

Improvement of Accuracy and Handling of Missing Value Data in the Naive Bayes Kernel Algorithm

Bijanto^{*1}, Ryan Yunus²,

^{1,2}*Technical College of Pati / Pati-Tayu Km 3.5 Pati, telp/fax of institution/affiliation*

*E-mail : biyantokakoi@gmail.com^{*1}, riyanyunus@sttp.ac.id²*

**Corresponding author*

Zainal Abidin³

³*Technical College of Pati / Pati-Tayu Km 3.5 Pati, telp/fax of institution/affiliation*

E-mail : zainal.frsd@yahoo.com³

Abstract – The lost impact on the research process, can be serious in classifying results leading to biased parameter estimates, statistical information, decreased quality, increased standard error, and weak generalization of the findings. In this study, researchers discuss the problems that exist in one of the algorithms, namely the Naive Bayes Kernel algorithm. The Naive Bayes kernel algorithm has the disadvantage of not being able to process data with the mission value. Therefore, in order to process missing value data, there is one method that we propose to overcome, namely using the mean imputation method. The data we use is public data from UCI, namely the HCV (Hepatitis C Virus) dataset. The input method used to correct missing data so that it can be filled with the average value of the existing data. Before the imputation process means, the dataset uses yahoo bootstrap first. The data that has been corrected using the mean imputation method has just been processed using the Naive Bayes Kernel Algorithm. From the results of the research tests that have been carried out, it can be obtained an accuracy value of 96.05% and the speed of the data computing process with 1 second.

Keywords – *missing value, bootstrap, mean imputation, naive bayes kernel*

1. INTRODUCTION

In the current era, the term data mining has become a term that is often used in the medical literature, especially medicine along with computer science. Data processing can be used as a large data model to find association patterns that have not been recognized. The data input is used as an evaluation system, which results in useful data analysts. [1]. Of course this is very much needed accuracy in obtaining results when processing data, especially in the medical or medical field. In the data mining process, the data that has been stored and compiled will be more and more and then processed to produce new information stored in the data set. This is what will make the data very influential on data mining to produce a conclusion or decision. In cases that occur, missing value data greatly affects the pattern recognition (classification) process in data mining. The missing value problem greatly affects the pattern formed from the classification process [2]. This paper presents a kernel-based naive bayesian model that is used in case of problems using a supervised classification model for classification.

Naive bayes plus network tree, full graph classifier, and k-dependence bayesian classification model confirmed to an innovative kernel based naive bayes paradigm, in addition to strong naive bayes classifier ability to predict/classify. Supervised classification is good

training in recognizing patterns. This type of classification requires the creation of a classification algorithm, which is defined by a function that provides the required tags or class-specific identification labels to the instances described by the variable group described in a single dataset. Naive Bayes classification as a basic probabilistic classification based on Bayes theory. naive bayes algorithm process or not process any property that is different from the provided class also known as fixed attribute is not affected by the appearance of any additional features present in the data.

The main benefit of using naive bayesian algorithms for small amounts of data used for training is to train the system to measure variance as well as the means of all variables provided in the dataset, an important requirement for classification. Only the variance of the variables for each individual label needs to be evaluated and not the entire covariance matrix for the data provided as input to the system on the grounds that all independent variables are grouped in the Naive Bayes kernel class implemented on the numeric attributes in contrast with the Naive Bayesian classification. The function used in the nonparametric estimation process is known as the kernel. Different kernels are applied to the kernel process density prediction technique for the estimation of the density function of the random variables, in the kernel regression mechanism for the expected conditional estimation of the random variables in the data. This kernel density prediction has a special class of estimation functions known as nonparametric density estimators, The process applied for medical prediction includes the following steps:

Data collection retrieval.

Data pre-processing.

- A. Data cleaning.
- B. Replace missing values.
- C. Outlier identification and noise removal.
- D. Data transformation (scaling, conversion, and normalization).
- E. Machine learning algorithm (kernel based Naïve Bayes classifier).

The results of this study can be used as recommendations and input for health experts in making predictions of hepatitis disease, helping higher education administrations to provide early warning and early guidance for students who may not graduate on time. The scope of this study is limited to the use of the mean imputation method, in predicting hepatitis disease and comparing the accuracy of the method [3]. This study uses the naive Bayes algorithm to predict HCV. Naïve Bayesian strengths and weaknesses [4] Strengths: a. Easy to implement. b. Gives good results for most cases. Weaknesses: a. Must assume that between features are not related (independent) In reality, the relationship exists This relationship cannot be modeled by the Naïve Bayesian Classifier.

2. RESEARCH METHOD

2.1. Missing Value

Missing value is the incompleteness of data or components in the dataset which causes the dataset to be imperfect. Incomplete values are caused by several factors, including human error, lost data due to viruses in the database. In research data is very important. Missing or incomplete data in the dataset will cause the results obtained to be inaccurate[2],[5]. In the missing value data, there are several causes that result in the loss of data in records and attributes. This is because there is no response in one part or several parts of the data source and can be one of the factors that result in missing values in the dataset. [6]. The problem of missing data itself, in data processing will greatly affect the classification results [7]. The data we use in this study is HCV (Hepatitis C Virus) data taken from UCI. The dataset consists of 615 records, 12 attributes and 1 label [8]. In the HVC dataset, there are some missing data values.

Because the dataset used has a missing value, a problem arises in the algorithm that we will use to process the classification results using the Naive Bayes kernel algorithm. Below is an example of a table with missing values:

Table 1. Example of a dataset table with missing values

Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT	Category
23	0	47	19.1	38.9	164.2	17	7.09	3.2	79.3	90.4	70.1	3
25	0	42	38.2	63.3	167.7	14	6	4.28	66.9	40.2	70.5	3
27	0	45	27.5	10.5	37.8	10	8.77	3.2	55.2	35.9	74.5	3
29	0	49	NA	53	39	15	8.79	3.6	79	37	90	3
30	0	45	NA	66	45	14	12.16	6.1	86	43	77	3
31	0	45	34.6	44.3	96.2	16	10.11	6.28	81.2	48.1	82.1	3
32	0	41	34.4	12.1	60.9	6	13.8	5.48	45.4	33.1	71.1	3

In the dataset because there are missing data values, there are several methods used, including filling in the minimum, average and maximum values from the existing data. Before the process of recharging with the method already mentioned, there is another step, namely bootstrap. After processing the new bootstrap, it will proceed to the process of replacing missing values and proceed again to processing using the Naive Bayes kernel algorithm.

2.2. Bootstrap

Before entering the process of replacing missing values with the mean replace missing value method, the data will be processed using a bootstrap. The bootstrap method is used to estimate the acquisition of an unknown group or population with real gains obtained from the repeat pilot process. Returning the original sample or sample is a technique used by the bootstrap method. The results of observations that are considered as if the population is turned into a sample which is called the original sample.

The use of the bootstrap method aims to obtain parameter estimates based on minimal data. In statistics, scanty data and data that do not match certain assumptions or data that do not have any meaning about their acquisition can be referred to as minimal data.

The following are some assumptions about the bootstrap method:

1. The sample used to represent the population is an appropriate sample that has been owned.
2. The method that is used to estimate the acquisition of an unknown group or population with the actual gain obtained from the re-sampling process is called the bootstrap method. So, each Bootstrap sample provides each other. However, each Bootstrap sample is independent of each other or unaffected by the others.

The training data is retrieved and then returned to the initial dataset so that it has the opportunity to retrieve it, that's what the bootstrap method does. If the first dataset has N data, it can be seen that the average bootstrap sample has a value of N data in the range of 63.2% originating from the original dataset.

This method is in accordance with the fact that the bootstrap sample performs data probability selection, namely:

$$1-(1-1/N)^N \quad (1)$$

The probability will approach $1-e^{-1} = 0.632$ with the asymptotic technique, when the value of N is large. Data that is not categorized as a bootstrapping sample will automatically be part of the testing data. The model that is formed from training data will be implemented in data

testing which is used to produce Bootstrap sample accuracy, namely (ci) To generate b bootstrap samples, samples are taken many times as many as b [9].

There are many ways to approach the bootstrap sample regarding all the accuracy of the arithmetic classification. A commonly used example is Bootstrap 0.632 where in finding the accuracy of everything by combining the accuracy of each bootstrap sample (ci) with the accuracy sought from training data where the data is known for its class label (accs). For this, the formula is:

$$acc_{boot} = \frac{1}{b} \sum_{i=1}^b (0.632 \times ci + 0.368 \times accs) \quad (2)$$

When processing data, the number of attributes and the amount of data is large, it greatly affects the performance of the computer that must be used for processing. This is very reasonable because in processing so many data or attributes, it takes a long time because they have to process each data one by one. The more data that is processed, the higher the computer specifications used. Therefore, we need a method used to reduce the amount of data randomly. The Sample Bootstrapping method is used to reduce the amount of training data to be processed [10]. With the reduced amount of data processed, the shorter the time required. In addition, the required computer specifications are also not as high as when the data is still intact or has not been reduced. To reduce the amount of training data to be processed, the Bootstrapping Method can be used [11]. Thus, one of the methods used to reduce the amount of data or attributes that are many randomly, can use the Bootstrapping Method. In addition, basically the bootstrap method is used to reduce standard errors. It can be seen the difference before and after bootstrapping.

Basically a bootstrap method, not as an error reducer. However, it is used to predict errors. Thus, the standard error (SE) will be obtained in the dataset. To estimate the standard error, you can use the following equation::

$$SE = \frac{\sigma}{\sqrt{n}} \quad (3)$$

σ = standard deviation

n = number of subjects

SE = standard error

So the larger the value of n, the smaller the error value.

The processed data after the bootstrap data still has missing values. Therefore, there are more steps needed to fix the missing data. In the dataset because there is missing data, there are several methods that can be used, including filling it back with minimum, average and maximum values. In this study the proposed method is to fill back with the average value of.

2.3. Missing Value Mean Imputation

Imputation is a way to solve the problem of missing values. Where the process is carried out by eliminating values that do not match the data set, looking for missing values in the data set by making estimates based on certain methods or in other ways. [12],[13]. There are several imputation techniques used, including listwise deletion, mean imputation and K-NN imputation. [12]. Mean imputation is one method of the average value of the existing values for each variable is calculated and the missing values for these variables and calculated with this average. There are many methods used for imputation such as average imputation to some more robust method based on the relationship between attributes [14]. This method works by calculating the mean or median of the non-missing values in a column and then replacing the missing values

in each column separately and independently of the others. It can only be used with numeric data types.

As for how to find out the mean or average value, you can use the equation below::

$$\text{Mean Value} = \frac{\text{number of values}}{\text{number of data}} \quad (4)$$

From the calculation of the average value or the mean of one of the ALB attributes, for example, before there was a value of 31.5, there were only values between 31, 39, 23 and 33. The average value was obtained from the sum of the values 31 + 39 + 23 + 33 then in for the number of available values before the missing value is filled, that is 5 records. After the data is imputed using the average value, it will continue with the process using the Naive Bayes kernel algorithm.

2.4. Naive Bayes Kernel

Naive Bayes kernel method was used to implement work on predictive data analysis, applying kernel based Naïve Bayes classifier to validate some lessons learned to predict disease probability. The kernel-based Naive Bayes algorithm implemented for the classification process produces higher accuracy than the classic Naive Bayes network [15],[16],[17].

The Naive Bayes classifier is described as a basic probabilistic classifier built on the application of Bayes' theorem along a set of assumptions specific to conditional independence. 'Mode of independent features' would be considered a clearer term describing the original probability model. Directly, the algorithm behind the work of the Naive Bayesian classification process assumes that the occurrence or non-occurrence of any distinct property of the provided class also known as a fixed attribute is unaffected by the occurrence (or non-occurrence) of any additional features present in the data.

When studying Bayesian network-based classification. Continuous variables are usually handled by discretization or assumed by a Gaussian distribution. In addition, nave Bayes added a Bayesian k-dependence classifier tree and a complete graph classifier adapted to the new kernel-based Baesian network pattern. Naïve Bayes plus flexible trees seem to have superior behavior for supervised classification [18].

Possible continuous attribute patterns for the naive Bayesian classifier can be approximated by kernel density estimates, letting each pattern influence the shape of the probability density resulting in an accurate estimate. KDE suffers from computational costs making it impractical in many real-world applications. Smooths the density using splines so that it requires fewer coefficients for estimation than the entire training set [19]. The possible continuous feature patterns required for probabilistic inference in a Bayesian network classifier can be calculated by kernel density estimation, letting each pattern influence the shape of the probability density. We smoothed the density using a spline so that it required estimation of only very few coefficients than the entire training set allowing fast implementation of BNC without compromising classifier accuracy. All rights reserved [20].

Nonparametric density estimation has wide application in computational finance especially in cases where high frequency data is available. Given the number of kernels estimating density, the current method takes time directly to sum kernels to perform a single density query. In on-line algorithms where points are constantly added to the density, the cumulative run time for the number of queries makes it very expensive, if not impractical, to calculate the density for large n. The run time for the density query is reduced to variable X or even time constant, depending on the selected kernel, and, accordingly, the cumulative run time

is reduced to X , respectively. Our results show that the MODE algorithm provides a dramatic advantage over a direct approach to density evaluation[21].

The main benefit of using the Naive Bayesian classification algorithm is that only a small amount of data is needed as training data to train the system to estimate the variance as well as the average of all the variables provided in the data set, which is an important condition for the classification process. It's just the variance of the variables for each individual label that needs to be evaluated and not the entire covariance matrix for the data provided as input to the system because all independent variables are grouped into classes. Nave Bayes net kernels can also be implemented on different numeric attributes with Naive Bayesian classifiers. With the rapid development of the Internet and the rapid development of big data analysis technology, data mining has played a positive role in promoting industry and academia. Classification is an important problem in data mining. According to the scale and characteristics of the data, different solution spaces are selected, and the solutions of the multiple problem are transformed to the original space classification surface to increase the speed of the algorithm. Research Process. The speed of the algorithm can be increased by transforming the solution of the multiple problem into the classification surface of the original space [22].

The weighting function used in the nonparametric estimation procedure is known as the kernel. Different kernels are implemented to process kernel density estimation techniques for estimation of the density function of random variables, or in kernel regression mechanisms to estimate conditional expectations of random variables in the dataset. These kernel density estimators belong to a special class of estimator functions known as nonparametric density estimators. Unlike parametric estimators where density estimators have a fixed functional form and these function parameters are the only information required to be stored, there is no fixed structure for nonparametric density estimators and they rely on data points provided in the data to reach estimates.

In this step, we perform two tasks. The first task is usually to build a machine learning model with the selected data set, which is called the training data set then the second task is to test the built model using another invisible data set, which is called the test data set. The proposed model used in our methodology is the KNB classifier. Actually, KNB is Nave Bayes with kernel density estimation (KDE). The following subsection provides an explanation of the proposed KNB.

Suppose X is the set of data values, $(x_1=x_1, x_2, \dots, x_N)$ and C is the topic set assuming Naive Bayes, the probability of a topic is C . Given the features x_1, x_2, \dots, x_N can be calculated by the following equation:

$$\begin{aligned}
 C &= \max_{c_j \in C} P(c_j | x_1 \dots x_n) & (5) \\
 &= \max_{c_j \in C} \frac{P(x_1, x_2, \dots, x_n | c_j) P(c_j)}{P(c_1, c_2, \dots, c_n)} \\
 &= \max_{c_j \in C} P(x_1, x_2 \dots x_n | c_j) P(c_j)
 \end{aligned}$$

The more common:

$$P(x_1, x_2, x_3 \dots x_n | c_j) = \prod_i P(x_i | c_j) \quad (6)$$

The probability, $P(x_i/c_j)$, that the feature value of a value equal to x if given the topic j (class j) equals c_j , was estimated using KDE from a training data set labeled (X, C) . KDE that is:

$$P(x_i|c_j) = \frac{1}{Nh} \sum_{y=1}^N \text{guKernel}(x_1, x_{vi}), \text{guKernel}(a, b) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(a-b)^2}{2h^2}} \quad (7)$$

One non-parametric way of estimating the population probability density function Probability), $P(x_i/c_j)$ is estimated using the Equation as above i.e. guKernel is a Gaussian function kernel with variance 1 and mean zero, N is the number of input data X belonging to class j where, c_j , x_{vi} are the word feature values at the i -th position of the v -th input $X = (x_{1i}, x_{2i} \dots x_{Ni})$ in class j , and h is the bandwidth, or smoothing parameter. To optimally estimate the conditional probability, h is optimized on the training dataset.

The process applied to predictions on a dataset includes the following steps:

1. Dataset retrieval.
2. Data pre-processing.
 - a. Bootstrap processing on dataset
 - b. Replace the missing value with the mean or average value.
 - c. Data transformation (scaling, conversion, and normalization).
3. Machine learning algorithm (kernel based Naïve Bayes classifier).
4. Prediction and calculation of performance

2.5. Dataset

The dataset used is HCV (Hepatitis C Virus) public data taken from the UCI repository. The dataset consists of 615 records, 12 attributes and 1 label. In the data, there is still data that is missing value, so it needs improvement so that the missing data can have a complete value.

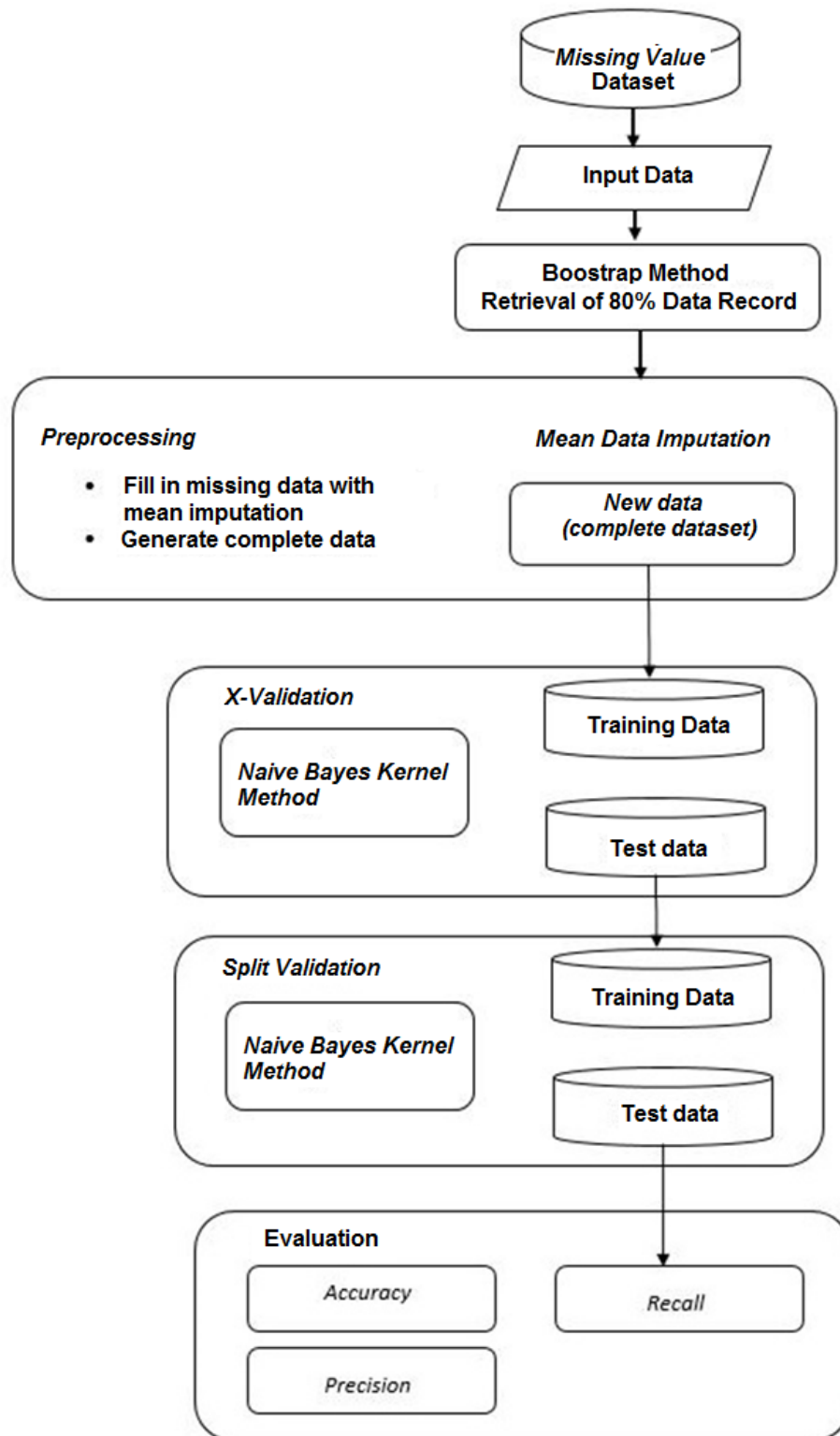


Figure 2. Flowchart of data processing and testing of the Naive Bayes kernel algorithm

3. RESULTS AND DISCUSSION

We took the HVC data set contained at UCI in 2021 data consisting of 615 records, 12 attributes and 1 label. In the data there are some data that is missing value. Therefore, special handling is needed so that no data is lost. In the trial of the missing data, we made improvements by filling in the data from the calculation of the average value for each attribute. However, before filling in the missing value data, we filter the original data using the bootstrap method which is used to select the data that has weight for the mean imputation process..

Table 2. Data that still has Missing Value

Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT	Category
0	31	86.3	5.4	95.4	117	1.57	3.51	60.5	53.6	68.5	5
0	38	102.9	27.3	243.2	15	5.38	4.88	72.3	400.3	73.4	5
0	NA	NA	40	54	13	7.5	NA	70	107	79	5
0	23	34.1	2.1	90.4	22	2.5	3.29	51	46.8	57.1	5
1	33	79	3.7	55.7	200	1.72	5.16	89.1	146.3	69.9	3

Table 3. The missing value data has been filled with the calculated value from the average

Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT	Category
0	31	86.3	5.4	95.4	117	1.57	3.51	60.5	53.6	68.5	5
0	39	102.9	27.3	143.2	15	5.38	4.88	72.3	400.3	73.4	5
0	31.5	75.575	40	54	13	7.5	4.21	70	107	79	5
0	23	34.1	2.1	90.4	22	2.5	3.29	51	46.8	57.1	5
1	33	79	3.7	55.7	200	1.72	5.16	89.1	146.3	69.9	5

The data that has been corrected with the new mean imputation is tested by using the Naive Bayes kernel algorithm. The differences between using the traditional naive bayes algorithm and the naive bayes kernel are as follows:

Table 4. differences in accuracy test results

Classifiers	X-Validation	Split Validation
Naive Bayes Kernel	96.05%	96.72%
Naive Bayes	89.92%	90.76%

4. CONCLUSION

In our research using the missing value HVC dataset, we can get outstanding results in the Naive Bayes kernel algorithm. Problems encountered in naive Bayes kernel processing can be solved by means of imputation solutions to fill in missing data values. The test results with X-Validation training modeling on the Naive Bayes kernel got an accuracy value of 96.05% and Split validation was 96.72% larger than the traditional Naive Bayes model, namely the X-Validation model was only 89.92% and Split Validation by 90.76% only.

REFERENCES

- [1] D. Khanna and A. Sharma, *Kernel-Based Naive Bayes Classifier for Medical Predictions*. Springer Singapore, 2018.
- [2] A. F. Sallaby, "Analysis of Missing Value Imputation Application with K-Nearest Neighbor (K-NN) Algorithm in Dataset," vol. 5, no. 2, pp. 141–144, 2021.
- [3] Eka Wulansari Fridayanthie, "Analisa Data Mining Untuk Prediksi Penyakit Hepatitis

- Dengan Menggunakan Metode Naive Bayes dan Support Vector Machine,” vol. 3, no. 1, p. 2015, 2015.
- [4] E. Siswanto, “Optimasi Metode Naïve Bayes Dalam Memprediksi Tingkat Kelulusan Mahasiswa Stekom Semarang,” *Jurikom*), vol. 6, no. 1, pp. 1–6, 2019.
- [5] 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.
- [6] Mawarsari, “Imputasi missing data dengan k-nearest neighbour dan algoritma genetika,” *J. Ilm. Pendidik. Mat. Ilmu Mat. dan Mat. Terap.*, vol. 6, no. 1, 2016.
- [7] Y. Dong and C. Y. J. Peng, “Principled missing data methods for researchers,” *Springerplus*, vol. 2, no. 1, pp. 1–17, 2013.
- [8] R. Sarmiento, E. Text, and M. Visualization, “Hepatitis C Records - A Complete Statistical Analysis,” no. January, 2021.
- [9] M. S. and V. K. T. Pang-Ning, “Introduction to data mining,” *Libr. Congr.*, 2006.
- [10] S. A. Setiawan. T. A., Wahono. R. S., “Integrasi Metode Sample Bootstrapping dan Weighted Principal Component Analysis untuk Meningkatkan Performa k Nearest Neighbor pada Dataset Besar,” *J. Intell. Syst.*, p. 796, 2015.
- [11] Sahibsingh A. Dudani, “The Distance-Weighted k-Nearest-Neighbor Rule,” *IEEE Trans. Syst. Man. Cybern.*, vol. SMC-6, pp. 325–327, 1976.
- [12] Q. Song, M. Shepperd, X. Chen, and J. Liu, “Can k-NN imputation improve the performance of C4.5 with small software project data sets? A comparative evaluation,” *J. Syst. Softw.*, vol. 81, no. 12, pp. 2361–2370, 2008.
- [13] 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.
- [14] G. E. A. P. A. Batista and M. C. Monard, “A study of k-nearest neighbour as an imputation method,” *Front. Artif. Intell. Appl.*, vol. 87, no. January, pp. 251–260, 2002.
- [15] M. Aladjem, “Projection pursuit mixture density estimation,” *IEEE Trans. Signal Process.*, vol. 53, pp. 4376–4383, 2005.
- [16] J. Bilmes, “A gentle tutorial on the EM algorithm and its application to parameter estimation for gaussian mixture models,” *Int. Comput. Sci. Inst.*, 1998.
- [17] Christopher M. Bishop, “Neural Networks for Pattern Recognition,” *Oxford Univ. Press. Inc.198 Madison Ave. New York, NYUnited States*, p. 482, 1995.
- [18] A. Pérez, P. Larrañaga, and I. Inza, “Bayesian classifiers based on kernel density estimation: Flexible classifiers,” *Int. J. Approx. Reason.*, vol. 50, no. 2, pp. 341–362, 2009.
- [19] Y. Gurwicz and B. Lerner, “Rapid Spline-based Kernel Density Estimation for Bayesian Networks,” pp. 5–8.
- [20] Y. Gurwicz and B. Lerner, “Bayesian network classification using spline-approximated kernel density estimation,” vol. 26, pp. 1761–1771, 2005.
- [21] C. G. Lambert, S. E. Harrington, C. R. Harvey, and A. Glodjo, “Efficient on-line nonparametric kernel density estimation,” *Algorithmica (New York)*, vol. 25, no. 1, pp. 37–57, 1999.
- [22] B. Gaye, D. Zhang, and A. Wulamu, “Improvement of Support Vector Machine Algorithm in Big Data Background,” *Math. Probl. Eng.*, vol. 2021, 2021.