and the absence of previous studies that directly compare the three statistically based imputation methods and kNN Imputation. The four methods are the methods that the author considers the simplest in handling missing value numeric data.

This research is structured as follows: In part 2 there is a review of the literature containing previous works and related research. The basic principles for testing several imputation methods as well as an explanation of the methods used in section 3. Section 4 contains an explanation of the dataset to be used and the techniques for using it. The results of the experiment are discussed in section 5. Finally, part 6 gives a conclusion [10].

The results of testing error values with RMSE and MAPE showed all error calculation results show that kNNI imputation results are much better than statistical-based imputations. Although kNNI is better than statistical-based imputation, it needs to choose the right k value to get the best imputation results.

Table 1. Dataset

Alumni (A)	C1	C2	C3	C4	C5	C6	C7
A1	5	5	4	5	5	5	5
A2	4	4	4	5	4	4	4
A3	4	4	4	4	4	4	4
A4	4	4	3	4	5	4	4
A5	5	4	4	4	4	4	4
A6	4	4	3	4	4	4	3
A7	4	4	4	5	4	4	4
A8	5	5	4	5	5	5	5
A9	5	5	3	5	5	5	5
A10	4	4	3	4	5	5	4
..	..	..	..	..	..	..	..
A91	5	5	5	5	5	5	5
A92	4	4	3	4	4	4	3
A93	5	5	4	4	5	5	5
A94	4	4	3	4	4	4	4
A95	4	4	3	2	3	3	4
A96	5	4	3	3	5	5	5
A97	4	4	3	4	3	3	3
A98	5	5	3	5	5	5	4
A99	5	5	3	5	5	5	5
A100	5	4	2	3	4	4	4

2. RELATED WORKS

Missing value is a problem in most scientific studies such as Medical [11, 12] this study examines the impact of missing value imputation on the classification of clinical trial data and breast cancer, Troyanskaya et al. [13] tested the classification accuracy of DNA data with and without kNN imputation, or Climate Science [14], this research is more about testing the missing value imputation approach to data climate as a data trial because the data climate has large numbers and variations even traffic data [10, 15], Noyunsan et al. [16], Choudhury, and Pal [17] resulted in the application of missing value imputation in data traffic can reduce the risk of accidents. Several studies related to the use and comparison of imputation methods to fill the missing value have also been carried out several times, both studies examined the impact of missing value imputation on the classification process, where both studies resulted in an increase in the accuracy of the classification process after imputation with several methods.

Research conducted by Acuna and Rodriquez [18] and Dixon [19] shows almost the same result under the impact of four methods in handling missing value. These methods are recording deletion, and three imputation methods: Mean imputation, median imputation, and k-nearest neighbor imputation (kNNI). Classification is carried out using two methods: linear discriminant analysis (LDA) and kNN. Their results showed that imputation had a less significant impact on classification accuracy. The weakness in these two studies is testing on datasets with missing values below 20%.

Research conducted by Batista and Monard [20] has tested the classification accuracy of two classification methods, namely C4.5 [21] and CN2 [22], and three imputation methods, namely mean imputation, imputation mode, and kNNI. The missing data is entered only in a few selected attributes. The results show that kNN imputation produces good accuracy, but only when the attributes are not highly correlated with each other. Over time from this study the distribution of missing values varies less in each attribute.

Research conducted by Grzymala-Busse and Hu [23] look at the classification accuracy of datasets that have missing values in ten datasets using five imputation methods (Imputation mode, C4.5, LERS and two non-traditional imputation methods) and non-traditional imputation machine learning methods) and rough set theory. The results showed that the imputation of missing data before classification is useful for improving the accuracy of classification results. The weakness of this study is that only one examined only one classifier and the missing value rate was small (1%-13%).

Research conducted by Mundfrom and Whitcomb [24] used two classifiers (linear discriminant function and logistic regression) to test three imputation methods (Mean, Hot deck and regression imputation) and the results showed that the mean imputation method achieved the best performance. However, this method only makes comparisons on one dataset. Therefore, the conclusion is not very convincing.

3. THE PROPOSED METHOD

3.1 Principle of the analysis

The analysis was carried out on four simple imputation methods that are most often used, namely mean imputation, median imputation and mode imputation which represent a statistic-based imputation method and k Nearest Neighbor Imputation (kNNI) from machine learning-based methods. The imputation method will test three datasets. Three datasets come from a single dataset by having the entire full attribute value without missing value. Dataset 1 is set to have a missing value on the final 10 records and on one of the attributes randomly. Dataset 2 is set to have a missing value on the final 20 records and on one of the attributes randomly. Dataset 3 is set to have a missing value on 30 records and on one of the

attributes randomly. Giving a missing value to only one of the attributes in each record shows that analysis is carried out for single imputations, according to conditions that often occur in the dataset. The following dataset can be seen in Figure 1.

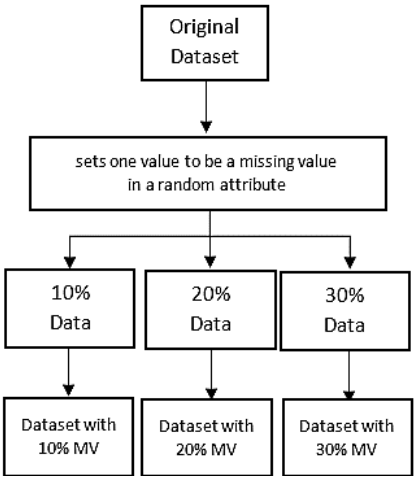

Figure 1. Dataset preparation

Orginal dataset from user satisfaction survey data for graduates. The data collected is the level of satisfaction for 100 graduates from 2017 to 2021. Each entity of the dataset which is hereinafter referred to as a data point represents one graduate who has seven attributes. Each attribute or criterion represents aspects that are assessed to show the level of user satisfaction with graduates, namely C1=Ethics, C2=Main Competencies, C3=Foreign Language Ability, C4=Use of Information Technology, C5=Communication Ability, A6=Cooperation, and C7=Self-Development. The value for each attribute uses a likert scale with a scale 1-5 [9]. The details are 1=Unsatisfied, 2=Unsatisfied, 3=Moderately Satisfied, 4=Satisfied, and 5=Very Satisfied.

Three datasets will be tested with four imputation methods, namely three statistical-based imputations and the kNNI method. The results of the imputation prediction were eight, namely three from statistical-based imputations (mean, median and mode) and five from kNNI (k=1, k=5, k=10, k=15, and k=20). Eight imputation prediction results will be measured by comparing with the complete dataset by measuring the error rate. The following development model is seen in Figure 2.

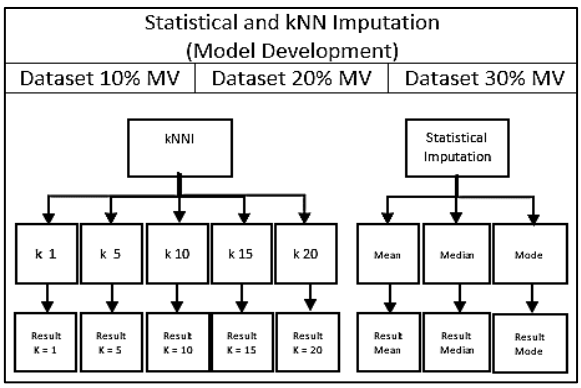

Figure 2. Model development

Error rate measurement uses two methods, namely Root Mean Squared Error (RMSE) to measure the entire error value in each dataset and Mean Absolute Percentage Error (MAPE) is used to measure the ratio of error values on each record that has a missing value then on average to find out the error value in each dataset. The following Model Evaluation can be seen in Figure 3.

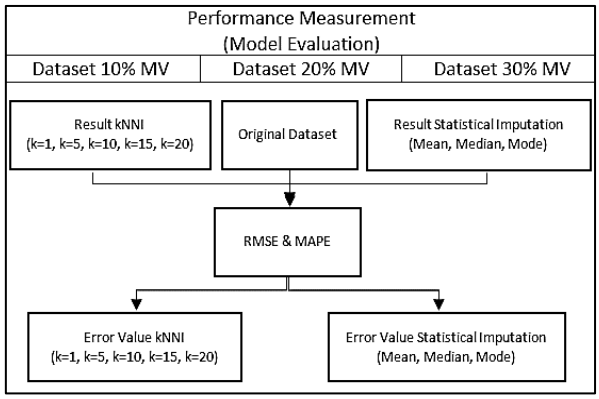

Figure 3. Model evaluation

The smallest error value will be selected as the best method in this study, both overall and in each group of simple imputation methods / statistical based imputation methods and k selection in kNNI. The impact of the number of missing values in each dataset on the error value will also be analyzed how much it affects. The magnitude of the difference between simple imputation methods and machine learning, which in this case is represented by kNNI, will also be analyzed. All imputation prediction results are processed using the scikit-learn library. The following best model selection is seen in Figure 4.

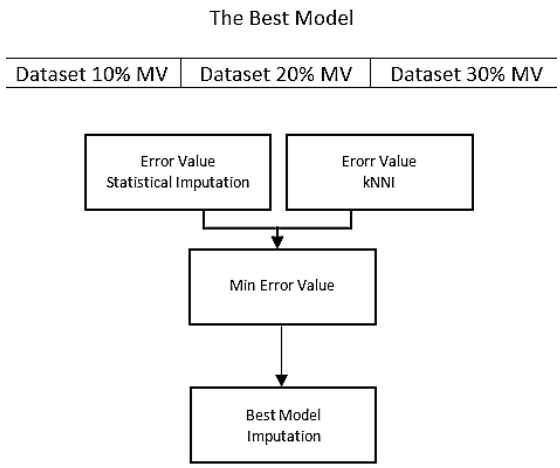

Figure 4. Best model choice

3.2 Statistical imputation

The statistical-based imputation method has a quick approach to missing values by replacing them with mean, median, mode value [25]. Mean Imputation, this imputation method uses an average statistical approach [1]. The average taken is the average of the attributes that have a missing value in the testing data against all training data. This method is very simple and fast but only works on numerical data. If there is so much data variation, it can be ascertained that this imputation method is not effective because this method uses all the values in the training data.

The median imputation is almost the same as the mean imputation using a statistical approach. The imputation median sorts all values in the training data in the missing value in the testing data and then looks for the middle value [26]. The median value is the result of the imputation prediction. This method is also very simple and fast but only effective on numerical data. Both median imputation and mean imputation algorithms are supported by SimpleImputer function in scikit-learn library.

The mode imputation or often called the "most frequent" imputation [23] is another imputation method available in SimpleImputer function of the scikit-learn library. The most frequent imputation predicts value of imputation by looking for the value that most often appears in the training data in the attribute that has a missing value in the testing data. This method is also relatively simple and fast. In contrast to the mean and median, mode imputation can be used on the entire dataset not only numeric datasets.

3.3 k-Nearest Neighbor (kNN)

k-Nearest Neighbor (kNN) is typical method in machine learning for classification. The basic principle starts from NN or 1-NN, this concept exists because of the large amount of data that causes a lack of classification accuracy. The 1-NN process is to calculate the proximity of the distance between the testing data and all training data with the distance calculation equation as commonly used by euclidean distantace, manhatan distance or others, then the class of one training data with the closest distance is the testing data class [27]. Here are the most popular distance calculations used because of their simplicity, namely euclidean distance.

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

where, d_E=euclidean distance; x=testing data, y=training Value, i=Number of atribute, d=Max of atribute.

The idea of k-Nearest Neighbor (kNN) is to group data into groups that have similar properties to it [27]. This classification algorithm is also called a lazy algorithm because it does not learn how to categorize data, but only remembers existing data. kNN searches for k feature vectors with similar properties, then groups the new data into that group of feature vectors.

3.4 k-Nearest Neighbor Imputation (kNNI)

k-Nearest Neighbor Imputation (kNNI) is a variance of kNN method. It uses the basic concept of kNN in calculating the proximity of the distance between testing data that has missing value and training data that has complete data [28]. The only difference is that attributes that have a missing value in the testing data are not counted in the distance calculation. The distance calculation used in this study is euclidian distance can be seen in equation 1. The result of the distance calculation will be sorted from the smallest. This study uses five k values for each dataset. Dataset 1, dataset 2 and dataset 3 are gone through kNN with k=1, k=5, k=10, k=15 and k=20.

This study uses numerical datasets so that the approach used mean / average for results with a value of k more than 1 [1]. Examples such as k=5 mean that the imputation value is the average of 5 attribute values (attributes that have missing values in the testing data) in the average training data. Similarly, k=1, k=10, k=15, and k=20.

3.5 Performance measurement

Performance measurement is done by measuring the error rate by using two methods Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). RMSE is one of the methods for evaluating forecasting techniques or measuring the accuracy of a model's forecasting results. RMSE expresses the average value of the sum of squares of a forecast model. The small RMSE value gives a clue that the variation (diversity) of the values produced by the forecast model is close to the variation in its observation value. One of the measures of error in forecasting is the middle value of the square root or Root Mean Square Error (RMSE), here is the RMSE equation [29].

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\mathrm{A}_t - \mathrm{F}_t)^2}{n}} \quad (2)$$

where, *RMSE*=Root Mean Square Error, n=Number of Samples, A_t=Actual Value, F_t=Prediction Value.

Mean Absolute Percentage Error (MAPE) is a percentage measure of the error of the predicted result. The smaller the MAPE value, the smaller the error of the prediction result, on the contrary, the greater the MAPE value, the greater the error of the prediction result. The imputation result is very good if the MAPE value is <10%, while the imputation result is good if the MAPE value is between 10% and 20%. MAPE can be calculated by the following formula [30].

$$MAPE = \frac{\sum_{i=1}^{n}\left|\frac{X_i - F_i}{X_i}\right|}{n} 100\% \quad (3)$$

where, X_i=actual data, F_i=predicted result, n=Number of Samples.

4. DATASETS

The graduate user satisfaction rate dataset has seven attributes with a likert scale of 1-5. The original dataset contains 100 data points with all of its attributes having values or not missing values. This original dataset is duplicated into three datasets. Furthermore, each dataset is modified so that it has a missing value with a different pattern. For Dataset 1, out of 100 data points modified so that 10 data points have missing values while the other 90 data points are complete. Successively, Dataset 2 has 20 data points with missing value and Dataset 3 has 30 data points with missing value. Because this study only focuses on single imputation, modifications to data points that have missing values are carried out by removing the value of only one attribute. The determination of which data point in the dataset will have a missing value and which attribute of the data point whose value will be missing is done randomly. This research will develop a model to determine the value of each missing value. Three models were developed with statistical methods (mean, median, mode) and one model with machine learning methods using kNN imputation with five different values of k. The development of each model used the three datasets in order to find the best

model of missing value imputer especially for the case of single imputation. Details of the dataset to be used in this study can be seen in Table 2.

Table 2. Datasets

Datasets	Data Point With MV	Data Point Without MV	Total Data Point	Percentage of MV
Dataset 1	10	90	100	10%
Dataset 2	20	80	100	20%
Dataset 3	30	70	100	30%

5. EXPERIMENT RESULT AND ANALYSIS

Duplication of datasets into three is carried out to test the accuracy of the prediction results of each dataset against the initial dataset whose entire record does not have a missing value. Each dataset produces eight imputation prediction results, three result from statistical approaches (mean, median, mode) and five result from kNNI with five different k values. The results of imputation prediction will compare the three datasets that have different percentages of missing values, in the statistical approach will also be compared the results of the mean, median and mode on the three datasets, in kNN imputation will also be compared which five k values have the best accuracy in the three datasets, and finally compare all imputation prediction results in both ways, namely the statistical approach and the imputation kNN on the entire dataset.

5.1 Statistical-based imputation result

The prediction of missing value in this research uses scikit-learn library and python programming language. The statistical approach prediction employs SimpleImputer function of the library. Such function supports three statistical-based imputation methods, namely mean, median and mode. Figure 5 shows prediction missing value with mean method. The first step declares numpy library into a variable, numpy is used to hold the numeric array value of the dataset to be used. The second step creates a Simple imputer function for mean imputation. The third step identifies the missing value used with "np.nan" and the methods or strategies used such as mean, median and mode. The fourth step inserts the dataset into an array variable. The final step prints the imputation prediction results with predefined methods and datasets.

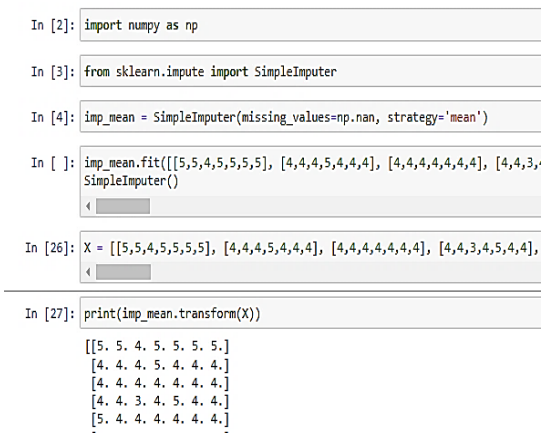

Figure 5. Simple imputer (Scikit-Learn)

The results of statistical-based imputation on the three datasets can be seen in Table 3, Table 4 and Table 5. The results shown are partial due to writing limitations.

Table 3. Statistical-based imputation on dataset 1

Alumni	Attribute	Ori value	Mean	Median	Mode
A77	C1	5	4.58	5	5
A47	C2	4	4.47	4	4
A54	C3	4	3.54	3	3
A41	C4	4	4.29	4	4
A3	C5	3	4.46	4	5
A99	C6	5	4.47	5	5
A1	C7	3	4.33	4	4
A100	C6	5	4.47	5	5
A78	C5	5	4.46	4	5
A42	C4	3	4.29	4	4

Table 4. Statistical-based imputation on dataset 2

Alumni	Attribute	Ori value	Mean	Median	Mode
A41	C3	3	3.54	3	3
A3	C4	4	4.28	4	4
A90	C5	5	4.45	4	4
A97	C6	4	4.47	5	5
A42	C7	4	4.34	4	4
..	..	..	,,	..	..
A4	C5	3	4.45	4	4
A99	C6	5	4.47	5	5
A5	C7	3	4.34	4	4
A100	C6	5	4.47	5	5
A96	C5	5	4.45	4	4
A46	C4	3	4.28	4	4

Table 5. Statistical-based imputation on dataset 3

Alumni	Attribute	Ori value	Mean	Median	Mode
A33	C5	4	4.46	4	5
A95	C6	4	4.48	5	5
A34	C7	4	4.34	4	4
A96	C6	4	4.48	5	5
A35	C5	4	4.46	4	5
..	..	..	..	..	..
A37	C1	4	4.58	5	5
A38	C2	5	4.47	4	4
A39	C3	3	3.52	3	3
A3	C4	4	4.28	4	4
A88	C5	5	4.46	4	5
..	..	..	..	..	..
A4	C5	3	4.46	4	5
A99	C6	5	4.48	5	5
A5	C7	3	4.34	4	4
A100	C6	5	4.48	5	5
A94	C5	5	4.46	4	5
A44	C4	3	4.28	4	4

The mean, median and mode imputation results of dataset 1 are seen in Table 3, dataset 2 is seen in Table 4, and dataset 3 is seen in Table 5. The mean imputation result is in fractional number because it uses the concept of averaging all training data while the median concept is to sort all values in the data then look for the middle value, and the simplest mode is to find the value that most often appears from the data. In order to find the most accurate prediction of the missing value this research to uses two error value tests, namely Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). The smallest error value is the best result obtained. The following are the results of testing error values with RMSE and MAPE

by comparing the mean, median and mode imputation results against the original values in the three datasets, can be seen in Table 6. Figure 6 and Figure 7 exhibit graphical comparisons of the error test using RMSE and MAPE respectively.

Table 6. Statistical-based imputation

Method (Statistic-Based)	RMSE			MAPE		
	Dataset 1	Dataset 2	Dataset 3	Dataset 1	Dataset 2	Dataset 3
Mean	0.84	0.69	0.68	20.71%	16.14%	14.99%
Median	0.71	0.71	0.80	14.50%	12.75%	13.67%
Mode	0.84	0.71	0.84	15.83%	12.75%	14.44%

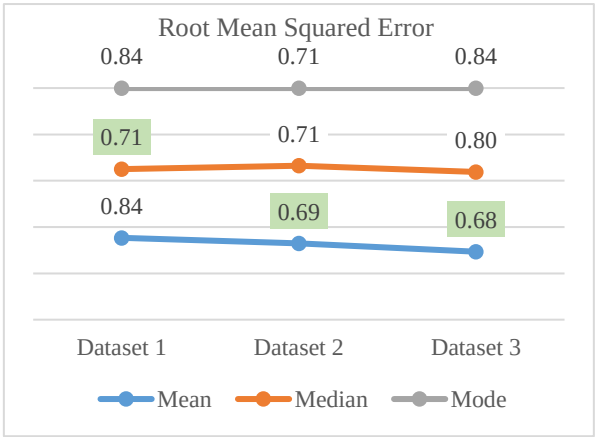

Figure 6. RMSE to imputation statistic-based

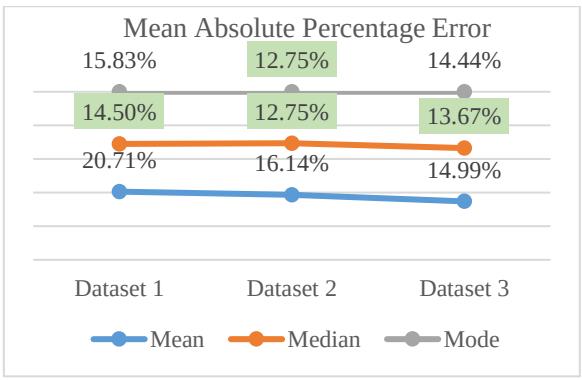

Figure 7. MAPE to imputation statistic-based

Testing with RMSE showed that dataset 1 median has the lowest error value, in dataset 2 mean is superior to median and mode while in dataset 3 mean is again superior to median and mode. However, if we look at the error value, it is more stable in the median with a difference that is not too far in each dataset while the mean although superior in two datasets, the error value is very large for dataset 1, where dataset 1 is the dataset with the least missing value and the most training data. It can be concluded that the large amount of training data greatly affects the impunity with the mean.

Testing with MAPE showed that dataset 1 median has the lowest error value while in dataset 2 median and mode have same and the lowest value. In dataset 3, median is lead ahead of mean and mode. This test is even more convincing when compared to the RMSE test that the median has the most stable accuracy of each existing dataset meaning that imputation with the median does not have much influence on the amount of training data and the number of missing values. In the second order of the best imputations using a statistical basis we chose the mean and then the imputation with the last mode of the three static-based imputation methods because the average had the largest error of the two tests.

5.2 k-Nearest Neighbor imputation (kNNI)

Missing value prediction with kNNI uses the KNN Imputer function from the scikit-learn library. The function is very simple, just input the dataset and then determine the k value used. Figure 8 shows the command line for missing value prediction with kNNI with k=15. The first step declares the numpy library into a variable. The second step inserts the dataset into the array variable. The third step determines the value of k used. The last step prints the imputation prediction results with the specified k value and dataset as shown in Tables 7-9.

```
In [1]: import numpy as np

In [2]: from sklearn.impute import KNNImputer

In [106]: X = [[5,5,4,5,5,5,5], [4,4,4,5,4,4,4], [4,4,4,4,4,4,4],

In [114]: imputer = KNNImputer(n_neighbors=15)

In [115]: imputer.fit_transform(X)

Out[115]: array([[5.  , 5.  , 4.  , 5.  , 5.  , 5.  , 5.  ],
                 [4.  , 4.  , 4.  , 5.  , 4.  , 4.  , 4.  ],
                 [4.  , 4.  , 4.  , 4.  , 4.  , 4.  , 4.  ],
                 [4.  , 4.  , 3.  , 4.  , 5.  , 4.  , 4.  ],
                 [5.  , 4.  , 4.  , 4.  , 4.  , 4.  , 4.  ],
```

Figure 8. kNN Imputer (Scikit-Learn)

Table 7. kNNI in dataset 1

Alumni	Attribute	Ori Value	k 1	k 5	k 10	k 15	k 20
A77	C1	5	5	5	5	5	5
A47	C2	4	4	4	4	4	4.1
A54	C3	4	3	4	3.9	4.1	4.1
A41	C4	4	4	4	4	3.9	4
A3	C5	3	4	4	4.1	4.1	4.1
A99	C6	5	4	5	4.7	4.6	4.7
A1	C7	3	3	3	3.7	3.7	3.7
A100	C6	5	5	5	4.7	4.8	4.8
A78	C5	5	5	5	5	4,9	4,9
A42	C4	3	3	3	4	4	4

Table 8. kNNI in dataset 2

Alumni	Attribute	Ori Value	k 1	k 5	k 10	k 15	k 20
A41	C3	3	3	2.8	3	3.07	3.25
A3	C4	4	4	3.6	3.7	3.80	3.85
A90	C5	5	5	5	5	5	5
A97	C6	4	5	4	4.1	4.13	4.2
A42	C7	4	4	4	4	4	3.9
..	..	..	..	..	..	..	..
A4	C5	3	3	4	4	4.13	4.1
A99	C6	5	4	4.8	4.9	4.67	4.65
A5	C7	3	3	3.6	3.7	3.67	3.7
A100	C6	5	5	4.8	4.9	4.87	4.8
A96	C5	5	5	4.8	4.9	4.93	4.95
A46	C4	3	3	3.8	4	4.07	3.95

Table 9. kNNI in dataset 3

Alumni	Attribute	Ori Value	k 1	k 5	k 10	k 15	k 20
A33	C5	4	5	4	4.1	4.07	4.1
A95	C6	4	4	4.4	4.3	4.20	4.2
A34	C7	4	5	4.6	4.4	4.33	4.3
A96	C6	4	5	4.8	4.7	4.53	4.5
A35	C5	4	5	4.8	4.6	4.53	4.6
..	..	..	..	..	..	..	..
A37	C1	4	4	4.2	4.2	4.20	4.25
A38	C2	5	4	4.2	4.1	4.07	4.15
A39	C3	3	3	3	3.2	3.33	3.25
A3	C4	4	4	3.6	3.7	3.80	3.85
A88	C5	5	5	5	5	5	5
..	..	..	..	..	..	..	..
A4	C5	3	4	3.8	3.9	4,13	4.05
A99	C6	5	4	5	4.8	4.73	4.7
A5	C7	3	3	3.6	3.6	3.67	3.7
A100	C6	5	5	4.8	4.9	4.93	4.85
A94	C5	5	5	5	4.9	4.93	4.9
A44	C4	3	3	3.8	4.1	4	3.95

The result of kNNI method on the three datasets then undergo RMSE and MAPE test. Testing the error value is necessary because when viewed from the data the prediction results show quite varied results for the best k value. Table 10 shows the output of RMSE and MAPE test for the three datasets while Figure 9 and Figure 10 present the graphical comparisons of the output.

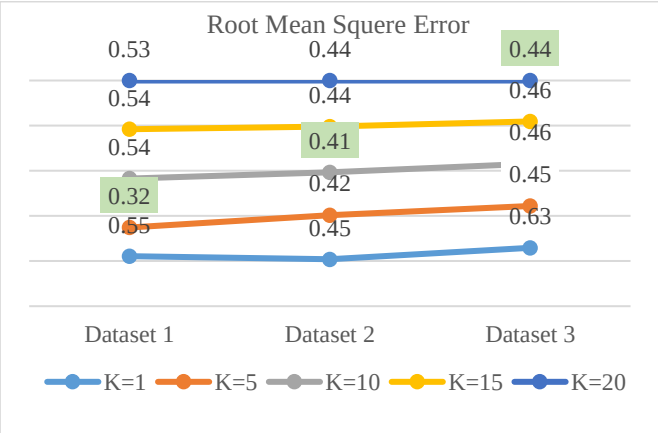

Figure 9. RMSE to kNNI

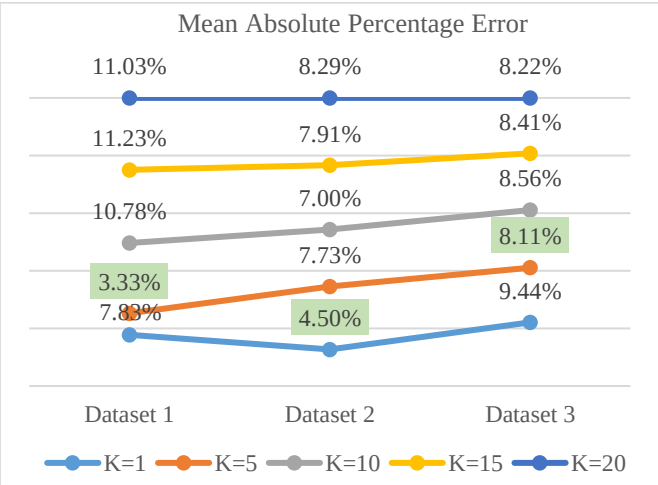

Figure 10. MAPE to kNNI

Table 10. k Nearest Neighboar imputation

Method (kNNI)	RMSE			MAPE		
	Dataset 1	Dataset 2	Dataset 3	Dataset 1	Dataset 2	Dataset 3
k=1	0.55	0.45	0.63	7.83%	4.50%	9.44%
k=5	0.32	0.42	0.45	3.33%	7.73%	8.11%
k=10	0.54	0.41	0.46	10.78%	7.00%	8.56%
k=15	0.54	0.44	0.46	11.23%	7.91%	8.41%
k=20	0.53	0.44	0.44	11.03%	8.29%	8.22%

Figure 9 and Figure 10 show the imputation with kNNI k=5 the majority of the superior accuracy as evidenced by the lowest error value, followed by kNNI k=10 where in RMSE superior in dataset 2, kNNI k=1 in MAPE superior also in dataset 2, and kNNI k=20 in RMSE superior in dataset 3. In terms of datasets, in dataset 1 which has a small missing value while large training data has the smallest error value at one k value, namely k=5 but the majority have the largest error value compared to other datasets, namely k=10, k=15, and k=20, this means that if the training data is large, you must choose the right k value to get the smallest error value. Dataset 2 shows fairly flat error values. Dataset 3, where the missing value is the most and the training data is the least compared to other datasets shows that the greater the k value, the lower the error value except for MAPE kNNI k=5.

5.3 Comparison methods

The machine learning imputation method with kNNI is a better imputation approach than the three statistical-based imputation methods. The results of testing error values with RMSE and MAPE showed that the effect of calculating distances was so great on the imputation results. All error calculation results show that kNNI imputation results are much better than statistical-based imputations. Based on the standards used in the MAPE test, the kNNI test results (error values) are almost entirely very good because the error value is <10% except for three test results in dataset 1 at k=10, k=15 and k=20, while the statistical-based imputation results are only good because the error value is between 10% and 20%, even one of the results exceeds 20%. A comparison of error value testing with RMSE and MAPE on all three datasets can be seen in Table 11 and Figures 11 and 12.

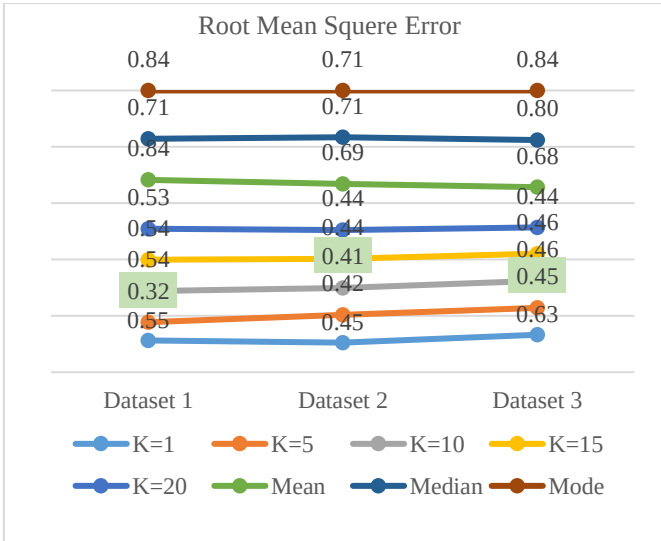

Figure 11. RMSE to Imputation Statistic-Based and kNNI

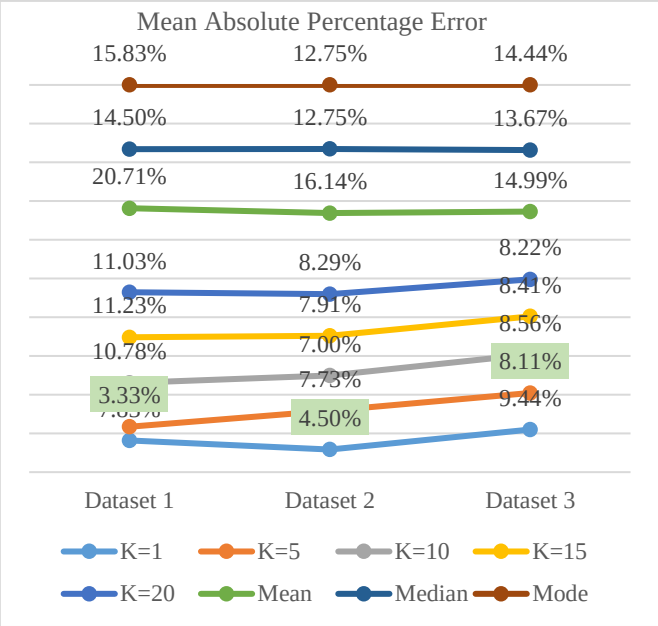

Figure 12. MAPE to imputation statistic-based and kNNI

Table 11. Result imputation

Method	RMSE			MAPE		
	Dataset 1	Dataset 2	Dataset 3	Dataset 1	Dataset 2	Dataset 3
Statistic-based						
Mean	0.84	0.69	0.68	20.71%	16.14%	14.99%
Median	0.71	0.71	0.80	14.50%	12.75%	13.67%
Mode	0.84	0.71	0.84	15.83%	12.75%	14.44%
kNNI						
k=1	0.55	0.45	0.63	7.83%	4.50%	9.44%
k=5	0.32	0.42	0.45	3.33%	7.73%	8.11%
k=10	0.54	0.41	0.46	10.78%	7.00%	8.56%
k=15	0.54	0.44	0.46	11.23%	7.91%	8.41%
k=20	0.53	0.44	0.44	11.03%	8.29%	8.22%

6. CONCLUSION AND DISCUSSION

Imputation is the preferred choice in handling missing values as opposed to eliminating the entire record which will reduce the number of datasets, especially in the case of single imputation or only one attribute of one record is missing. This study examines several simple imputation methods to find the best method to predict the missing value. The methods involve three statistical based imputation and a machine learning based imputation using kNNI algorithms. because the dataset tested only has one attribute from one record that has a missing value so that with simple imputation it will speed up the process of predicting the value to be imputed. Most guided classification methods or unguided/clustering classifications are created without a mechanism for handling missing values, so other methods or ways are needed for it. Imputation is expected to improve the accuracy of the dataset management process as follows, such as classification.

The results of testing the error value with RMSE and MAPE on statistical-based imputation show that the median has a fairly stable error value on the three datasets with different amounts of training and testing data, the mean also shows fairly good results but is highly dependent on the amount of training and testing data available, while the mode has the highest error value.

The results of testing error values with RMSE and MAPE on kNNI showed imputation with kNNI k=5 the majority were superior in accuracy as evidenced by the lowest error values, followed by kNNI k=10 where in RMSE superior in dataset 2, kNNI k=1 in MAPE also excelled in dataset 2, and kNNI k=20 in RMSE superior in dataset 3. In terms of datasets, in dataset 1 which has a small missing value while large training data has the smallest error value at one k value, namely k=5 but the majority have the largest error value compared to other datasets, namely k=10, k=15, and k=20, this means that if the training data is large, you must choose the right k value to get the smallest error value. Dataset 2 shows fairly flat error values. Dataset 3, where the missing value is the most and the training data is the least compared to other datasets shows that the greater the k value, the lower the error value except for MAPE kNNI k=5.

The results of testing error values with RMSE and MAPE showed all error calculation results show that kNNI imputation results are much better than statistical-based imputations. Based on the standards used in the MAPE test, the kNNI test results (error values) are almost entirely very good because the error value is <10% except for three test results in dataset 1 at k=10, k=15 and k=20, while the statistical-based imputation results are only good because the error value is between 10% and 20%, even one of the results exceeds 20%. Although kNNI is better than statistical-based imputation, it needs to choose the right k value to get the best imputation results. From all test results, the author suggests STMIK PPKIA Tarakanita Rahmawati to use the kNN imputation method with k = 5 as a solution to handling missing value in graduate user satisfaction level data in order to get the best results in the clustering process.

This study has tested three variations of numeric datasets on handling missing values with three types of statistical-based methods and kNN Imputation with 5 variations of k values with two error value measurement methods, namely RMSE and MAPE. It is expected that further research will add variations both in terms of datasets, imputation methods or methods of measuring error / accuracy values, and can test imputed datasets in the classification and / or clustering process.

REFERENCES

[1] Yan, Y., Wu, Y., Du, X., Zhang, Y. (2021). Incomplete data ensemble classification using imputation-revision framework with local spatial neighborhood information. Applied Soft Computing, 99: 106905. https://doi.org/10.1016/j.asoc.2020.106905

[2] Jadhav, A., Pramod, D., Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. Applied Artificial Intelligence, 33(10): 913-933. https://doi.org/10.1080/08839514.2019.1637138

[3] Schmitt, P., Mandel, J., Guedj, M. (2015). A comparison of six methods for missing data imputation. Journal of Biometrics & Biostatistics, 6(1): 1. https://doi.org/10.4172/2155-6180.1000224

[4] Ngueilbaye, A., Wang, H., Mahamat, D.A., dan Junaidu, S.B. (2021). Modulo 9 model-based learning for missing data imputation. Applied Soft Computing, 103: 107167. https://doi.org/10.1016/j.asoc.2021.107167

[5] Luengo, J., García, S., Herrera, F. (2012). On the choice of the best imputation methods for missing values considering three groups of classification methods. Knowledge and Information Systems, 32: 77-108. https://doi.org/10.1007/s10115-011-0424-2

[6] Ghorbani, S., Desmarais, M.C. (2017). Performance comparison of recent imputation methods for classification tasks over binary data. Applied Artificial Intelligence, 31(1): 1-22. https://doi.org/10.1080/08839514.2017.1279046

[7] Xu, X., Chong, W., Li, S., Arabo, A., Xiao, J. (2018). MIAEC: Missing data imputation based on the evidence chain. IEEE Access, 6: 12983-12992. https://doi.org/10.1109/ACCESS.2018.2803755

[8] Nanni, L., Lumini, A., Brahnam, S. (2012). A classifier ensemble approach for the missing feature problem. Artificial Intelligence in Medicine, 55(1): 37-50. https://doi.org/10.1016/j.artmed.2011.11.006

[9] Joshi, A., Kale, S., Chandel, S., Pal, D.K. (2015). Likert scale: Explored and explained. British Journal of Applied Science & Technology, 7(4): 396. https://doi.org/10.9734/bjast/2015/14975

[10] Deb, R., Liew, A.W.C. (2016). Missing value imputation for the analysis of incomplete traffic accident data. Information Sciences, 339: 274-289. https://doi.org/10.1016/j.ins.2016.01.018

[11] Little, R.J., D'Agostino, R., Cohen, M.L., Dickersin, K., Emerson, S.S., Farrar, J.T., Stern, H. (2012). The prevention and treatment of missing data in clinical trials. New England Journal of Medicine, 367(14): 1355-1360. https://doi.org/10.1056/nejmsr1203730

[12] Jerez, J.M., Molina, I., García-Laencina, P.J., Alba, E., Ribelles, N., Martín, M., Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial Intelligence in Medicine, 50(2): 105-115. https://doi.org/10.1016/j.artmed.2010.05.002

[13] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Altman, R.B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics, 17(6): 520-525. https://doi.org/10.1093/bioinformatics/17.6.520

[14] Schneider, T. (2001). Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. Journal of Climate, 14(5): 853-871. https://doi.org/10.1175/1520-0442(2001)014<0853:AOICDE>2.0.CO;2

[15] Deb, R., Liew, A.W.C. (2016). Missing value imputation for the analysis of incomplete traffic accident data. Information Sciences, 339: 274-289. https://doi.org/10.1007/978-3-662-45652-1_28

[16] Noyunsan, C., Katanyukul, T., Saikaew, K. (2018). Performance evaluation of supervised learning algorithms with various training data sizes and missing attributes. Engineering and Applied Science Research, 45(3): 221-229. https://doi.org/10.14456/easr.2018.28

[17] Choudhury, S.J., Pal, N.R. (2019). Imputation of missing data with neural networks for classification. Knowledge-Based Systems, 182: 104838. https://doi.org/10.1016/j.knosys.2019.07.009

[18] Acuna, E., Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. In Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, July 15-18, 2004, pp. 639-647. https://doi.org/10.1007/978-3-642-17103-1_60

[19] Dixon, J.K. (1979). Pattern recognition with partly missing data. IEEE Transactions on Systems, Man, and Cybernetics, 9(10): 617-621. https://doi.org/10.1109/TSMC.1979.4310090

[20] Batista, G.E., Monard, M.C. (2003). An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence, 17(5-6): 519-533. https://doi.org/10.1080/713827181

[21] Shanthi, J., Rani, D.G.N., Rajaram, S. (2022). A C4. 5 decision tree classifier based floorplanning algorithm for System-on-Chip design. Microelectronics Journal, 121: 105361. https://doi.org/10.1016/j.mejo.2022.105361

[22] Swe, S.M., Sett, K.M. (2019). Approaching rules induction CN2 algorithm in categorizing of biodiversity. International Journal of Trend in Scientific Research and Development, 3(4): 1581-1584. https://doi.org/10.31142/ijtsrd25153

[23] Grzymala-Busse, J.W., Hu, M. (2001). A comparison of several approaches to missing attribute values in data mining. In Rough Sets and Current Trends in Computing: Second International Conference, RSCTC 2000 Banff, Canada, October 16-19, 2000, pp. 378-385. https://doi.org/10.1007/3-540-45554-X_46

[24] Mundfrom, D.J., Whitcomb, A. (1998). Imputing missing values: The effect on the accuracy of classification. https://files.eric.ed.gov/fulltext/ED419817.pdf

[25] Zhang, Z. (2016). Missing data imputation: focusing on single imputation. Annals of translational medicine, 4(1): 9. https://doi.org/10.3978/j.issn.2305-5839.2015.12.38

[26] Donders, A.R.T., Van Der Heijden, G.J., Stijnen, T., Moons, K.G. (2006). A gentle introduction to imputation of missing values. Journal of clinical epidemiology, 59(10): 1087-1091. https://doi.org/10.1016/j.jclinepi.2006.01.014

[27] Cover, T., Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1): 21-27. https://doi.org/10.1109/TIT.1967.1053964

[28] Jonsson, P., Wohlin, C. (2004). An evaluation of k-nearest neighbour imputation using likert data. In 10th International Symposium on Software Metrics, 2004. Proceedings, pp. 108-118. https://doi.org/10.1109/METRIC.2004.1357895

[29] Hodson, T.O. (2022). Root-Mean-Square Error (RMSE) or Mean Absolute Error (MAE): When to use them or not. Geoscientific Model Development, 15(14): 5481-5487. https://doi.org/10.5194/gmd-15-5481-2022

[30] Al-Khowarizmi, R.S., Nasution, M.K., Elveny, M. (2021). Sensitivity of MAPE using detection rate for big data forecasting crude palm oil on k-nearest neighbor. International Journal of Electrical and Computer Engineering (IJECE), 11(3): 2696-2703. https://doi.org/10.11591/ijece.v11i3.pp2696-2703