K-Nearest Neighbor (K-NN) based Missing Data Imputation

1st Della Murbarani Prawidya Murti
Electrical Engineering Department
State University Of Malang
Malang, Indonesia
dellamurbarani4@gmail.com

2nd Aji Prasetya Wibawa
Electrical Engineering Department
State University Of Malang
Malang, Indonesia
aji.prasetya.ft@um.ac.id

3rd Muhammad Iqbal Akbar
Electrical Engineering Department
State University Of Malang
Malang, Indonesia
iqbal.elektro.um@gmail.com

4th Utomo Pujianto
Electrical Engineering Department
State University Of Malang
Malang, Indonesia
utomo.pujianto.ft@um.ac.id

***Abstract*— The performance of the classification algorithm depends on the quality of the training data. Data quality is an important factor that affects the data mining classification results. However, one problems that often found is missing data. Effect many missing data is a less optimal classification model. Because it is can deletes important information that affect the performance of the algorithm. One method used to recover missing data is to fill it, as known as imputation. This study uses the K-NN method as an imputation carried out in several cases that have different mechanisms and missing data model. On these imputed dataset then apply classification with Naive Bayes algorithm. In this study, analyzes the performance of imputation method using Naive Bayes algorithm on the basis of accuracy for handling missing data. The results, handling missing data with K-NN-based imputation can reach the accuracy of complete data in each case with a low accuracy difference.**

Keywords*— *Missing Data, Imputation, k-Nearest Neighbor, Naive Bayes Classifier

I. INTRODUCTION

The missing data values on certain attributes could be problematic. It may caused program errors, human error, missed questions in the questionnaire [1]. In the case of classification or data analysis can only work on complete data [2]. In other words, missing data may reduce the classification performance.

The simplest approach to handling this problem is remove records containing missing data [3]. However, this method is less effective because it can eliminate important information in data [4], [5]. Other weaknesses make bias values if using this approach [6].

Another method is the imputation. Imputation replaces missing data with an estimate of the correct value to obtain complete data, so that data analysis procedures that require complete data can be applied [7]. The main reason for imputation is to reduce the non-response bias that occurs because missing data so it's better than removing missing data [8].

The invented imputation techniques include mean [9], regression [10], hot deck [11], and Expectation Maximization (EM) [12]. The mean method is the easiest method to handling missing data by the average for a attribute [13]. But, sometimes the value from this method does not match with actual data so it can resulting an error estimate [14]. [15] shows that imputation with the mean method does produce parameters that are biased so this method is recommended to be used when the correlation attribute is low and the percentage of missing data less than 5%. Regression uses prediction models based on attribute relation to fill in missing data [16]. However, in some cases attribute relation are sometimes not linear, so it resulting in bias parameters [2]. The Hotdeck method replaces missing data with values from similar cases [11]. But, it will be repetition if there are many missing data so that it also bring to bias parameters [17]. EM method is used to predict maximum likelihood parameters from known data distribution, if there are missing data [12]. So the method requires large samples and the mechanism of data is lost randomly [13].

One of the developed the machine learning methods for handling missing data is K-NN. This method uses the distance between training data in classification [18]. K-NN is a flexible method in both continuous data and discrete data [19]. This method can be used as in multiple missing data filler [20], does not require a prediction model for each attribute [21]. [5] shows that non-parametric imputation (K-NN) is more efficient than parametric (EM) imputation for medium data size. While the disadvantage of the K-NN method is that time complexity is very high because this method looking for similar data in all datasets [22].

This study was made with a different cases. In the previous study [19], the correlation between attributes were neglected. In this study, the Journal Rankings dataset has a correlation. This study has a 80% inclompete dataset, which is much higher then the previous research [14]. Multi missing data in each attributes is also considered in this study. This may differ with the previous study [23], only used one attributes missing data. These various cases, are used to analyse the performance of K-NN as imputation method. The success factor is reached when the classification imputation using result is close to the baseline. The baseline is the Journal Rankings dataset.

II. METHODS

This research consists of five steps which shows the study steps. The steps is presented on the Figure I.

978-1-7281-2380-6/19/$31.00 ©2019 IEEE

1. Start
2. Collecting Dataset Journal Rankings,
3. Data Preprocessing,
4. Simulation of missing data,
5. Imputation with K-Nearest Neighbor Method,
6. Performance evaluation of K-Nearest Neighbor Imputation Method,
7. End

Figure I. Steps of the research methodology

In the first step data collection. The dataset is secondary data from one of the Journal Ranking sites, SCImago Journal Rank (SJR). The field is computer science range 2014 until 2017 consisting of 7191 instances. However, in this study using 8 attributes were shown in Table I. Attribute of the Best Quartile SJR is a label class attribute that has four journal quality classes Q1, Q2, Q3, and Q4.

TABLE I. DATASET ATTRIBUTE LIST

Attribute	Data Type	Range of Value
SJR Best Quartile	Nominal	(Q1, Q2, Q3, Q4)
H Index	Integer	(0-318)
Total Docs. (2017)	Interger	(0-20858)
Total Docs. (3 years)	Interger	(0-66063)
Total Refs	Integer	(0-415920)
Total Cites (3 years)	Integer	(0-58176)
Citable Docs. (3 years)	Integer	(0-61823)
Cites/Doc (2 years)	Real	(0-19.990)
Ref. / Doc.	Real	(0-269)

In the second step, data preprocessing is used because the data obtained is not included in the journal category, but distributed in other categories such as book series, conferences and procedures and trade journals. This study only focuses on data in the journal category. So the data that originally consisted of 7191 instances was reduced to 1441 instances.

The next step is simulation missing data. In this simulation made several cases which each case was formed with different mechanisms and model missing data. The process of removing data from each case was designed and then imputed using the K-NN method. In addition, the amount of data used is also different for each case where in the first case 1441 instance is used. While in the second to fourth cases the amount of data used was reduced to 200 instances. The 200 instances include 50 data in each of the quartile classes Q1, Q2, Q3, and Q4. These cases are as follows.

The first case begins by removing a small amount of data from the all data used. Removal mechanism is doing randomly where each quartile class has at least two missing data. Meanwhile, the missing data model was doing on all attributes alternately so that in this case there were 8 experiment based the number of attributes. Based the design, this case only remove five data where there is one missing data in Q1, Q2 and Q4 class . Whereas in Q3 class there are two missing data. The five data are 250, 500, 750, 1000 and 1250 data.

The next case the data removal is doing by one attribute. The selection of these attributes is based on the correlation value of each attribute to the class. Correlation to class can be doing through the SJR attribute. Because the SJR attribute is an indicator that determines the quartile class. Based on the rules of the correlation coefficient [24] obtained the highest and lowest correlation coefficients are shown in Table II.

TABLE II. DATASET CORRELATIONS WITH CLASS

Attribute	Correlation Coefficient	Description
Cites / Doc. (2years)	0,754118	Strong Correlation
Cites / Doc. (2years)	0,754118	Weak Correlation

In the second case, the attribute used has the highest correlation coefficient. Based on Table II, the attribute is Cites / Doc. (2 years) with a correlation coefficient 0,754118. After the attribute is determined as data experiment, then determined mechanism of missing data. Data removal was doing by 20 data consisting of 5 data for each class Q1, Q2, Q3, and Q4. While the missing data model in this case were doing sequentially. The experiment was doing 10 times according to the amount of data used.

In the third case data removal is doing on two attributes. The selection of the two attributes is based on the value of the correlation between attributes with other attributes. Based on the rules of the same correlation coefficient can be known attributes that have a relationship or not based on the highest and lowest correlation values as shown in Table III.

TABLE III. DATASET CORRELATIONS

Attribute	Correlation Coefficient	Description
Total Docs. (3years) dan Citable Docs. (3years)	0,999	Correlation
Ref./ Doc dan Total Docs. (2017)	0,003	Not Correlation

Based on Table III, the attribute that has the highest correlation coefficient is the Total Docs (3years) and Citable Docs. (3 years) with a correlation value of 0.999. After the attribute is determined as experiment data, then made the missing data mechanism. Data removal is doing on both attributes where each attribute has 20 missing data consisting of 5 data for each class Q1, Q2, Q3, and Q4. While the model of data removal is doing alternately. The experiment was doing 5 times according to the amount of data used.

The fourth case of data removal is doing on all attributes. The different with the first case where the model of missing data is doing by removing the data on each attribute alternately. In this case, missing data is doing periodically where one by one each attribute will be missing. While the model of data removal is alternately for each attribute. The amount of data that was removed is 20 data consisting of 5 data for each class Q1, Q2, Q3, and Q4.The experiment was doing 8 times according to the number of attributes.

Based on all of these cases, the K-NN method can applied as an imputation method. The procedure for the K-NN method is as follows [24]:

1) Determine the parameter k
In this study used parameters k = 1,3,5,7. There is no specific method, in determining the k value for the K-NN method. If the value of k is too small, there will be a lot of noise which reduces the level of accuracy in the classification, but if it is too large it can also cause errors in limiting the value taken and indirectly affecting the accuracy [25].

2) Calculate the Euclidian distance between missing data instances and complete data by equation (1) [26].

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

Where :
$d_{(x,y)}$: Euclidian distance,
j : data attribute with *j*=1,2,3,.....*s*
s : data dimensions
x_{aj} : value from *j*- attribute containing missing data,
y_{bj} : value from other *j*- attribute containing complete data,

3) Based on the distance information obtained the minimum Euclidian distance based on the parameter k that has been determined as the estimated value in the missing data. The value imputation is calculated using Weight Mean Estimation in equation (2) [27].

$$x_j = \frac{\sum_{k=1}^{K} w_k v_k}{\sum_{k=1}^{K} w_k} \qquad (2)$$

Where :
x_j :Weight Mean Estimation,
K :number of parameters k used with k=1,2,3,....K
w_k :nearest neighbor observation value K
v_k :value in complete data on attribute containing missing data based on parameter k

Here, the formula of w_k is calculated in equation (3).

$$w_k = \frac{1}{d_{(x,y)}{}^2}, \qquad (3)$$

Where :
$d_{(x,y)}$: Euclidian distance on each parameter k,

To measure the estimated value of good or bad by imputation method, so this study doing an analysis based on accuracy that was get by many classification methods [28]. The chosen classification method is Naive Bayes. Because this method has been used in the Journal Rankings classification with the accuracy of 71.60% [29].

Quality measurement of a classification method is built using confusion matrix [30]. Testing is done using 10-fold cross validation because 10-fold is often used in research in the use of confusion matrix [31]. To get the of accuracy values, Weka tools can be used in several machine learning techniques, one of which is classification [32]. Meanwhile, the evaluation is also doing by comparing the accuracy value from the imputation data with the complete data. The resulting accuracy difference can help find out an imputation error.

III. Result and Discussion

Based on simulations missing data, the results each case shown in Figures 2 - 5. The accuracy value shown consist the average of each experiment and all the k parameters used. Figure 2 shows the accuracy of the imputation results for each attribute.

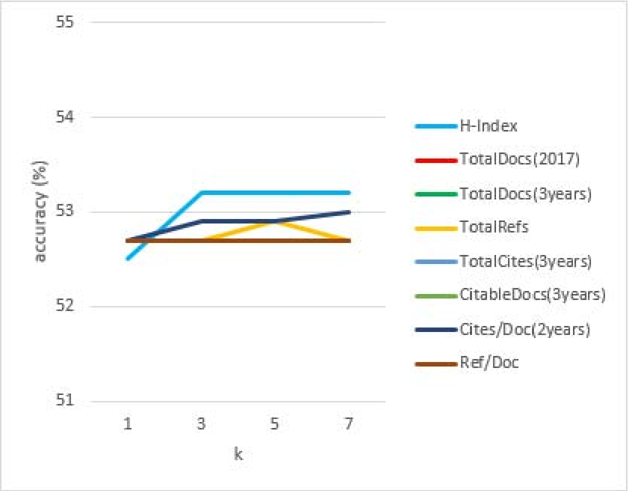

Figure 2. Cases 1

From Figure 2, the attributes that increase the accuracy value after imputation, this is H-Index attribute and the Cites / Docs attribute (2years) with an increase in accuracy of 0.3%. While the other attributes tend not to increase or constant at an accuracy of 52.7%. This shows that the removal of one attribute which is then imputated does not affect the classification results. With this case, the case doing by [23] did not prove effective. However, this may also occur because the dataset and the characteristics of the data used are different. So that the implementation of the case needs to be reviewed. While Figure 3 shows the accuracy of the imputation results on the one attributes of the class.

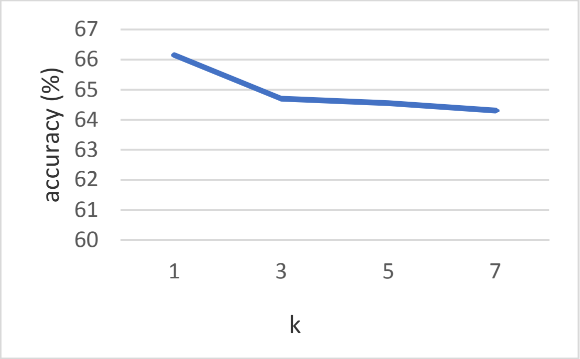

Figure 3. Cases 2

The graph in Figure 3, shows different accuracy for each k parameter used. In the parameter k = 1 the average accuracy value is 66,15%. While the accuracy of the other k parameters decreases with the lowest accuracy of 64,3% in the parameter k = 7. Based on the accuracy from imputation, in this case imputation uses the K-NN method with all k parameters result less optimal accuracy. The decrease in accuracy can be caused the dataset used has a high variance. This can affect the results of imputation and classification so that there are differences in the results of the accuracy of each parameter k. The difference in the accuracy value of each experiment depends on the missing data. Figure 4 shows the accuracy of imputation results on two correlated attributes.

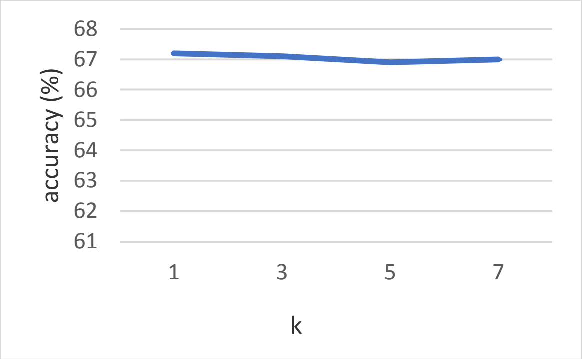

Figure 4. Cases 3

Based on Figure 4 the accuracy from the imputation using the K-NN method, does not show much difference. The results of the classification using the Naive Bayes algorithm shows the data removal on the two correlated attributes give not effect when compared to the overall data. Different when data removal occurs on one attribute that is correlated with class. This can occur because the attribute has strong correlated so if the value of one attribute has increasing the other value of will also increase. So, through the process in the case that can be found in the K-NN method as an imputation that is able to handling the missing data that has a correlate between attributes. Figure 5 shows the results of the attribute results on the attribute.

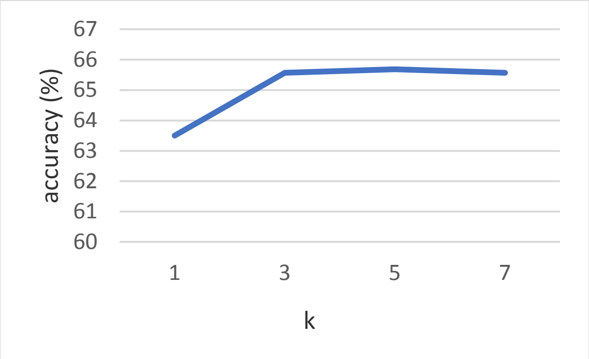

Figure 5. Cases 4

The graph in Figure 5 shows a decrease in the accuracy value from each parameter k when compared with complete data. The decrease reason is the availability of complete data which reduced along with the increasing number of missing data for each experiment. Even during the data filling process in experiment 8, the number of attributes used was also reduced because missing values in the same data. As a result of the reduction in the amount of complete data and the number of attributes, this results in a decrease in accuracy. Based on Figure 5 it can also be seen that the data removal on all attributes can affects the classification accuracy value.

The differences each experiment and the k parameters used are to missing data model. Because the characteristics of datasets and missing data model, it can also affect the performance of imputation and classification methods [21], [34]. Therefore, consideration needs to be taken in determining the model of missing data. Meanwhile the dataset used has a high data variant. This can affect the of Euclidian and the estimated value obtained. The K-NN does not have criteria for choosing the best k value so that the best k value is obtained empirically [35].

The difference in the accuracy of each parameter k used because to a high dataset variation. This means that the Journal Rangkings dataset has data with a large range between attributes in one observation. So it will affect the acquisition of Euclidian distance and the estimated value obtained. In addition, based on the conclusion of the second case, the results of accuracy for each of the parameters k are influenced by the results of imputation. However, it maybe different in other cases to the same reason when there is a difference in accuracy for each parameter k used. Because of the characteristics of datasets and the model of missing data, it can also affect the performance of imputation and classification methods so that it needs to be considered in data analysis [21], [33]. Meanwhile, K-NN method does not have theoretical criteria to choose the best k value so that the best k value is obtained empirically [34].

The best accuracy results from each case then compared with the classification accuracy values using complete data. From these comparasion, it can be known the difference in accuracy. A comparison of the two accuracy values is shown in Table IV.

TABLE IV. ACCURACY COMPARISON

Case	Accuracy (%)		
	Complete Data	Treated Data	Difference
1	52,7	52,8	0,1
2	67	66,15	0,85
3	67	67,2	0,2
4	67	65,59	1,41
Average			0,64

Based on Table IV, it can be concluded that using the K-NN method can reach complete data with a low average accuracy difference . In the 4th case the resulting difference in accuracy is higher. This happens when the 4th accuracy value decreases when the number of missing data increases. The reduction in complete data because the imputation value on the attribute or previous experiment cannot be used in the next data filling process because this will affect the estimation results. Estimates are based on actual data. According to [21] the efficiency of the K-NN method as a depends on the number of available complete data in the data set.

Thus, high resistance to the mechanism and model of missing data was makes the K-NN method as an one approach to handling missing data. This resistance is also seen by imputation using different k parameters.

IV. CONCLUSION

Based on the results and discussion of this study, it can be concluded that the K-NN imputation method is able to be one of the approaches that can be used when the dataset missing data. This can be seen from the performance of the K-NN method which is measured based on the accuracy value where the accuracy of the imputation results approaches the accuracy of the complete data with different missing data model. Meanwhile, the resulting accuracy difference shows that the K-NN method as an imputation can be applied in the case in this study.

In future studies, it is recommended to consider missing data model such as using of tools that can generate random missing values so that data removal is not done manually. Using another dataset with a low data variant can also be used as a comparison to method used determine the stability.

REFERENCES

[1] J. Kaiser, "Dealing with Missing Values in Data," *J. Syst. Integr.*, pp. 42–51, 2014.

[2] D. Umamaheswari and N. Shyamala, "Data Mining Techniques to Fill the Missing Data and Detecting Patterns," *Int. J. Sci. Technol. Eng.*, vol. 2, no. 1, pp. 39–42, 2015.

[3] P. D. Allison, "Missing Data Techniques for Structural Equation Modeling," *J. Abnorm. Psychol.*, vol. 112, no. 4, pp. 545–557, 2003.

[4] A. Izzah and N. Hayatin, "Imputasi Missing data Menggunakan Algoritma Pengelompokan Data K-Harmonic Means," *Semin. Nas. Mat. dan Apl.*, 2013.

[5] S. Rawal, S. . Gupta, and S. Singh, "Predicting Missing Values in a Dataset : Challenges and Approaches," *Int. J. Recent Res. Asp.*, vol. 4, no. 3, pp. 34–38, 2017.

[6] G. King, J. Honaker, A. Joseph, and K. Scheve, "Analyzing Incomplete Political Science Data : An Alternative Algorithm for Multiple Imputation," *Am. Polit. Sci. Rev.*, vol. 95, no. 1, pp. 49–69, 2001.

[7] M. Pampaka, G. Hutcheson, and J. Williams, "Handling missing data : analysis of a challenging data set using multiple imputation," *Int. J. Res. Method Educ.*, vol. 39, pp. 19–37, 2016.

[8] G. B. Durrant, "Imputation methods for handling item-nonresponse in practice: methodological issues and recent debates," *Int. J. Soc. Res. Methodol.*, vol. 12, no. 4, pp. 293–304, 2009.

[9] S. Martha and A. K-means, "Perbandingan Imputasi Missing Data Menggunakan Metode Mean dan Metode Algoritma K-Means," *Bul. Ilm. Mat Stat dan Ter.*, vol. 4, no. 3, pp. 305–312, 2015.

[10] S. Kim, A. Sugar, and T. R. Belin, "Evaluating model-based imputation methods for missing covariates in regression models with interactions," *Stat. Med.*, no. January, 2015.

[11] R. R. Andridge and R. J. A. Little, "A Review of Hot Deck Imputation for Survey Non-response," *Int. Stat. Rev.*, pp. 40–64, 2010.

[12] J. Sari and R. Yendra, "Teknik Mengatasi Data Hilang dengan Metode Algoritma EM," *J. Sains Mat. dan Stat.*, vol. 3, no. 1, pp. 70–74, 2017.

[13] H. Entin, "Classification Of Missing Values Handling Method During Data Mining : Review," *Sigma Epsil.*, vol. 21, no. 2, 2017.

[14] E. Hartini, "Implementation of missing values handling method for evaluating the system/component maintenance historical data," *J. Teknol. Reakt. Nukl.*, vol. 19, no. 1, pp. 11–18, 2018.

[15] P. Lodder, "To Impute or not Impute : That's the Question," *Pap. Methodol. Advice*, pp. 1–7, 2013.

[16] A. N. Baraldi and C. K. Enders, "An introduction to modern missing data analyses," *J. Sch. Psychol.*, vol. 48, no. 1, pp. 5–37, 2010.

[17] K. Pada, H. Deck, and H. Deck, "Kajian Metode Imputasi Dalam Menangani Missing Data," *Pros. Semin. Nas. Mat. dan Pendidik. Mat. UMS*, pp. 637–642, 2015.

[18] J. Maillo, S. Ram, I. Triguero, and F. Herrera, "kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors Classifier for Big Data," *Knowledge-Based Syst.*, 2016.

[19] G. E. A. P. . Batista and M. C. Monard, "A Study of K -Nearest Neighbour as an Imputation Method," *Soft Comput. Syst. - Des. Manag. Appl.*, 2002.

[20] B. Suthar, H. Patel, and A. Goswami, "A Survey : Classification of Imputation Methods in Data Mining," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 2, no. 1, 2012.

[21] D. Priya and S. Sivaraj, "A Review Of Missing Data Handling Methods," *Int. J. Eng. Technol. Sci.*, vol. 2, no. 2, pp. 58–68, 2015.

[22] E. Acuna and C. Rodriguez, "The treatment of missing values and its effect in the classifier accuracy." p. 9, 2004.

[23] S. Susanti, S. Martha, and E. Sulistianingsih, "K-Nearest Neigbor Dalam Imputasi Missing Data," *Bul. Ilm. Mat Stat dan Ter.*, vol. 7, no. 1, pp. 9–14, 2018.

[24] M. Minakshi, R. Vohra, and G. Gimpy, "Missing Value Imputation in Multi Attribute Data Set," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 4, pp. 5315–5321, 2014.

[25] N. Dany and O. Setyawati, "Perbaikan Missing Value Menggunakan Pendekatan Korelasi Pada Metode K-Nearest Neighbor," *J. Inform.*, vol. 9, no. 3, 2017.

[26] M. Bazmara and S. Jafari, "K Nearest Neighbor Algorithm for Finding Soccer Talent," *J. Basic Appl. Sci. Res.*, no. April 2013, 2014.

[27] W. Ling and F. D. Mei, "Estimation of Missing Values Using a Weighted K-Nearest Neighbors Algorithm," *Int. Conf. Environ. Sci. Inf. Appl. Technol.*, no. 2, pp. 660–663, 2009.

[28] J. Luengo, S. García, and F. Herrera, "On the choice of the best imputation methods for missing values considering three groups of classification methods," *Knowl. Inf. Syst.*, vol. 32, pp. 77–108, 2012.

[29] A. P. Wibawa *et al.*, "Naive Bayes Classifier for Journal Quartile Classification," *Int. J. Recent Contrib. from Eng. Sci. IT*, vol. 7, no. 2, 2019.

[30] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond Accuracy , F-Score and ROC : A Family of Discriminant Measures for Performance Evaluation," *Conf. Artifical Intell.*, pp. 1015–1021, 2006.

[31] D. Li, Y. Fang, and Y. M. F. Fang, “The data complexity index to construct an efficient cross-validation method,” *Decis. Support Syst.*, vol. 50, no. 1, pp. 93–102, 2010.

[32] S. B. Jagtap, “Census Data Mining and Data Analysis using WEKA,” *Int. Conf. Emerg. Trends Sci. Technol. Manag.*, vol. 2013, pp. 35–40, 2013.

[33] A. W. Liew, N. Law, and H. Yan, “Missing value imputation for gene expression data : computational techniques to recover missing data from available information,” *Brief. Bioinform.*, vol. 12, no. 5, pp. 498–513, 2010.

[34] J. Luengo, S. Garcia, and F. Herrera, *Imputation of Missing Values Methods ’ Description*. .