The IJICS (International Journal of Informatics and Computer Science)
Vol 5 No 2, July 2021, Page 141-144
ISSN 2548-8384 (online), ISSN 2548-8449 (print)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/ijics/index
DOI 10.30865/ijics.v5i2.3185

Analysis of Missing Value Imputation Application with K-Nearest Neighbor (K-NN) Algorithm in Dataset

Achmad Fikri Sallaby[1,*], Azlan[2]

[1] Universitas Dehasen, Bengkulu, Indonesia
[2] STMIK Triguna Dharma, Medan, Indonesia
Email: [1,*]Fikrisallaby@gmail.com, [2]Azlansaja19@gmail.com
Coressponding Author: Fikrisallaby@gmail.com
Submitted: **18/06/2021**; Accepted: **30/07/2021**; Published: **31/07/2021**

Abstract−Missing value is a problem that is still often found in many studies. Missing value is where data or data features are not available completely and intact. This still happens a lot in datasets that will be used in research. The missing value is caused by many factors such as human error, unavailable data or even from a virus in the database. Data is important for research, incomplete data will affect the results obtained. Data mining is a process that is very influential on data, including the classification process. Classification in data mining can be done if the data is complete. These problems can be overcome by the Imputation process by combining it with the K-Nearest Neighbor process or the process can be called K-Nearest Neighbor Imputation (K-NNI). In the research that has been done the K-Nearest Neighbor Imputation algorithm can overcome the problem of missing values in the dataset. This can be seen from the level of accuracy obtained where the accuracy of the classification process before handling the missing value is 77.01% while after the imputation process the accuracy is 78.31%.

Keywords: Missing Value; Data Mining; Datasets; Imputation; K-Nearest Neighbor Imputation (K-NNI)

1. INTRODUCTION

In many studies, it is often found that the problem of loss of value in the dataset used is called missing value. Missing value is the unavailability of data or features in the dataset which causes the dataset to be incomplete. Missing value is caused by many factors such as human error, data that is not available when collected or data that is lost due to a virus in the database. Data is a very important part of research. Loss of data in the dataset will cause inaccuracies in the results obtained[1]–[3].

Data mining is a process that is very influential on the data. In data mining, data that has been stored and piled up becomes a lot and then processed to get new information stored in the data set. This is what causes the data to be very influential on data mining in getting results. In many problems that occur, the missing value will greatly affect the pattern recognition process (classification) in data mining. The missing value problem really determines the pattern obtained from the classification process.

Missing value is a major problem in the classification process where the entire classification process can be carried out if the available data is complete. From these problems, it is encouraging to solve problems related to missing values, namely imputation. Imputation is a way to solve the missing value problem. Where the process is carried out by eliminating values that do not match the data set, looking for missing values in the dataset by making estimates based on certain methods and other ways[4]–[6].

The K-Nearest Neighbor (K-NN) algorithm is one method that can be combined with Imputation to find values and overcome the missing value problem or commonly called K-Nearest Neighbor Imputation (K-NNI). K-NNI is part of the data mining classification process that performs pattern discovery based on the distance to its closest neighbors. To measure the distance to the nearest neighbor based on the value of Euclidean Distance. K-NNI imputed the missing value based on information from the closest observation that had similarities with the missing value.

Several studies have been carried out on the K-Nearest Neighbor Imputation (K-NNI) method, such as that conducted by Iman Jihad Fadillah and Siti Muchlisoh, the K-Nearest Neighbor Imputation (K-NNI) algorithm has a better accuracy value than the Hot-Nearest Neighbor Imputation (K-NNI) algorithm. Deck Imputation[7]. Another study was conducted by Taufiq Rizaldi, et al. The performance of the K-Nearest Neighbor Imputation (K-NNI) algorithm has a better performance level than the Naïve Bayes Imputation algorithm[6].

In this study, research will be carried out by conducting a process to restore the missing values in the dataset and then comparing the final results of the performance of the K-NNI algorithm by comparing the accuracy of the classification process carried out.

2. RESEARCH METHODOLOGY

2.1. Research Methodology

The research methodology is the steps taken in solving the problem. Where the process starts from analysis, process, to conclusions and research reports. The methodology in this research is used as a flow to solve problems. The research methodology is also a problem-solving framework so that the processes carried out in research are not repeated which causes the research time to take longer.

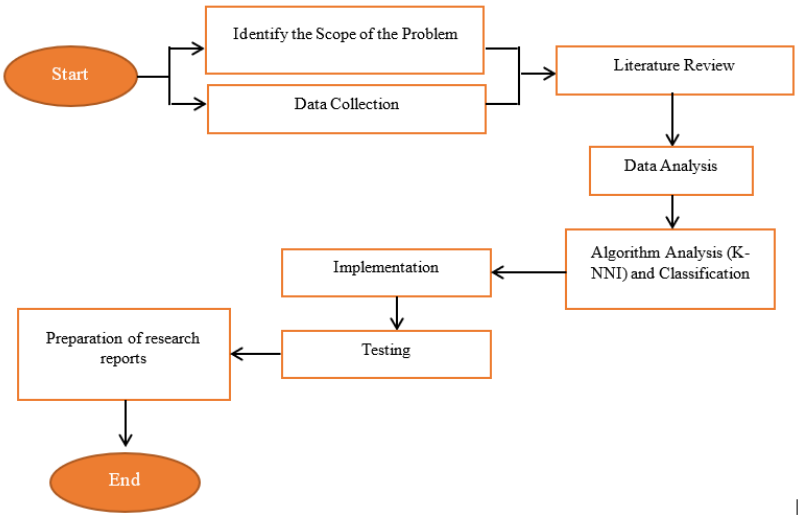

Figure 1. Research Methodology

2.2 Data Mining

Data mining is a process that uses statistics, artificial intelligence, and machine learning to extract and identify useful ones. Data mining is defined as the process of finding patterns in data. Based on the task, data mining is grouped into description, prediction estimation, classification, clustering, and association. A process of mining important information from a data. This important information is obtained from a very complicated process such as using artificial intelligence, statistical techniques, mathematics, machine learning and so on[8]–[11]. So, based on the explanation above, it can be concluded that data mining is a process of mining data in very large amounts of data using statistical, mathematical methods, to utilizing the latest artificial intelligence technology. According to experts, the purpose of data mining is to extract and identify data for certain information related to a large database or big data.

2.3 K-Nearest Neighbor Imputation (K-NNI)

Nearest Neighbor (NN) is a method that uses a supervised learning algorithm. Supervised learning aims to find new patterns in the data by connecting existing data patterns with new data. There are two types of NN algorithms, namely 1NN and KNN. 1NN or Nearest Neighbor is an approach that performs classification on the nearest 1 data, while KNN is an approach that performs classification on the nearest K data, with K > 1. KNN is a method used to classify objects based on some data that is closest to the object. In classification, KNN works by calculating the distance between new data (testing data) and data whose class is known (training data) using Euclidian distance.

Handling missing data with KNN begins with determining a number of nearest neighbors or closest observations symbolized by K, then calculating the smallest distance from each observation that does not contain missing data. The steps for imputing missing data with the KNN method are as follows[12]–[14]:

1. Determine the parameter K, K is the number of closest observations or nearest neighbors to be used.
2. Calculate the distance between observations containing missing data and complete observations on the jth variable that does not contain missing data with other j variables that correspond to the Euclidian distance formula, namely:

   $$d(xa,xb) = \sqrt{\sum_{j=1}^{m}(x_{aj} - x_{bj})^2} \quad (1)$$

   $d(x_a,x_b)$ is the distance between observations containing missing data and observations that do not contain missing data, is the value of the j-th variable in each observation containing missing data with j = 1,2,…,m,x_{bj} is the value of the other variables in each observation that does not contain missing data with = 1,2,…,m
3. Sort the distance based on the observation that has the largest distance value to the observation that has the smallest distance value.
4. Determine the closest K observations based on the smallest distance value.
5. Impute missing data by calculating the weight mean estimation value on the closest K observations that do not contain missing data values with the formula:

   $$\bar{x}j = \frac{\sum_{k=1}^{K} wk\ vk}{\sum_{k=1}^{K} wk} \quad (2)$$

   where $\bar{x}j$ is the estimated weighted average, vk is the value of the complete data on the variable containing missing data based on observations from k, K is the number of closest observations used, k is the observation of K, wk is the weight of the K-th closest neighbor observation with formula wk = $\frac{1}{d(x_{ak}x_{bk})^2}$, where d(xa,xb) is the observation distance K.

3. RESULT AND DISCUSSION

At this stage, we will discuss the implementation of the K-NNI algorithm in restoring lost data in the dataset. The first step in this research is to find out the pattern in the classification and measure the level of accuracy of the dataset which is still not completely available for the data value. Datasets have data structures

Table 1. Dataset

Atribut	Kelas
Survived	C
Pclass	Q
Age	S
SibSp	
Parch	
Tiket	
Fare	
Cabin	

From the dataset above with incomplete data, then classification is carried out and then measuring the performance results carried out in the classification process proses

Table 2. Classification Performance

Accuracy: 77,01%

	True S	True C	True Q	Class Precision
Pred. S	641	129	57	77,51%
Pred. C	14	49	0	77,78%
Pred. Q	6	6	20	62,50%
Class Recall	96,97%	26,63%	25,97%	

In the process that has been carried out on the dataset by measuring performance in the classification, the results obtained an accuracy rate of 78.31%. Then carry out the process to find the missing value in the missing value dataset by using the K-Nearest Neighbor Imputation (K-NNI) algorithm and then again measure the performance of the classification results.

Table 3. Classification Performance + K-NNI

Accuracy: 78,31%

	True S	True C	True Q	Class Precision
Pred. S	633	114	47	79,72%
Pred. C	17	59	0	77,63%
Pred. Q	11	11	30	57,69%
Class Recall	95,76%	32,07%	38,96%	

Tabel 4. Comparison of Performance Results

	Akurasi
Klasifikasi	77,01%
Klasifikasi + K-NNI	78,31%

4. CONCLUSION

Based on the research conducted, it can be concluded that the K-Nearest Neighbor Imputation algorithm can overcome the problem of missing values in the dataset. The K-Nearest Neighbor Imputation Algorithm can also help improve performance than data mining classification on datasets that do not have complete value features in it. This can be seen from the level of accuracy obtained where the accuracy of the classification process before handling the missing value is 77.01% while after the imputation process the accuracy is 78.31%.

REFERENCES

[1] J. Seita, "Missing data Missing data," *Reclaiming Child. Youth*, vol. 19, no. March, pp. 7–8, 2007.
[2] I. J. Fadillah and C. D. Puspita, "Pemanfaatan Metode Weighted K-Nearest Neighbor Impution (Weighted KNNI) Untuk Mengatasi Missing Data: Penerapan pada Data Indeks Produksi Triwulanan Industri Mikro Kecil (IMK) Tahun 2016-2019," *Semin. Nas. Off. Stat. 2019 Pengemb. Off. Stat. dalam mendukung Implementasi SDG's*, pp. 511–518, 2020.
[3] I. Bagus and G. Narinda, "Missing Value Imputation Using KNN Method Optimized With Memetic Algorithm," *e-Proceeding Eng.*, vol. 3, no. 1, pp. 1098–1105, 2016.
[4] Susanti, S. Martha, and E. Sulistianingsih, "K-Nearest Neighbor Dalam Imputasi Missing Data," *Bul. Ilm. Math. Stat. dan Ter.*, vol. 07, no. 1, pp. 9–14, 2018.

[5] A. Izzah, S. Ramadhan, and P. D. K. Means, "Imputasi Missing data Menggunakan Algoritma Pengelompokan Data K- Harmonic Means Related papers Imputasi Missing data Menggunakan Algoritma."
[6] T. Rizaldi, F. E. Purnomo, and A. S. Arifianto, "Perbandingan Metode K-Nn Dan Bayes Pada Missing Imputation," *J. Teknol. Inf. dan Terap.*, vol. 5, no. 2, pp. 85–90, 2019, doi: 10.25047/jtit.v5i2.84.
[7] I. J. Fadillah and S. Muchlisoh, "Perbandingan Metode Hot-Deck Imputation Dan Metode Knni Dalam Mengatasi Missing Values," *Semin. Nas. Off. Stat.*, vol. 2019, no. 1, pp. 275–285, 2020, doi: 10.34123/semnasoffstat.v2019i1.101.
[8] D. Nofriansyah and G. W. Nurcahyo, *Algoritma Data Mining Dan Pengujiannya*. Yogyakarta: Deepublish, 2015.
[9] D. Nofriansyah, *Konsep Data Mining Vs Sistem Pendukung Keputusan*. Yogyakarta: Deepublish, 2014.
[10] B. Efori, *Data Mining Untuk Perguruan Tinggi*. Yogyakarta: Deepublish, 2020.
[11] E. Prasetyo, *Data Mining : Konsep dan Aplikasi Menggunakan Matlab*. Yogyakarta: CV. Andi Offset, 2012.
[12] U. Mawarsari, "IMPUTASI MISSING DATA DENGAN K-NEAREST NEIGHBOR DAN ALGORITMA GENETIKA," *AdMathEdu*, vol. 6, no. 1, pp. 77–86, 2016.
[13] Moch. Lutfi and Mochamad Hasyim, "Penanganan Data Missing Value Pada Kualitas Produksi Jagung Dengan Menggunakan Metode K-Nn Imputation Pada Algoritma C4.5," *J. Resist. (Rekayasa Sist. Komputer)*, vol. 2, no. 2, pp. 89–104, 2019, doi: 10.31598/jurnalresistor.v2i2.427.
[14] E. Sartika, "Analisis metode k nearest neighbor imputation (knni) untuk mengatasi data hilang pada estimasi data survey," *Tedc*, vol. 12, no. 3, pp. 219–227, 2018.