New Method Based Pre-Processing to Tackle Missing and High Dimensional Data of CRISP-DM Approach

Joko Suntoro¹*, Ahmad Ilham², Handini Arga Damar Rani³
¹Informatics Engineering, Faculty of Technology Information and Communication, Universitas Semarang, Semarang, Indonesia
²Department of Informatics, Universitas Muhammadiyah Semarang, Semarang, Indonesia
³Informatics Education, Faculty of Science and Technology, IVET University, Semarang, Indonesia

Email: jokosuntoro@usm.ac.id

Abstract. The kidneys are one of the most important organs including the excretion system in humans. The kidneys are responsible for maintaining blood concentrations to remain constant (homeostatic) and help to control blood pressure (BP). If the task of the kidney is not functioning properly it will cause kidney failure. In the past decade, data mining methods have been used to diagnose kidney failure. The dataset used to predict kidney failure was successfully summarized by Soundarapandian, and was named the Chronic Kidney Disease (CKD) dataset. But the data in the CKD dataset contains missing value and high dimension data (original data) so that it affects the evaluation results on classification. This research proposes methods in preprocessing data, namely modus in every class (MEC) method to solve missing value problems, and the weight information gain (WIG) method for solving high dimensional data problems, the proposed method is named the MEC + WIG method. The MEC + WIG method will be compared with the original method and the MEC method and evaluated based on the accuracy of the traditional classification method (k-NN, Naïve Bayes, C4.5, and CART). The results showed that the average accuracy of the MEC + WIG method was better than the original method and the MEC method, with the average accuracy of the MEC + WIG method at 98.13%, while the average value of the accuracy of the original method and MEC respectively amounting to 88.56% and 92.88%. There were significant differences between the three methods when tested using Friedman test with a p-value of 0.02. It can be concluded that the MEC + WIG method can improve the performance of traditional methods k-NN, Naïve Bayes, C4.5 and CART by overcoming the problem of missing value and data high dimension.

1. Introduction
The kidneys are one of the most important organs including the excretion system in humans. The kidneys play a role in the metabolism of the human body, which is responsible for maintaining blood concentrations to remain constant (homeostatic) and help to control blood pressure (BP) [1]. In addition, the acid and base balance is also controlled through the kidney organs by removing ammonia, creatine, and urea. If the task of the kidney is not functioning properly it will cause kidney failure. Total kidney failure is caused by progressive loss of kidney function [2].
In the past few decades, research in the healthcare field has applied data mining methods to diagnose diseases [3]. A large amount of data is collected by researchers in the healthcare field to find hidden information, then from that information is used to predict of disease diagnosis and is used for effective decision making[4]. Research conducted by [3]–[5] for two months successfully collected a dataset used for predicting kidney failure, the dataset was named the Chronic Kidney Disease (CKD) dataset.

The data in the CKD dataset contains missing value. Missing value is defined as a value that is missing or not available on an object [6]. The causes of missing values are equipment errors, human errors, and incorrect data transmission [7]. Missing value is a problem in classification, because some classification methods in data mining require prediction on the missing value to produce more accurate evaluation results [8]. CKD datasets also contain high dimensional data. In addition to missing values, CKD datasets also contain high-dimensional data. High-dimensional data is defined as a dataset that has a large sample size or number of attributes [9]. Data with a large number of attributes causes classification performance to be low [10], [11].

Several methods have been proposed and published to overcome data containing missing value and high dimension data, the method applied is data preprocessing[12]–[14]. According to Han[15], the data preprocessing technique can be used to solve the problem of data containing missing value using the data cleaning method, while the data reduction problem is used to solve high-dimensional data problems.

Data cleaning method has been used to overcome data containing missing value [16]. This method works by removing lines containing missing value data, filling in data manually (random), and filling in data automatically (based on the mean value of the sample with the same class where the sample with the same class is based on the probability value). The data reduction method has also been used to deal with high-dimensional data with a feature selection[17]. Attributes with large size will be reduced based on the highest attribute weights [18]. The feature selection method is divided into 3 (three) approaches, namely filter, wrapper, and embedded. Filter techniques use relevance between attributes based on the intrinsic nature of the data, the wrapper technique selects attributes based on classifier performance evaluations [19] while embedded techniques are often given hybrid names by combining filter techniques and wrapper techniques.

The main contributions in this research are: (1) CKD datasets containing missing values in the CKD dataset will be overcome using data cleaning techniques using the Modus in Every Class (MEC) method. (2) Next, use the filter method in the feature selection to deal with high-dimensional data. The filter method used is the attribute selection method based on the information gain weight value. The research question is how well the standard classification methods work when integrated with the MEC and WIG methods in the preprocessing stage.

This paper is organized as follows. In section 2, Data gathering, proposed method and theory implementation of k-NN, NB, C4.5, and CART method are explained. The experimental results of comparing the proposed method with others are presented in section 3. Finally, our works of this paper is concluded in the last section.

2. Material and methods
2.1. Data gathering
The dataset used to evaluate the proposed method is the same as that used by [14]. This dataset is collected from Apollo Hospital in India. The detail source can refer to here [14]. Also, the CKD dataset can be downloaded from the UCI machine learning repository database. The CKD dataset consists of 25 attributes (11 numeric and 14 nominal) and 400 data records. The CKD dataset has two
class labels used for diagnostic activities: (1) non-CKD (250) and (2) CKD (159). Information data used in this research, can be seen in Table 1

| No | Attribute name | Description      | Data type   | Value                        | Missing |
|----|----------------|------------------|-------------|------------------------------|---------|
| 1  | age            | usiadalamtahun   | numeric     | 2 – 90 tahun                 | 9       |
| 2  | bp             | blood pressure   | numeric     | 50 – 180 mm/Hg               | 12      |
| 3  | sg             | specific gravity | polynomial  | 1.005, 1.010, 1.015, 1.020, 1.025 | 47      |
| 4  | al             | Albumin          | polynomial  | 0,1,2,3,4,5                  | 46      |
| 5  | su             | Sugar            | polynomial  | 0,1,2,3,4,5                  | 49      |
| 6  | rbc            | red blood cells  | binominal   | normal, abnormal             | 152     |
| 7  | pc             | pus cell         | binominal   | normal, abnormal             | 65      |
| 8  | pcc            | pus cell clumps  | binominal   | present, notpresent          | 4       |
| 9  | ba             | Bacteria         | binominal   | present, notpresent          | -       |
| 10 | bgr            | blood glucose    | numerik     | 22 – 490 mgs/dl              | 44      |
| 11 | bu             | blood urea       | numerik     | 1.5 – 391 mgs/dl             | 19      |
| 12 | sc             | serum creatinine | numerik     | 0.4 – 76 mgs/dl              | 17      |
| 13 | sod            | Sodium           | numerik     | 4.5 – 163 mEq/L              | 87      |
| 14 | pot            | potassium        | numerik     | 2.5 – 47 mEq/L               | 88      |
| 15 | hemo           | hemoglobin       | numerik     | 3.1 – 17.800 gms             | 52      |
| 16 | pcv            | packed cell volume| numerik    | 9 – 54                       | 71      |
| 17 | wbcc           | white blood cell count | numerik | 2.200 – 26.400 cells/cumm   | 106     |
| 18 | rbcc           | red blood cell count | numerik | 2.1 – 8 millions/cmm        | 131     |
| 19 | htn            | hypertension     | binominal   | yes, no                      | 2       |
| 20 | dm             | diabetes mellitus| binominal   | yes, no                      | 2       |
| 21 | cad            | coronary artery disease | binominal | yes, no                      | 2       |
| 22 | apet           | Appetite         | binominal   | good, poor                   | 1       |
| 23 | pe             | pedal edema      | binominal   | yes, no                      | 1       |
| 24 | ane            | Anemia           | binominal   | yes, no                      | 1       |
| 25 | class          | Class            | binominal   | ckd, notckd                  | -       |

2.2. Proposed method

Figure 1, shows the method we proposed. As mentioned earlier, missing value problems can be overcome by data cleaning techniques to eliminate noise and correct inconsistencies in data [20], which can improve classification performance [13]. The method used is modus in every class (MEC). This technique was introduced by [21] to overcome the missing value problem in the data because of possible human error factors. MEC define as Equation (1).

\[
\text{mod } us = t b + \left( \frac{D F_1}{D F_1 + F_2} \right) p 
\]

Secondly, CKD dataset has many attributes and some of its attributes are irrelevant, so it needs to be removed [22] or minimized [12] based on the highest weight [16]. Weight Information Gain (WIG) is applied to minimize the attribute information needed for classification. The WIG method was introduced by [22] to overcome high-dimensional data. The calculation of the WIG method can be seen in Equation (2) and (3). The proposed method is named MEC+WIG.

\[
Entropy(S) = \sum_{i=1}^{n} -p_i \times \log_2 p_i
\]
\[ WIG(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times Entropy(S_i) \] (3)

Figure 1. Framework of the proposed method

### 2.3. Classification methods for testing our proposed

The proposed methods will combine and compared with the standard classification methods such as kNN [23], Naive Bayes[24], [25], C4.5[23], and CART [12], [26]. According by [27] the four methods have been included in one of the top 10 methods used in data mining.
2.4. Model validation & performance measure
We use 10 Cross Validation for learning and testing data where we divide the training data into 10 equal parts and then do the learning process 10 times. Every time, we select another dataset for testing and use the remaining nine parts to study. After that, we calculate the average value and the deviation value of ten different test results. We use 10-fold cross validation, because this method has become a standard and sophisticated validation method in practical terms [28]. Some tests also show that the use of stratification slightly increases the performance of classifiers [12]. In the results of this experiment, we evaluate the performance of classifiers using accuracy, see Equation (4).

\[
\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}
\]  

(4)

3. Results and Discussion
3.1. Deep frameworks & configuration setting
Data on CKD datasets containing missing value (MV) and high dimension data (HDD) will be solved by data preprocessing method. Data containing missing value will be filled with values based on the calculation of the mode on each attribute of each class (mode in every class). After the missing value data is filled with data, the next step is to select the highest-weighted attribute with a number of top-10, based on the calculation of weight information gain (WIG). This method is named the MEC + WIG method, Fig. 1, shows the proposed method framework, namely the MEC + WIG method. The experiment was conducted using a computing platform based on the Intel Celeron 2 Duo 2.2 GHz CPU, 8 GB RAM, and Microsoft Windows 1064-bit as the operating system and RapidMiner version 8 as a data analysis tool. The experiment was carried out using the auxiliary program RapidMiner version 7.4, Microsoft Excel 2013 and XLSTAT 2016. Rapid miner will produce model performance as a result of calculating accuracy.

3.2. Experiment & method evaluation
First of all, we conducted an experiment using only the classical method without preprocessing or original method and MEC [21] method combine with classification methods such as kNN, NB, C4.5 and CART without proposed method. The experimental results, taken from the results of RapidMiner calculations, are shown in Table 2. To accuracy as shown in Table 2, the accuracy method by [21] are 73.75, 99.50, 99.50, and 98.75 respectively. The accuracy of this classification is very good because has a good accuracy average is 92.88% rather than the original method (combination classical method without preprocessing). Second experiment, we apply the proposed method where we combine MEC+WIG. MEC method applied to handle missing data and Weight Information Gain (WIG) applied to minimize the required attribute information. The original and MEC methods will compare with the proposed method. As shown in Table 2, the proposed method is superior rather than another method is 94.50, 99.75, 99.50, and 98.75 respectively. The accuracy of the proposed method is excellent accuracy in all of the classical method except kNN and NB. However, we cannot say the proposed method is the best implemented because NB and kNN can still be improved by choosing the right weighting method. Figure 2, shows a diagram comparison of accuracy in the original method, MEC, and MEC+WIG. After calculating the value of accuracy in each method, a validation test is needed to know significant differences between methods. Before validity test is carried out, a normality test is conducted. In this study, data normalization test used is Shapiro-Wilk. The value of α used is 0.05. If the Shapiro-Wilk p-value one method is less than the value of α (0.05), it is concluded that the data is not normally distributed, but vice versa, if the Shapiro-Wilk p-value values are more than the value of α (0.05), it is concluded that the data is normally distributed. Table 3, shows the p-value in the Shapiro-Wilk test of all methods. The result is p-value by an original method and MEC is less than the value of α, so it can be concluded that the data on the accuracy of Table 3 are normally distributed if data is not normally distributed, a nonparametric validation test is used [29]. Nonparametric validation test used was Friedman test [29]. Friedman test was used to determine significant differences in more than two methods [30]. In this study the value of α was set at 0.05, if
the p-value is greater than the value of α, there is no significant difference between the models. Whereas if the p-value is smaller than the value of α then there are significant differences between the models. Table 4 shows the results of the Friedman test in this study, the p-value in this study is 0.024 (less than the value of α), so it can be concluded that there are significant differences between the original method, MEC, and MEC+WIG.

**Table 2.** Comparison Summary of Original Method, MEC method with Proposed Method as a pre-processing method in each classification methods based on accuracy

| Classic  | Without proposed | Proposed |
|----------|------------------|----------|
|          | Original | MEC     | MEC+WIG |
| kNN (k=10) | 61,00   | 73,75   | 94,50   |
| NB       | 99,25   | 99,50   | 99,75   |
| C4.5     | 96,50   | 99,50   | 99,50   |
| Chart    | 97,50   | 98,75   | 98,75   |
| Average  | 88,56   | 92,88   | 98,13   |

**Figure 2.** Comparison diagram of the average value on accuracy for each classification method combined with pre-processing methods.

**Table 3.** Summary of p-value by Shapiro-Wilk test for all method.

| Method   | Shapiro-Wilk | Information |
|----------|--------------|-------------|
|          | p-value      | p-value     |
| Original | 0.008        | <α          |
| MEC      | 0.003        | <α          |
| MEC+WIG  | 0.060        | >α          |

*α = 0.05

**Table 4.** Summary of Friedman test for all method.

| Q (Observed value) | 7,429 |
|--------------------|-------|
| Q (Critical value) | 5,991 |
| DF                 | 2     |
| p-value            | 0.024 |
| alpha              | 0.050 |
In the Friedman test, it was concluded that there were significant differences between the models. The next step is to do the Bonferroni-Dunn test [31], [32] to compare the MEC + WIG method with original method and MEC method. In general, the performance of the two classifications differs significantly if the average rank value (AR) is smaller than the critical difference (CD) value. Table 4 shows the average rank rating of the proposed method (MEC + WIG) with other methods. The calculation of CD values is based on Equation (5), with values \(K = 3\), \(N = 4\), and \(q_a = 2.241\), so the CD value in this study is 1,584.

\[
CD = q_a \sqrt{\frac{k(k+1)}{6N}}
\]

(5)

After the AR rank and CD value are known, the next step is to create an average rank difference table between the proposed method (MEC + WIG) and the original method and the MEC method. The Table 6, shows the differences in the average rank method. It is seen that the MEC + WIG method differs significantly from the original method, because the difference value is greater than the CD value (1,750 > 1,584), while the MEC + WIG method cannot be said to be significant with the MEC method, because the difference value is smaller than the CD value (0.500 < 1,584).

The average value of accuracy in the MEC + WIG method is greater than the original method and the MEC method. This answers the research question, that the MEC + WIG method produces good evaluation results for handling CKD datasets containing missing value and high dimension data. At the deployment stage, a knowledge-based application will be developed for the diagnosis of kidney failure. The dataset used in the deployment stage is the CKD dataset that has been preprocessed with the MEC + WIG method. The programming language used in the deployment stage is the PHP programming language. Source code can be downloaded via the link https://github.com/jokosuntoro/ckd_mec_wig.

Table 5. Rank average of each pre-processing method in each classification methods based on accuracy.

| Classical M | Original | MEC | MEC+WIG |
|------------|----------|-----|---------|
| k-NN       | 3        | 2   | 1       |
| Naïve Bayes| 3        | 2   | 1       |
| C4.5       | 3        | 1.5 | 1.5     |
| CART       | 3        | 1.5 | 1.5     |
| Avg. Rank  | 3        | 1.75| 1.25    |

Table 6. Summary of rank average based on accuracy and Friedman test of all method.

| Classical M | Average Rank | Difference |
|------------|--------------|------------|
| MEC+WIG    | 1,250        | 0          |
| MEC        | 1,750        | 0,500      |
| Original   | 3            | 1,750      |

4. Conclusion

The results showed that the average accuracy of the MEC + WIG method was better than the original method and the MEC method, with the average accuracy of the MEC + WIG method at 98.13%, while the average value of the accuracy of the original method and MEC respectively amounting to 88.56% and 92.88%. There were significant differences between the three methods when tested using Friedman test with a p-value of 0.02. This study has contributed to the handling of missing values, data containing missing values will be filled with the mode values for each attribute of each class.
Moreover, our work also contributes to attribute selection, namely by selecting the attribute with the highest weight of a number of top-10 based on calculating weight information gain.

Acknowledgment
We would like to thank the Intelligent Systems Research Group for a warm discussion about this study and fellow lecturers at the Informatics Engineering at Universitas Semarang, Department of Informatics at Universitas Muhammadiyah Semarang, and fellow lecturers at Informatics Education at IVET University.

References
[1] S. Murabito and B. F. Hallmark, “Complications of Kidney Disease,” Nurs. Clin. North Am., vol. 53, no. 4, pp. 579–588, 2018.
[2] J. C. He, P. Y. Chuang, A. Ma’Ayan, and R. Iyengar, “Systems biology of kidney diseases,” Kidney Int., vol. 81, no. 1, pp. 22–39, 2012.
[3] M. H. Tekieh and B. Raahemi, “Importance of Data Mining in Healthcare: A Survey,” in International Conference on Advances in Social Networks Analysis and Mining Importance, 2015, pp. 1057–1062.
[4] V. Kunwar, K. Chandel, A. S. Sabitha, and A. Bansal, “Chronic Kidney Disease Analysis using Data Mining Classification Techniques,” 2016 6th Int. Conf. - Cloud Syst. Big Data Eng., vol. 2, no. 4, pp. 508–511, 2017.
[5] Soundarapandian, L. J. Rubini, and Eswaran, “UCI Machine Learning Repository Chronic Kidney Disease,” 2015.
[6] C.-F. Tsai and F. Y. Chang, “Combining instance selection for better missing value imputation,” Clin. Neuroradiol., vol. 26, no. 4, pp. 481–483, 2016.
[7] R. Deb and A. W. C. Liew, “Missing value imputation for the analysis of incomplete traffic accident data,” Inf. Sci. (Ny.), vol. 339, pp. 274–289, 2016.
[8] F. O. De França, G. P. Coelho, and F. J. Von Zuben, “Predicting missing values with biclustering: A coherence-based approach,” Pattern Recognit., vol. 46, no. 5, pp. 1255–1266, 2013.
[9] M. Bennasar, Y. Hicks, and R. Setchi, “Feature selection using Joint Mutual Information Maximisation,” Expert Syst. Appl., vol. 42, no. 22, pp. 8520–8532, 2015.
[10] W. Qian and W. Shu, “Mutual information criterion for feature selection from incomplete data,” Neurocomputing, vol. 168, pp. 210–220, 2015.
[11] J. A. Sáez, J. Derrac, J. Luengo, and F. Herrera, “Statistical computation of feature weighting schemes through data estimation for nearest neighbor classifiers,” Pattern Recognit., vol. 47, no. 12, pp. 3941–3948, 2014.
[12] Han and Kamber, Data Mining Concepts and Techniques Third Edition, vol. 1. 2012.
[13] C. Jin, S. W. Jin, and L. N. Qin, “Attribute selection method based on a hybrid BPNN and PSO algorithms,” Appl. Soft Comput. J., vol. 12, no. 8, pp. 2147–2155, 2012.
[14] E. M. Mirkes, T. J. Coats, J. Levesley, and A. N. Gorban, “Handling missing data in large healthcare dataset : A case study of unknown trauma outcomes,” Comput. Biol. Med., vol. 75, pp. 203–216, 2016.
[15] R. S. Wahono, N. Suryana, and S. Ahmad, “Metaheuristic Optimization based Feature Selection for Software Defect Prediction,” J. Softw. Eng., vol. 9, no. 5, pp. 1324–1333, 2014.
[16] F. Sebastiani, “Machine Learning in Automated Text Categorization,” ACM Comput. Surv., vol. 34, no. 1, pp. 1–47, 2002.
[17] X. Sun, Y. Liu, M. Xu, H. Chen, J. Han, and K. Wang, “Feature selection using dynamic weights for classification,” Knowledge-Based Syst., vol. 37, pp. 541–549, 2013.
[18] P. Hajek and K. Michalak, “Feature selection in corporate credit rating prediction,” Knowledge-Based Syst., vol. 51, pp. 72–84, 2013.
[19] B. Lantz, Machine Learning with R. Birmingham: Packt Publishing Ltd, 2013.
[20] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*. Elsevier, 2012.
[21] G. E. A. P. A. Batista and M. C. Monard, “An analysis of four missing data treatment methods for supervised learning,” *Appl. Artif. Intell.*, vol. 17, no. 5–6, pp. 519–533, 2003.
[22] D. Jain and V. Singh, “Feature selection and classification systems for chronic disease prediction: A review,” *Egypt. Informatics J.*, vol. 19, no. 3, pp. 179–189, 2018.
[23] X. Wu, *The Top Ten Algorithms in Data Mining*, 1st ed. Francis: CRC Press Taylor & Francis Group, 2009.
[24] J. Suntoro, F. Wahyu, and H. Indriyawati, “Software Defect Prediction Using AWEIG + ADACOST Bayesian Algorithm for Handling High Dimensional Data and Class Imbalanced Problem,” *Int. J. Inf. Technol. Bus.*, vol. 1, no. 1, pp. 36–41, 2018.
[25] A. Ilham, L. Khikmah, A. Qahslim, I. B. A. Indra Iswara, F. E. Laumal, and R. Rahim, “A systematic literature review on attribute independent assumption of Naive Bayes: research trend, datasets, methods and frameworks,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 420, p. 012086, 2018.
[26] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*, Second. 2011.
[27] X. Wu et al., *Top 10 algorithms in data mining*, vol. 14, no. 1. 2008.
[28] R. S. Wahono, N. S. Herman, and S. Ahmad, “A comparison framework of classification models for software defect prediction,” *Adv. Sci. Lett.*, vol. 20, no. 10–12, pp. 1945–1950, 2014.
[29] I. Brown and C. Mues, “An experimental comparison of classification algorithms for imbalanced credit scoring data sets,” *Expert Syst. Appl.*, vol. 39, no. 3, pp. 3446–3453, 2012.
[30] J. Demsar, “Statistical Comparisons of Classifiers over Multiple Data Sets,” *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.
[31] H.-L. Dai, “Class imbalance learning via a fuzzy total margin based support vector machine,” *Appl. Soft Comput.*, vol. 31, pp. 172–184, 2015.
[32] J. Demšar, “Statistical Comparisons of Classifiers over Multiple Data Sets,” *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.