Clustering of Cardiovascular Disease Patients Using Data Mining Techniques with Principal Component Analysis and K-Medoids

Edy Irwansyah1 [1,*], Ebiet Salim Pratama [2] and Margaretha Ohyver [2]

[1] School of Computer Science, Bina Nusantara University; Jakarta 11480, Indonesia
[2] Department of Statistics, School of Computer Science, Bina Nusantara University; Jakarta 11480, Indonesia
* Correspondence: eirwansyah@binus.edu

Abstract: Cardiovascular disease is the number one cause of death in the world and Quoting from WHO, around 31% of deaths in the world are caused by cardiovascular diseases and more than 75% of deaths occur in developing countries. The results of patients with cardiovascular disease produce many medical records that can be used for further patient management. This study aims to develop a method of data mining by grouping patients with cardiovascular disease to determine the level of patient complications in the two clusters. The method applied is principal component analysis (PCA) which aims to reduce the dimensions of the large data available and the techniques of data mining in the form of cluster analysis which implements the K-Medoids algorithm. The results of data reduction with PCA resulted in five new components with a cumulative proportion variance of 0.8311. The five new components are implemented for cluster formation using the K-Medoids algorithm which results in the form of two clusters with a silhouette coefficient of 0.35. Combination of techniques of Data reduction by PCA and the application of the K-Medoids clustering algorithm are new ways for grouping data of patients with cardiovascular disease based on the level of patient complications in each cluster of data generated.

Keywords: data mining; cardiovascular diseases; cluster analysis; principle component analysis

1. Introduction

Cardiovascular disease is the number one cause of death in the world. Cardiovascular disease is a group of diseases caused by impaired heart and blood vessel function. Examples of diseases that are categorized as cardiovascular disease are coronary heart disease, cerebrovascular disease, arterial peripheral disease, rheumatic heart disease, congenital heart disease, and deep vein thrombosis and pulmonary embolism [1]. According to the Ministry of Health of the Republic of Indonesia, quoting from WHO, around 31% of deaths in the world are caused by cardiovascular diseases and more than 75% of deaths occur in developing countries [2].

Cardiovascular disease is a type of non-communicable disease caused by a combination of risk factors that cannot be modified and risk factors that can be modified. Risk factors that cannot be modified are risk factors that cannot be changed such as age, and gender. Modifiable risk factors are factors that can be changed through individual behavior such as smoking, alcohol consumption, poor diet and lack of physical activity. The combination of these two factors causes metabolic disorders such as increased glucose levels and cholesterol in the blood and will increase the risk of cardiovascular disease [3]. In the past 25 years, obesity and diabetes mellitus have overtaken cigarette smoking, dyslipidemia, and hypertension as risk factors for coronary heart disease [4]. Author also conducted an interview with interviewees who work as doctors. The results of the interview concluded that cardiovascular disease is a disease related to the metabolic system of the body that can be complicated by other organs such as the kidneys that function to filter dirty blood and dispose of it through urine.

One problem in the health sector is the large number of documents recorded by medical examinations of patients [5]. Based on the author's interview with the interviewees, examination of patients suffering from cardiovascular disease produces many medical records considering that cardiovascular disease can be caused by the functions of other organs. Through this medical record, several conclusions can be drawn that can be used for medical treatment based on illness, symptoms and treatment aimed at patients [5]. Therefore, it is necessary to develop a system so that the results of medical records, especially in patients with cardiovascular disease can be utilized optimally. The system in question is a computer-based decision making system that was developed using the process of pulling useful information from a collection of data and turning it into a new structure. This system is called data mining. The new structure generated through the data mining system can be used for further analysis [6].

One of the techniques contained in data mining is clustering. Clustering technique can be defined as the process of dividing a collection of data objects (observations) into a new subset so that data objects (observations) in that subset have similarities. One of the basic algorithms used in the clustering process is partitioning method which divides a group of objects into a number of groups that have been determined. One of the clustering algorithms included in the partitioning method is K-Medoids or Partitioning Around Medoids (PAM). The K-Medoids method is the development of K-Means to overcome the presence of outliers [7]. The K-Medoids method uses an object that will be called medoids as the focal point of the cluster.

One of the problems in clustering is high dimensional data or data that has many attributes. As dimensions increase, data becomes more scattered because data points are located in different dimensions [8]. Several methods are used to overcome this problem, one of which is to reduce the dimensions of the data using Principal Component Analysis (PCA) [9]. PCA method is a method is a process in which data that has many dimensions are linearly transformed into a collection of variables that are not related and sorted descending based on variance per component. By using the PCA method, large dimension data can be explained using a number of smaller components that have been formed [10].

Several studies on clustering that combine PCA and K-Medoids methods have been carried out. This combination of methods has been used in the energy field to increase the effectiveness of wind powered energy turbines [11]. PCA is used to reduce variable dimensions that are considered to make wind turbines generating energy ineffective work which is then followed by the selection of the best clustering method. The clustering method used is Fuzzy C-Means, K-Medoids and K-Means which will be evaluated using RMSE. The conclusion of the study stated that the best clustering method was K-Medoids because it produced the smallest RMSE value. In the health sector, a combination of PCA and K-Medoids methods has been used to improve medical diagnoses of patients suffering from cardiovascular disease, Parkinson disease and liver disorders [12]. This study uses PCA as the initial step of the study and then uses K-Medoids to group these patients. The study concluded that the combination of PCA and K-Medoids could help to improve medical diagnoses of patients suffering from cardiovascular disease, Parkinson disease and liver disorders. Through this explanation, this study combines the PCA method and the K-Medoids algorithm to cluster patients with cardiovascular disease. The purpose of using both methods is to determine the characteristics of patients with cardiovascular disease found in each cluster based on the level of complications. This is expected to improve the quality of treatment of patients with cardiovascular disease.

2. Materials and Methods

2.1. Data Sources and Research Variables

This study uses secondary data obtained from a private hospital in Jakarta. The data obtained are 644 observations. This study uses age variables as well as 8 variables of the results of blood tests of patients with cardiovascular disease presented in Table 1.

Table 1 Research Variables

Variable	Notation	Explanation of Variables	Scale of Measurement
Age	AGE	Age of the patient in units of year.	Ratio
Urea levels	UREA	Urea levels in the patient's blood. The unit of measurement of urea levels in this study is mg / dL.	Ratio
Keratin levels	CREA	Creatinine levels in the patient's blood. The unit of measurement of creatinine levels in this study is mg / dL.	Ratio
Uric Acid Levels	UA	Uric acid levels in the patient's blood. The unit of measurement for uric acid levels in this study is mg / dL.	Ratio
Cholesterol Levels	CHOL	Cholesterol levels in the patient's blood. The unit of measurement of cholesterol levels in this study is mg / dL.	Ratio
Triglyceride Levels	TRIG	Triglyceride levels in the patient's blood. The unit of measurement of glucose levels in this study is mg / dL.	Ratio
HDL levels	HDL	HDL levels in the patient's blood. The unit of measurement for HDL levels in this study is mg / dL.	Ratio
Glucose Levels	GLU	Glucose level in the patient's blood in the morning. The unit of measurement of glucose in this study is mg / dL.	Ratio
Glucose Levels Two Hours after Eating	GLU2J	Glucose levels in the patient's blood are measured two hours	Ratio

	after the patient consumes food. The unit of measurement used is mg / dL.

2.2. Normalization of Z-Score

Normalization is the process of transforming data to equate a range of values with a certain scale. Normalization of z-score is a normalization method based on the average value and standard deviation of the data [13]. The formula of z-score can be written as follows:

$$z = \frac{x_i - \bar{x}}{s} \qquad (1)$$

where:

z	=	Value of *z-score*
x_i	=	observation to-i
$\bar{x}$	=	average observation
s	=	deviation standard of data

2.3. Principal Component Analysis

Principal Component Analysis (PCA) is a multivariate analysis technique introduced first time in 1901 and developed by [14]. The main idea of making PCA techniques is to reduce the dimensions of data sets that have many interconnected variables, but still maintain as many variations as they appear in the data set. This dimension reduction is obtained through transformation to a new set of variables called principal components, which are uncorrelated and sequential so that some initial components retain variations that appear within the original variable [15].

The process of forming major components in PCA uses values of *eigen* and vector eigen. In determining the value of eigen of a matrix, such as the X matrix, the equation used is the following equation [16]:

$$\det(X - \lambda I) = 0 \qquad (2)$$

where:

X	=	matrix of size $n \times n$
λ	=	value of *eigen*
I	=	identity matrix

Calculation of vector eigen on matrix X can be used as follows:

$$(X - \lambda I)x = 0 \qquad (3)$$

where:

X = matrix of sized$n \times n$

λ = value of *eigen*

I = identity matrix

x = vector *eigen*

The value of eigen formed based on equation (2) is used to determine the proportional variance of each component formed. The proportion of variance is used to find out how big a component is to explain the diversity of data [17]. Calculation of the proportion of variance can use the following formula:

$$p_i = \frac{\lambda_i}{\sum_{j=1}^{D} \lambda_j} \tag{4}$$

where:

p_i = variance of proportions of components to -*i*

λ_i = Value of *eigen on* component to -*i*

There are several criteria that are used as a reference to determine the number of main components taken. One of these criteria is to take components that have cumulative proportional variance values between 80% to 90% [17]. Calculation of the value of the cumulative proportion can be obtained using the following formula:

$$pk_r = \frac{\sum_{i=1}^{r} \lambda_i}{\sum_{j=1}^{D} \lambda_j} \tag{5}$$

dengan $\lambda_1 > \lambda_2 > \cdots > \lambda_D$

where:

pk_r = the value of the variance of cumulative proportions from the first component to the component to -*r*

λ_i = value of *eigen on* component to-*i*

$\lambda_1, \lambda_2, \ldots, \lambda_D$ = Value of *eigen* from the first component to the component to -*D*

2.4. K-Medoids

K-Medoids is one of the basic algorithms used in clustering partitioning methods. Basic K-Medoid is an algorithm called Partitioning Around Medoids (PAM). In the PAM algorithm, object representations in clusters are called medoids [18]. K-Medoids is one algorithm that is better than K-Means if there are outliers in a data set. This is because the K-Means center is formed using an average of observations which if there is an outlier value will greatly affect the value of the center point. Whereas K-Medoids uses objects that are used as center points (medoids) so that they do not have much effect if there are outliers in the data set used [7] The stages of the K-Medoids algorithm are as follows [7]:

1. Determine the amount of *k,* which will be the number of clusters. In determining the amount *k,* can use silhouette coefficient calculations as in the following formula:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{6}$$

where:

$s(i)$ = *silhouette coefficient*

$a(i)$ = average distance of the object to -*i* with all objects that are in one cluster

$b(i)$ = minimum value of the average distance of the object to -*i* with other objects that are in another cluster

2. Randomly select number of *k* observation in *D* as a representation (*medoids*) from the cluster (o_j).
3. Calculate the distance of the observation to the chosen medoids and place the observations into the cluster closest to the medoids.
4. Randomly select the non-medoids observation from the data set D (o_{random}).
5. Calculate the total cost (*S*) from exchange of *medoids* o_j with o_{random}.
6. If S < 0, exchange o_j with o_{random} to form a group *k* new observations as *medoids.*
7. Repeat the second until the fifth step until there is no exchange.

2.5. Research Stages

The research process begins with cleaning out incomplete medical record data. The process is continued by normalizing the data using values of *z-score*. The next step is to reduce the dimensions of the data using Principal Component Analysis to produce new components that will be used for the clustering process using K-Medoids. The research process ends with an evaluation of the results of the clustering with interviewees to determine the level of complications of patients suffering from cardiovascular disease. The research process is illustrated in Figure 1.

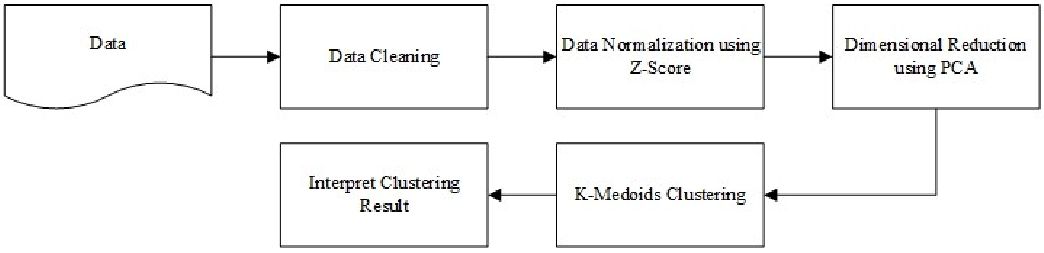
Data
Data Cleaning
Data Normalization using Z-Score
Dimensional Reduction using PCA
K-Medoids Clustering
Interpret Clustering Result

Figure 1 Research Process

3. Results

The entire calculation and analysis process in this study was carried out using R software. Descriptive statistical values (minimum value, quartile 1, median, mean, quartile 3, maximum value and standard deviation) of each variable are presented in the following Table 2:

Table 2 Descriptive Statistics of Research Data

Variable	Minimum	Q_1	Median	Average	Q_3	Maxi-mum	Standard Deviation
AGE	35	45.75	51	53.4	60.25	93	11.097
UREA	9	21	25	26.19	29	84	9.029
CREA	0.6	1	1,1	1.154	1,2	3.7	0.287
UA	2.4	5.2	6.1	6.204	7.1	11.9	1.425
CHOL	91	168	192.5	199.3	226	366	43.127
TRIG	35	87	133	150.5	178	789	93.67
HDL	22	36	43	43.62	48	88	10.138
GLU	63	86	93	103.7	99	532	38.789
GLU2J	67	112	124	138	144.2	599	58.024

Q1 : Value in 1st Quartile, Q3 : Value in 3th Quartile

Based on Table 2, all patients with cardiovascular disease are adults aged between 35 and 93 years. Through the value of quartile 1, 25% of patients with cardiovascular disease are under 45.75 years old and quartile 3 shows that 25% of patients with cardiovascular disease are aged over 60.25 years. Based on the average value (53.4) is greater than the median value (51) shows that the distribution of AGE data values is above the median value. The standard deviation value (11,097) indicates that the data value is quite varied. Explanation of descriptive statistics from other variables can follow the previous description.

In this study, the boxplot graph is used to detect the presence or absence of outlier values in each variable [19]. Boxplot of each variable is presented in Figure 2 below:

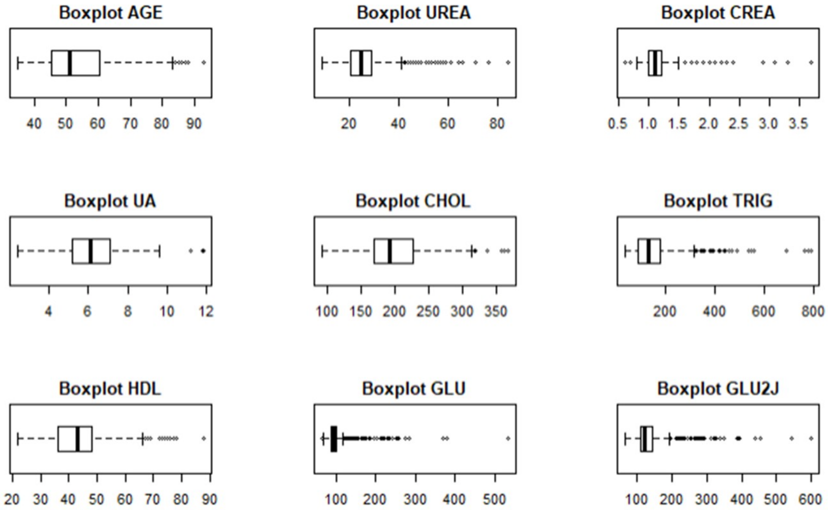

Figure 2 Boxplot of Research Variables

Through **Figure 2** all variables have outlier data which is indicated by points that are outside of the boxplot section.

The next step is to normalize the data using *z-score* (1). The initial stage of calculating the z-score is to determine the average value and standard deviation of the data with the example shown in **Table 3** below:

Table 3 Calculation of Z-Score

AGE Range	Avearage	Standard Deviation	*Z-Score*
35 - 93	53,4	11,1	-1,57
			-1,57
			-1,57
			-1,57
			-1,49

Calculation process of *z-score* is done to all variables with the same process as in **Table 3**. The five initial data normalized results are presented in **Table 4** below:

Table 4 Five Initial Data of Normalization Results

AGE	UREA	CREA	UA	CHOL	TRIG	HDL	GLU	GLU2J
-1,57	-0,02	-0,54	-0,00	0,43	-0,37	1,22	-0,4	-0,7
-1,57	-0,8	-0,54	-0,28	-0,96	-0,16	-1,44	-0,22	-0,01
-1,57	-0,58	-0,54	0,28	-0,42	4,22	-2,03	-0,3	-0,2
-1,57	-0,24	-0,19	-0,85	-0,75	-0,89	0,04	-0,3	-0,74
-1,48	-0,13	-0.18	-1,41	-0.7	-0,95	0,04	-0,28	-0,41

The next process is to reduce the dimensions of the normalized data using PCA. The PCA process will produce an *eigen* value or lambda (λ) with the proportion variance as well as the cumulative proportion variance presented in **Table 5** below:

Table 5. Values λ, Proportion Variance and Cumulative Proportion Variance

Component	λ	Varians Proportion	Cumulative Varians Proportion
PC1	2.69	0.2994	0.2994
PC2	1.66	0.1846	0.4839
PC3	1.33	0.1480	0.6319
PC4	1.08	0.1205	0.7524
PC5	0.71	0.079	0.8311
PC6	0.62	0.069	0.9
PC7	0.43	0.047	0.9477

PC8	0.29	0.032	0.9795
PC9	0.18	0.02	1

PC: Principal Component

The proportion variance for PC1 in Table 2 shows that PC1 can explain the diversity of data by 0.2994. PC2 can explain the diversity of data by 0.1846 and so on until PC9. The cumulative proportion variance value on PC1 includes diversity of 0.2994 while the cumulative proportion variance value will increase to 0.4839 if PC1 and PC2 are taken. The cumulative proportion value will be 1 if PC1 to PC9 is taken. In Table 2 it can be seen that PC1 through PC5 has illustrated the diversity of data of 0.8311 so that the components of PC1 through PC5 that are formed will be used for the clustering process using K-Medoids.

Determination of the best number of clusters uses the calculation of the value of the silhouette coefficient (6) and the value of the silhouette coefficient is obtained as shown in Figure 3 below:

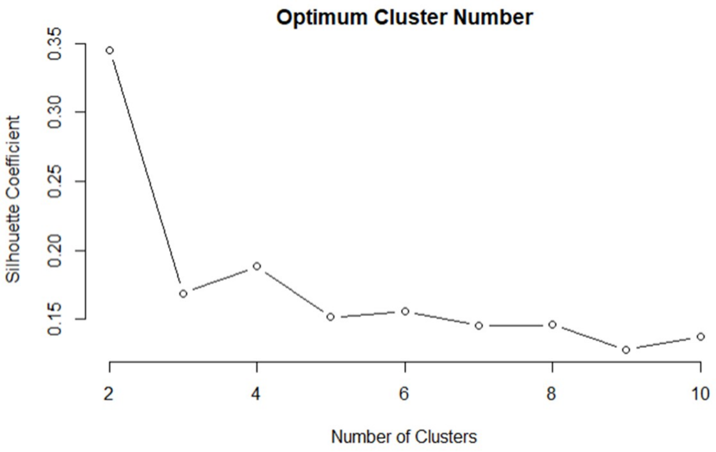

Figure 3 Value of Silhouette Coefficient

Through Figure 3, the number of the best clusters formed is two clusters because it has a value of silhouette coefficient of 0.35 and the highest among the number of other clusters. The next step is to carry out the clustering process using K-Medoids and this process classifies 503 patients into cluster 1 and 141 patients in the cluster 2. The distribution of variable data in each cluster is presented in Figure 4 below:

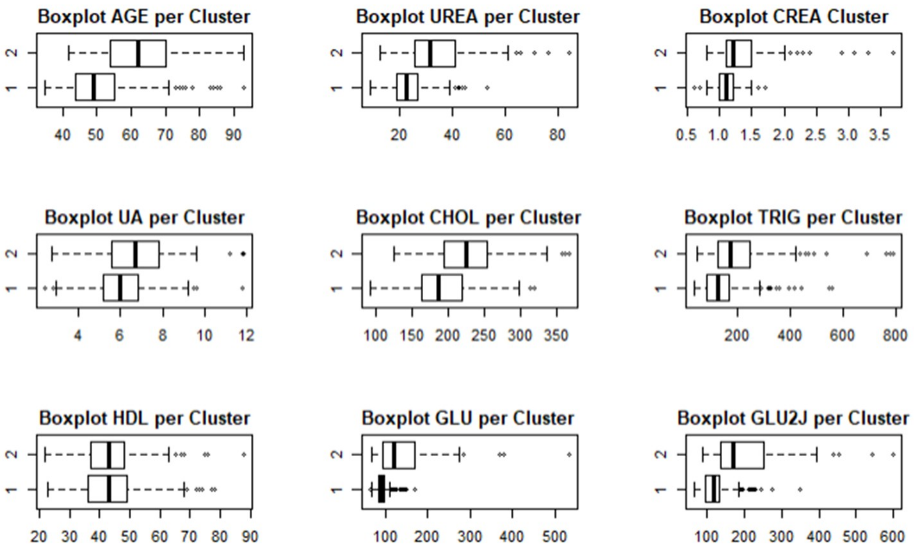

Figure 4. Distribution of Variable Data in Each Cluster

Through Figure 4 patients with cardiovascular disease who are categorized as cluster 1 have lower rates of cardiovascular disease complications. This is because the distribution of variable data of AGE, UREA, CREA, UA, CHOL, TRIG, GLU and GLU2J from patients in cluster 1 is lower than that of patients grouped into cluster 2. Through Figure 4, it can be seen that the HDL variables in cluster 1 and cluster 2 do not show significant differences. The results of the clustering have been validated by the knowledge of users engaged in the medical field. The mean and median values of each cluster are more clearly presented in the following Table 6:

Table 6 Mean Values and Median Variables in Each Cluster

Variable	Averages		Median	
	1st Cluster	2nd Cluster	1st Cluster	2nd Cluster
AGE	50.69	63.06	49	62
UREA	23.7	35.09	23	32
CREA	1.087	1.395	1.1	1.2
UA	6.032	6.817	6	6.7
CHOL	191.9	225.8	185	224
TRIG	134.1	209.2	123	174
HDL	43.63	43.61	43.63	43
GLU	93.36	140.5	91	119
GLU2J	121.9	199.1	119	171

Through Table 6 the mean and median values of the AGE variable in cluster 2 are higher than in cluster 1. This can be a concern because the higher age can cause plaque to stick to the walls of the heart and cause disruption of bloodstream through it [20]. The mean and median values of the UREA variable in cluster 2 are higher than cluster 1. However, both the mean and median values of the UREA variable in each cluster are still at normal levels or below 40 mg / dL [21]. The mean and median values of the CREA variable in cluster 2 are higher than cluster 1. However, both the mean and median values of the CREA variable in each cluster are still in normal numbers or below the level of 1.4 mg / dL such as previous research conducted by [21]. The mean and median values of the UA variable in cluster 2 are higher than cluster 1. In cluster 1 the mean and median values of the UA variable are still at normal numbers or below the level of 6.3 mg / dL but the mean and median values of the UA variable in cluster 2 have passed the normal limit. The mean and median values of the CHOL variable in cluster 2 are higher than cluster 1. In cluster 1 the mean and median values of the CHOL variable are at normal numbers or below 200 mg / dL but the mean and median values of the CHOL variables in cluster 2 have passed the normal limit. The mean and median values of the TRIG variable in cluster 2 are higher than cluster 1. In cluster 1 the mean and median values of the TRIG variable are normal or below 200 mg / dL but in cluster 2 the mean values are above the normal limit and the median does not cross the normal limit. The mean and median values of the HDL variable in cluster 1 are higher than cluster 2. But both the mean and median values of the CREA variable in each cluster are still at low levels or below 65 mg / dL. The mean and median values of the GLU variable in cluster 2 are higher than cluster 1. In cluster 1, the mean and median values of GLU variables are in normal number or below 110 mg / dL, but the mean and median values of GLU variables in cluster 2 have passed the normal limits. In the GLU2J variable, the mean and median values in cluster 2 are higher than cluster 1. In cluster 1, the mean and median values of the GLU2J variable are normal or below 140 mg / dL, but the mean and median values of the GLU2J variable in cluster 2 have passed the normal limit.

4. Conclusion

Data reduction technique with PCA from eight variable data of blood test of patients with cardiovascular disease and age variables obtained from 644 observations, can produce new five components with adequate data diversity

Through five new components resulting from data reduction, and the implementation of the K-Medoids algorithm, two data clusters can be produced for patients with cardiovascular disease with silhouette coefficient values which indicate low data density levels. Patients in the first cluster has lower rates of cardiovascular complications compared to patients in the second cluster, which is due to the lower distribution of data values for each variable AGE, UREA, CREA, UA, CHOL, TRIG, GLU and GLU2J.

The combination of data reduction techniques with PCA and the application of the K-Medoids clustering algorithm is a new way of data mining to group data of patients with cardiovascular disease to see the level of patient complications in each different data cluster.

There is still a lack of cluster evaluation results shown by the value of the Silhouette coefficient (SC) which has a weak structure with a value of 0.35 therefore it is necessary to develop further research methods to produce clusters with stronger structures.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: title, Table S1: title, Video S1: title.

Author Contributions: Conceptualization, E.I. and M.O.; methodology, I.E and M.O.; software, E.I and E.P.; validation, E..I, E.P. and M.O..; formal analysis, E.P.; investigation, E.P.; resources, E.I.; data curation, E.P.; writing—original draft preparation, E.P.; writing—review and editing, E.I.; visualization, E.P.; supervision, E.I.; project administration, M.O.; funding acquisition, M.O. All authors have read and agreed to the published version of the manuscript

Funding: This research received no external funding.

Acknowledgments: We gratefully acknowledge the support of institutions in the implementation of this research. Many thanks to the Mayapada Hospital, Jakarta, Indonesia, which supported the availability of data. We also thank the Research and Technology Transfer Office of Bina Nusantara University, which provided internal funding for the research until the publication of this article

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Health Organization. Cardiovascular Diseases (CVDs) [Internet]. Geneva, Switzerland: WHO; c2017 [cited at 2019 Apr 24]. Available from: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).
2. Ministry of Health Republic of Indonesia. Cardiovascular Hospital "Harapan Kita" a Cardivascular referrals [Internet]. Jakarta, Indonesia: Kementrian Kesehatan Republik Indonesia; c2018 [cited at 2019 Jul 25]. Available from http://www.depkes.go.id/article/print/18111200002/rs-jantung-harapan-kita-pengampu-rujukan-kardiovaskular.html.
3. Grundy, S. M. Metabolic syndrome: a multiplex cardiovascular risk factor. The J of Clinical Endocrinology and Metabolism 2007; 92(2): 399-404
4. Smith Jr, S. C. Multiple risk factors for cardiovascular disease and diabetes mellitus. The American J of med 2007;120(3): S3-S11.
5. Naaz E, Sharma D, Sirisha D, M Venkatesan. Enhanced K-Means Clustering Approach for Health Care Analysis using Clinical Documents. Int J of Pharm and Clinical Res 2016; 8(1):60-64.
6. Vijayashree J, Iyengar N Ch S N. Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques: A Review. Int J of Bio-Sci and Bio-Tech 2016; 8(4):139-148.
7. Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. 3rd ed. Amsterdam: Morgan Kaufmann;2012.
8. Singh N, Garg N, Pant J. A Comprehensive Study of Challenges and Approaches for Clustering High Dimensional Data. Int J of Comp Appl 2014; 92(4):7-10.
9. Pavithra M, Parvathi Dr R M S. A Survey on Clustering High Dimensional Data Techniques. Int J of Appl Eng Res 2017; 12(11):2893-2899.
10. Metsalu T, Vilo J. ClustVis: A Web Tool for Visualizing Clustering of Multivariate Data Using Principal Component Analysis and Heatmap. Nucleic Acids Res 2015; 43: 566-570.
11. Al-Shammari E T, Shamshirband S, Petkovic D, Zalnezhad E, Yee P L, Taher R S, Cojbasic Z. Comparative Study of Clusterig Methods for Wake Effect Analysis in Wind Farm. J Energy 2016; 95:573-579.
12. Peker M. A Decision Support System to Improve Medical Diagnosis using a Combination of K-Medoids Clustering based Attribute Weighting and SVM. J Med Sys 2016; 40(116): 1-16.
13. Nasution D A, Khotimah H H, Chamidah N. Perbandingan Normalisasi Data untuk Klasifikasi Wine menggunakan Algoritma K-NN. J of Comp Eng Sys and Sci 2019; 4(1):78-82.
14. Hotelling H. Analysis of a Complex of Statistical Variables Into Principal Components. J of Edu Psy 1933; 24(6):417-441.
15. Jolliffe IT. Principal Component Analysis, Second Edition. New York (NY): Springer; 2002.
16. Aryani F, Maisyitah R A D. Nilai Eigen dan Vektor Eigen dari Matriks Kompleks Bujursangkar Ajaib. J Sains Mat dan Stat 2015; 1(2):10-16.
17. Johnson Richard A, Wichern Dean W. Applied Multivariate Statistical Analysis Sixth Edition. Upper Saddle River(NJ): Pearson; 2007.
18. Kaufman L, Rousseuw P J. Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken (NJ): John Wiley & Sons; 2005.
19. Darsyah M Y. Penggunaan Stem and Leaf dan Boxplot untuk Analisis Data. J Karya Pend Mat 2014; 1(1): 55-67.
20. Ghani L, Susilawati M D, Novriani H. Faktor Risiko Dominan Penyakit Jantung Koroner di Indonesia. Ind Bull of Health Res 2016; 44(3): 153-164.
21. Miranda E, Irwansyah E, Amelga A Y, Maribondang M M, Salim M. Detection of Cardiovascular Disease Risk's Level for Adults Using Naive Bayes Classifier. Healthc Inform Res 2016; 22(3): 196-205.