Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 6, No. 2, May 2021, Pp. 109-116

ClusterMix K-prototypes algorithm to capture variable characteristics of patient mortality with heart failure

Raditya Novidianto[1], Hardianto Wibowo[*2], Didih Rizki Chandranegara[3]
Institut Teknologi Sepuluh Nopember, Indonesia[1]
Universitas Muhamadiyah Malang, Indonesia[2,3]

Article Info
Keywords:
Internet of Things Platform, Internet of Things, Message Queuing Telemetry Transport, MQTT Broker Server

Article history:
Received: January 27, 2021
Accepted: April 27, 2021
Published: May 31, 2021

Cite:
Novidianto, R. ., Wibowo, H., & Chandranegara, D. R. (2021). ClusterMix K-Prototypes Algorithm to Capture Variable Characteristics of Patient Mortality With Heart Failure. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 6(2). https://doi.org/10.22219/kinetik.v6i2.1209

*Corresponding author.
Hardianto Wibowo
E-mail address:
ardi@umm.ac.id

Abstract
Cardiovascular Disease (CVD) is one of the leading causes of many death worldwide, leading to heart failure incidence. The World Health Organization (WHO) says the number of people dying from cardiovascular disease from heart failure each year has an average of 17,9 million deaths each year, about 31 percent of the total deaths globally. Identify the mortality factors of heart failure patients that need to be formed, which reduces death due to heart failure. One of them is by using variable mortality due to heart failure by applying the k-prototypes algorithm. The clustering result is formed 2 clusters that are considered optimal based on the highest silhouette coefficient value of 0,5777. The results of the study were carried out as segmentation of patients with variable mortality of heart failure patients, which showed that cluster 1 is a cluster of patients who have a low risk of the chance of mortality due to heart failure and cluster 2 is a cluster of patients with a high risk of mortality due to heart failure. The segmentation is based on the average value of each variable of heart failure mortality factor in each cluster compared to normal conditions in serum creatine variables, ejection fraction, age, serum sodium, blood pressure, anemia, creatinine phosphokinase, platelets, smoking, gender, and diabetes.

1. Introduction

Cardiovascular Disease (CVD) is one of the leading causes of death and disability worldwide due to disorders of the heart and blood vessels, including coronary heart disease, stroke, heart failure, and other pathology types. The perception between doctors and patients is needed to do more attention so that the cure rate of this disease can be appropriately handled [1]. Cardiovascular disease is a deadly disease because data from the World Health Organization (WHO) states that people die of cardiovascular disease each year, having an average of 17,9 million per year, about 31 percent of the total deaths globally [2]. The cardiovascular disease makes the heart muscle work faster, causing heart failure. The state of the heart organs in humans will slowly weaken, and the longer it will be harder to pump blood properly when the heart condition weakens, certain substances will be released in the blood. The patient has a history of congenital diseases such as anemia, diabetes, blood pressure, other diseases, and other factors. Certain substances have a toxic effect on the blood to cause heart failure conditions [3].

Heart failure is a condition in which the heart wall muscles begin to loosen, enlarge, and restrict blood pumping to the heart [4]. The cause of heart failure can be due to ejection fraction, which is the proportion of blood pumped out of the heart during one contraction with percentage values ranging between 50 percent and 75 percent. Some of the causes of heart failure can be reduced ejection fraction (HFrEF), commonly known as heart failure due to left ventricular systolic dysfunction or systolic heart failure, characterized by ejection fractions smaller than 40 percent. Furthermore, heart failure with a stable ejection fraction (HFpEF), commonly referred to as diastolic heart failure or heart failure with an average ejection fraction. In this case, the left ventricle contracts normally during the systole, but the ventricle stiffens and fails to relax normally during the diastole, thereby interfering with the filling [5].

The heart is the most important vital organ because its function is related to a person's survival chances. Analyzing the survival of heart failure patients is a priority for doctors who aim to improve their health condition. However, until now, heart failure patients' clinical healing actions tend to remain relatively minimal because the characteristics of heart failure patients are very difficult to detect [6].

Patient health records or commonly called Electronic Health Records (EHR) is a recorded record used as a source of information about the characteristics of heart failure patients so that it can be known or contained in the role of demographic characteristics and other variables both directly and indirectly in clinical practice of healing heart failure patients [7]. A study studied a common survival pattern that showed a high mortality intensity of heart failure patients in the early days and then increased gradually until the end of the study [8].

Heart failure patients' mortality factors can be modeled, taking into account age, ejection fraction, serum creatine, serum sodium, anemia, platelets, creatinine phosphokinase, blood pressure, gender, diabetes, and smoking status potentially contribute to death [4]. The large role of variable mortality factors of heart failure patients is illustrated through an algorithm in machine learning so that it obtained variables importance determining the incidence of sequenced heart failure mortality, namely serum creatine, ejection fraction, age, serum sodium, blood pressure, anemia, creatinine phosphokinase, platelets, smoking, gender and diabetes [9]. Further research is to determine a segment of heart failure patients' mortality factors to see the character of heart failure patients based on similarity levels or similarities with assumptions that appear in a cluster are patients with characteristics that tend to be homogeneous in clusters heterogenic between clusters [10].

When grouping datasets with large observation units, the method often used is the k-means or k-modes method to form a segmentation of each cluster's characteristics [11]. The inaction in using the methodology is the number of clusters that need to be determined before the algorithm is applied, and the k-means method can only be used in continuous data, and k-modes can only be used on categorical data [12]. Data types are very broad in the real world and even tend to be mixed data types, so there are k-means and k-modes modification algorithms to integrate the algorithm and then build the clusterMix K-Prototypes algorithm [13].

Based on the above background, the researcher is interested in discussing the characteristics of the mortality factors for cardiovascular disease patients who have heart failure by having mixed data information for analysis using the ClusterMix k-prototypes algorithm. The results of the analysis are very useful to provide information on the description of cardiovascular disease patients through the proximity of the patient's characteristics, which are divided into k groups that must be treated differently. The validity of the similarity measurement in this study is based on the silhouette coefficient to obtain the optimum k or number of groups. The purpose of this study was to obtain optimal grouping results in the process of grouping heart failure patients and forming patient segments based on similar variables for the benefit of further management of heart failure patients. The importance of information about the results of this cluster can help medical personnel in taking action based on the segmentation formed in heart failure patients so that the incidence of heart failure can be minimized. For the world of machine learning, this is a learning medium with a new method, namely ClusterMix, in forming segments with mixed data.

2. Research Method

Descriptive statistics is one of the basic methods used to describe a particular situation by collecting, processing, and disseminating data collection [14]. In this study, descriptive methods such as average, standard deviation, median, and mode describe the variables of patients who have heart failure. Descriptive statistics will be presented in tables, images, heatmaps, charts, and boxplots.

2.1 Dataset Collection

The data used in this study is Electronic Health Records (Medical Records) of 299 heart failure patients collected by Faisalabad Institute of Cardiology and Allied Hospital in Faisalabad (Punjab, Pakistan) from April until December 2015 [4][15]. Patients consisted of 105 women and 194 men, and ages ranged between 40 and 95 years. Following Table 1, this is a table listing the variables used in this study.

Table 1. Research Variables

Code	Variable	Unit	Limit	Data Type
X1	Age	Years	[40,..., 95]	Numerical
X2	Anemia	Boolean	0, 1	Categorical
X3	Blood Pressure	Boolean	0, 1	Categorical
X4	Creatinine phosphokinase (CPK)	mcg/L	[23,..., 7861]	Numerical
X5	Diabetes	Boolean	0, 1	Categorical
X6	Ejection fraction	Percentage	[14,..., 80]	Numerical
X7	Gender	Binary	0, 1	Categorical
X8	Platelets	kiloplatelets/mL	[25.01,..., 850.00]	Numerical
X9	Serum creatinine	mg/dL	[0.50,..., 9.40]	Numerical
X10	Serum sodium	mEq/L	[114,..., 148]	Numerical
X11	Smoking	Boolean	0, 1	Categorical
X12	Time	Days	[4,...,285]	Numerical

2.2 Preprocessing Data

Before conducting cluster analysis that aims to group a number of objects based on the similar characteristics

they have, it determines the number of groups. Objects clustered in one cluster have a high degree of similarity, and objects between clusters have a low degree of similarity [16]. The general steps are carried out in cluster analysis, namely determining the Size of the similarity, the method of clustering, doing the symbolization, and the last is the interpretation of the clustering results [17]. Determination of the number of groups is carried out using the silhouette coefficient.

The silhouette coefficient is a method often used in cluster analysis to determine the exact number of *k* (cluster count) in the clustering process [18]. The silhouette coefficient can also measure the quality of clusters that have been formed [19]. Measurement of silhouette coefficient is formulated as follows Equation 1 [20].

$$S_i = \frac{b_i - a_i}{max(a_i b_i)} \quad (1)$$

Where,

a_i : Average distance between i object and an entire object in the same cluster

b_i : Average distance between i object and all objects in the nearest cluster

The similarity measure is then used in cluster analysis using the distance between objects and the distance between clusters because this research uses mixed data, namely Euclidean distance and distance categorical data types. Euclidean distance is used to measure the distance between objects with numerical data. One use is on the k-means algorithm. The Euclidean distance between the i to i object and the j object with the p variable is as follows [21]. Following is the Euclidean distance Equation 2.

$$d_{ij} = \left[\sum_{k=1}^{p}(x_{ik} - x_{jk})^2\right]^{1/2} \quad (2)$$

With :

d_{ij} : Euclidean distance between i object and j object

x_{ik} : the value of the object to i in the variable to k

x_{jk} : the value of the object to j in the variable to k

p : the number of variables observed

Then calculated the k-mode algorithm to group all categorical data. The distance measure used by the k-modes algorithm follows Equation 3 [22].

$$d_2^{'}(X,Y) = \sum_{j=1}^{p}(x_j - y_j)^2 + \gamma \sum_{j=p+1}^{m}\delta(x_j, y_j) \quad (3)$$

Where,

$d_2(X, Y)$ = Size of the distance between objects X *and* Y (mixed data)

$\sum_{j=1}^{p}(x_j - y_j)^2$ = Distance size for numerically typed data

$\sum_{j=p+1}^{m}\delta(x_j, y_j$ = Distance size for categorized data

γ = Weighing parameters

2.3 K-Prototypes Cluster Mix Algorithm

The ClusterMix k-prototypes algorithm is built by combining the k-means and k-modes algorithms. The following is the calculation of the algorithm for each algorithm to form k-prototypes

2.3.1 K-Means Algorithm

The k-means algorithm is a non-hierarchical clustering method that determines each object's grouping based on the closest average value. The k-means algorithm steps are dividing objects into initial clusters, grouping objects into clusters that have the closest average value, recalculating the average value for clusters that receive new objects or lose objects, and repetition until there are no more objects movements [23].

2.3.2 K-Modes Algorithm

The k-modes algorithm uses a distance measure for categorical data. The stages of the k-modes algorithm are determining the initial model for each cluster, allocating objects to the cluster based on the closest mode, retesting the

object's distance to the last mode, reallocating if there are objects close to other clusters, and repeating until no objects change clusters [22].

2.3.3 K-Prototypes Algorithm

The k-prototypes algorithm uses a mixed distance measure characterized by γ. The variable γ is a weighting parameter used to balance the two distance functions' proportions for numeric and categorical data. The k-prototypes algorithm can be done in the following stages [24]:

1. Determine the number of clusters (k) to be formed. The minimum limit of Size k is $\sqrt{n}$ clusters, while the maximum limit of is or $n/2$ where n is the number of observations.
2. Determine k initials of the prototypes, namely $Z_1, Z_2, ..., Z_k$ as the cluster center in each cluster.
3. Calculate the distance of all observations in the dataset against the initial cluster initials. The distance measure used is mixed.
4. Allocating all observations into clusters that have the closest prototype distance to the object being measured.
5. Performs the calculation of the new cluster center point after all objects have been allocated.
6. Reallocate all observational data in the dataset to the new prototype.

If the cluster center point does not change or has converged, the algorithm process stops. But if the center is still changing significantly, the process returns to stages 2 to 5 until the maximum iteration is reached or there is no more object displacement**.**

3. Results and Discussion

The following is an analysis of the discussion regarding the characteristic variables of heart failure patients in a hospital in Pakistan that can be illustrated through the distribution of data based on descriptive statistics and in-depth analysis using the k-prototypes algorithm.

3.1 Result

Exploration results show that there is no missing value in the categorical data. Of the 299 objects studied, a visual-based analysis of the pattern of relationships between objects will be carried out using a heatmap so that the relationship between observations can be found based on the strength of each patient's characteristic variables so that researchers can see the cluster pattern earlier before entering the k-prototypes algorithm. Visualized data is data that has been standardized on numerical data, which aims to eliminate unit differences in numerical variables so that they do not look dominant. The grouping process is detecting observations through similarity using the distance matrix in Figure 1.

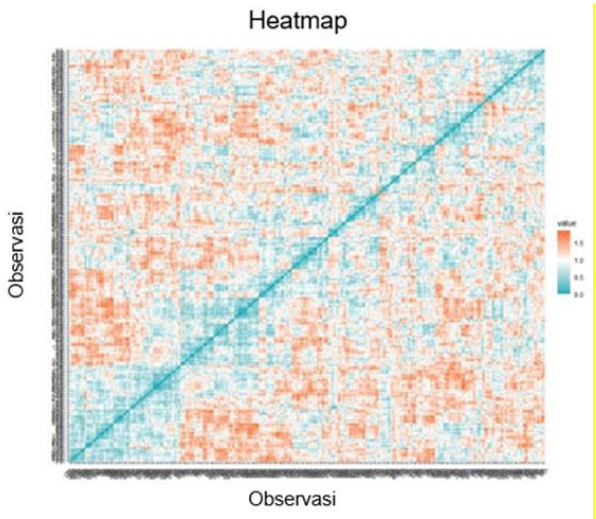

Figure 1. Heatmap Matrix Distance Between Observations

The process of grouping the data $n \times n$ with the number n is 299. It can be seen from the level of similarity by using a distance matrix using distance calculations, namely the Euclidean distance. Then the distance matrix is depicted visually using a heatmap as shown in Figure 1, showing the clustering process using a heatmap with graded colors. In the heatmap, it can be seen which objects have strong similarities between one observation and another. After identifying, it is continued using a grouping heatmap based on the similarity of the variables shown in Figure 2.

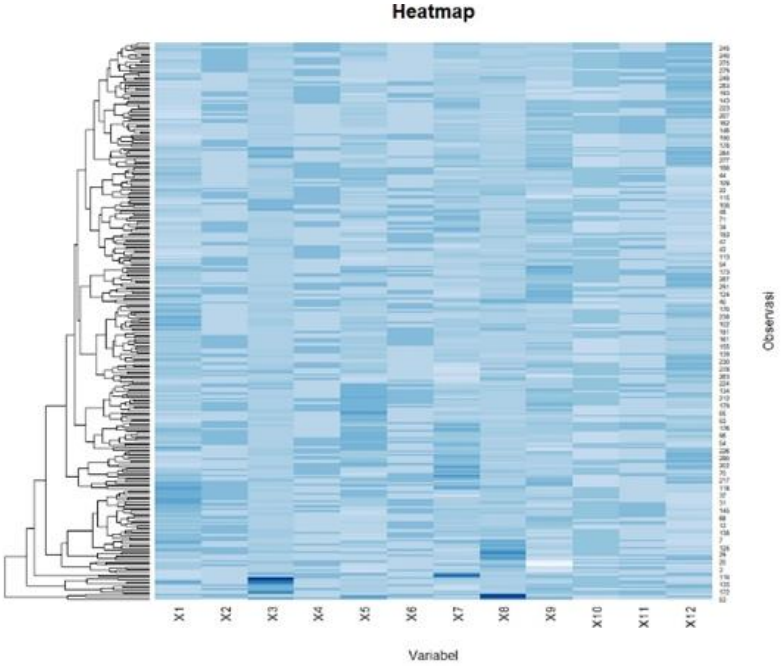

Figure 2. Heatmap of the Hierarchical Grouping of Observations

The determination of the number of clusters is obtained from the way of evaluating the calculation of silhouette coefficients in each cluster formed so that homogeneous clusters can be obtained in one cluster and heterogeneous between clusters. The greater the silhouette coefficient value, the optimal number of clusters assuming the more homogeneous the cluster is formed, the higher the level of correlation of objects inside so that the value of the silhouette coefficient will also be higher. The following is the calculation of the silhouette coefficient with the minimum calculation limit that the clusters formed are two and a maximum of 20 clusters formed in Figure 3.

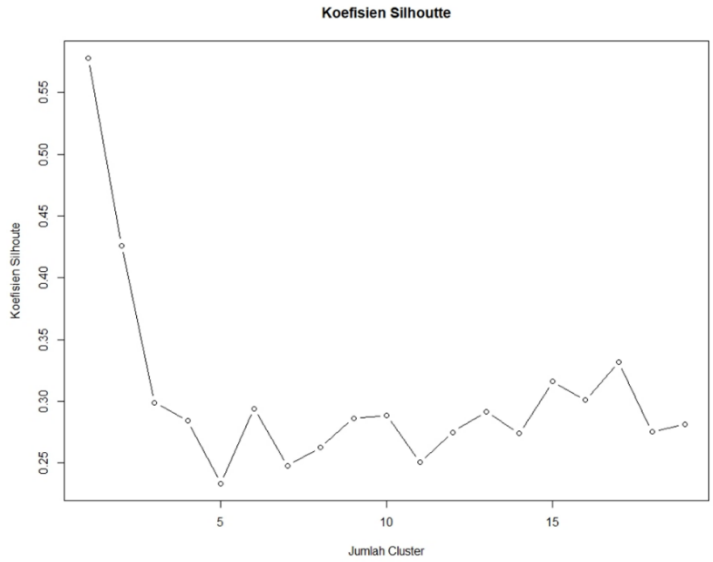

Figure 3. Silhoutte Coefficient

The selection of the number of clusters is carried out in stages starting from the number of clusters of 2 to 20 clusters. There was a fluctuation in the silhouette coefficient at each stage of the number of clusters or the k value. Figure 3 shows a decrease in the silhouette coefficient from k = 2 with a coefficient value of 0,5777 to k = 5 with a coefficient value of 0,23361, then an increase in k = 6 to k = 7, then tends to decrease from k = 8 to k = 11 and so on, according to Figure 3. The largest silhouette coefficient is generated at k = 2, so it is determined as the optimal number of clusters. The formation of clusters in the k-prototypes algorithm is also determined by the weighting coefficient (γ) according to Equation 3. Based on the processing results, the weighting coefficient (γ) is the same at each stage, the

number of clusters from $k = 2$ to $k = 20$, which is 2,1509. The value of the weighting coefficient (γ) is determined by the number of objects, the number of categorical variables, and numerical variables. The results of grouping the ClusterMix k-prototype for numerical variables are illustrated in Figure 4.

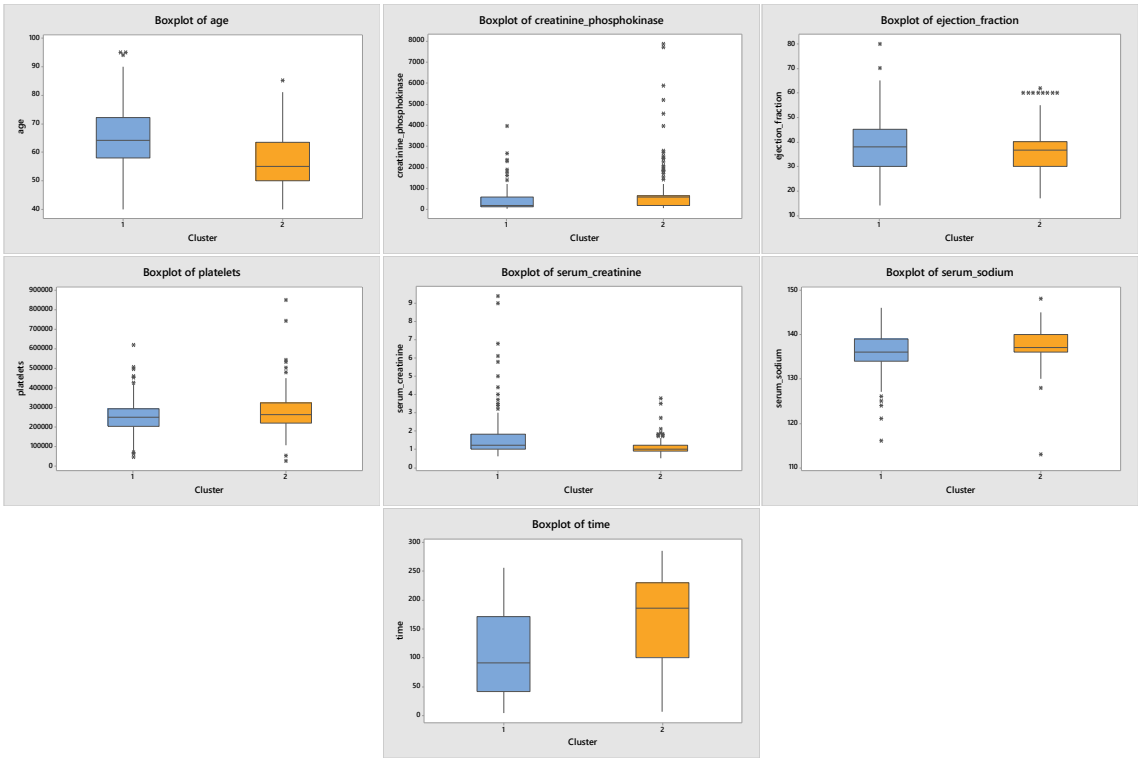

Figure 4. Boxplot k = 2 for Numeric Variables

Algorithm Cluster Mix k-prototype with $k = 2$ with a weighting coefficient of 2,1509. Two clusters were formed with 81 observations of cluster 1 members and 218 clusters 2 members. Visually visualizing the distribution in a boxplot shows that the difference is quite small in the distribution for numerical variables. The results of categorical variable grouping are illustrated in Figure 5.

The whole cluster formed will be compared with the parameters of the entire patient to identify the type of cluster formed, which is the cluster of patients at risk or patients who are not at risk of heart failure mortality factors. The summary of the comparison results is described as a whole for all variables in Table 2.

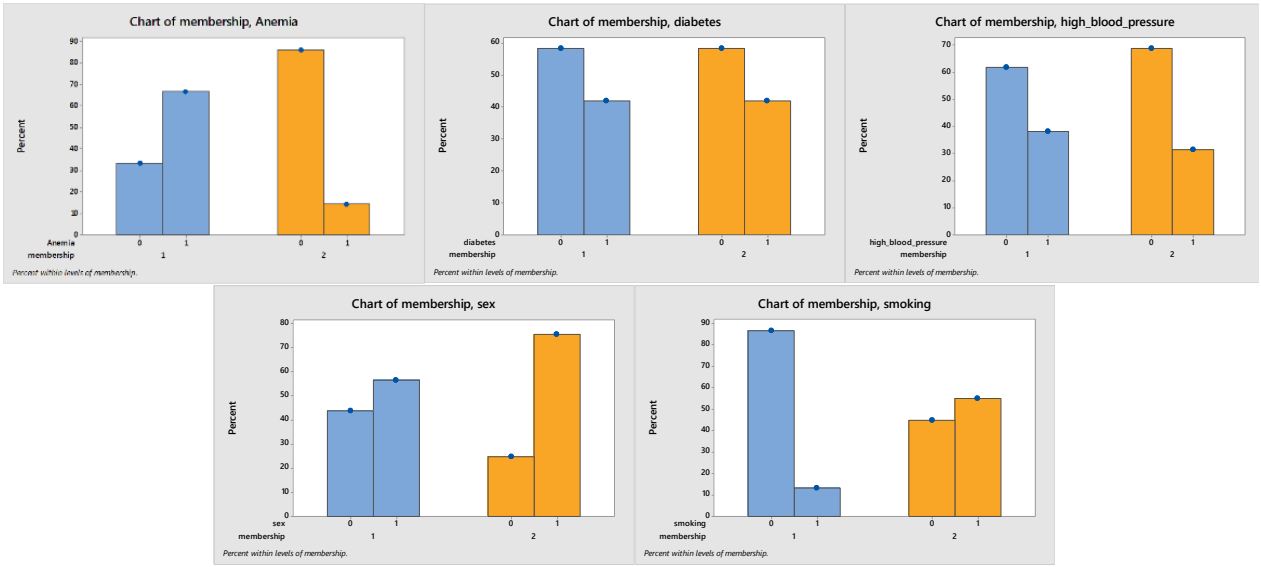

Figure 5. Barchart k = 2 for Categorical Variables

Table 2. Comparison of variables for all patients with clusters

Variable	Unit	All Patients	Cluster 1	Cluster 2
Usia	Years	60,83	64,711	56,06
Anemia	Boolean	0	1	0
High blood pressure	Boolean	0	0	0
Creatinine Phosphokinase (CPK)	mcg/L	581,84	387	822
Diabetes	Boolean	0	0	0
Ejection fraction	Percent	38,08	38,67	37,366
Gender	Binary	1	1	1
Platelets	kiloplatelets/mL	263.358,03	252.448	276.792
Serum creatinine	mg/dL	1,39	1,612	1,1255
Serum sodium	mEq/L	136,63	136,01	137,39
Smoking	Boolean	0	0	1
Time	Day	130,26	103,01	163,82

The age in cluster 1 showed an average of 64,71 years, and in cluster 2, it was 56,06 years, indicating in cluster 1, the patient's age was older than cluster 2. Young people are more often affected by heart failure due to lifestyle, lifestyle, heredity, and history of the disease, so the cluster formation tends to be the age variable as an indication of the lifestyle or lifestyle of patients with heart failure. Most of clusters 1 and 2 were dominated by patients with no congenital diseases such as blood pressure and diabetes, while anemia in cluster 1 was dominated by patients with congenital anemia, whereas in cluster 2, it was dominated by patients who did not have congenital anemia disease. In normal conditions, the CPK (Creatinine Phosphokinase) levels in the blood have a number of 20-200 mcg/L. In cluster 1, the dominant CPK condition of the patient is close to normal with an average of 387 mcg/L but in cluster 2 tends to be dominated by patients with low CPK levels. Tends to stay away from normal conditions with an average of 833 mcg/L. The ejection fraction has a normal limit of about 50-75 percent for adults. The lower it is, the less effective the heart's ability to pump blood. In cluster 1, it shows that the ejection fraction value is greater than in cluster 2, which is 38,08 percent and 37,36 percent so that both indicate that abnormal heart conditions pump blood throughout the body, but cluster 1 tends to be better versus cluster 2.

Platelets are the levels of platelets in the blood. Under normal circumstances, humans have a platelet level of 150.000-400.000 kiloplatelets / mL. High platelet levels in heart conditions cause various things, such as blood clots that can cause blood vessels to burst. Cluster 1 has a smaller platelet count than cluster 2, which is 252.448 kiloplatelets / mL and 276.792 kiloplatelets / mL, this causes cluster 1 to be in a pretty good condition compared to cluster 2.Creatine or serum creatine levels in humans are normal in men -men by 0,6 – 1,2 mg / dL and in women 0,5 – 1,1 mg / dL. High keratin levels usually occur in people who have kidney failure. Cluster 1 has higher levels of keratin than cluster 2, namely 1,612 mg / dL and 1,1255 mg / dL, so that cluster 2 is better than cluster 1.While sodium levels in the blood or serum sodium under normal conditions have a sodium level of 135 -145 mEq / L, so high levels of sodium have the potential for someone to easily experience hypogastrium. Cluster 1 has an average sodium level smaller than cluster 2, namely 136,01 mEq / L and 137,39 mEq / L, this causes cluster 1 to have a better sodium level than cluster 2. The patient's treatment time is getting better has long shown that the more complex the condition of a heart failure patient is. It can be seen in cluster 1, which has an average length of stay that is shorter than the average length of stay of patients in cluster 2. Most of the smoking habits are owned by cluster 2 compared to cluster 1, and this indicates that habits that can interfere with heart conditions are in cluster. 2 compared to cluster 1. Based on the information above, it can be concluded that, in general, cluster 1 is a characteristic of patients with low risk, while cluster 2 is a characteristic of patients with high risk.

4. Conclusion

The results of clustering observations using the k-prototypes algorithm from several experiments showed that the number of clusters formed was 2 clusters. Determination of the optimum cluster is by using a silhouette coefficient of 0,5777, which is used as an evaluation of the diversity within the cluster. The selection of the optimum cluster is based on the value of the largest silhouette coefficient in the number of other clusters. The results of the study were segmented patients with mortality characteristics of heart failure patients, which showed that cluster 1 was a group of patients who had a low risk of mortality due to heart failure, and cluster 2 was a group of patients with the characteristics of heart failure patients with a high risk of mortality due to failure. Heart. The segmentation is based on the mean value of each characteristic variable of heart failure mortality factor in each cluster, which is compared with normal conditions through variables serum creatine, ejection fraction, age, serum sodium, blood pressure, anemia, creatinine phosphokinase, platelets, smoking, type. Genital and diabetes.

In this study, there are suggestions regarding the differences in the characteristics between the two clusters so that there is a need for deeper analysis of medical action that refers to the characteristics of the two clusters so that it

can reduce the mortality rate due to heart failure. This study only uses cluster analysis so that other measurements using the mortality variable are needed to show the variables of importance to the mortality characteristics of heart failure patients so that these variables become determinants in identifying patients who are at risk of having heart failure. In this study, the method used is an unsupervised method, so it is necessary to compare it with a supervised method.

References

[1] J. Barallobre-Barreiro, Y.-L. Chung, and M. Mayr, "Proteomics and metabolomics for mechanistic insights and biomarker discovery in cardiovascular disease," *Rev. Española Cardiol. (English Ed.*, vol. 66, no. 8, pp. 657–661, 2013. https://doi.org/10.1016/j.rec.2013.04.009

[2] World Health Organization, "WHO."

[3] A. B. I. NATIONAL HEART, LUNG, "No Title."

[4] T. Ahmad, A. Munir, S. H. Bhatti, M. Aftab, and M. A. Raza, "Survival analysis of heart failure patients: A case study," *PLoS One*, vol. 12, no. 7, p. e0181001, 2017. https://doi.org/10.1371/journal.pone.0181001

[5] F. Meng *et al.*, "Machine learning for prediction of sudden cardiac death in heart failure patients with low left ventricular ejection fraction: study protocol for a retroprospective multicentre registry in China," *BMJ Open*, vol. 9, no. 5, p. e023724, 2019. https://doi.org/10.1136/bmjopen-2018-023724

[6] T. A. Buchan *et al.*, "Physician prediction versus model predicted prognosis in ambulatory patients with heart failure," *J. Hear. Lung Transplant.*, vol. 38, no. 4, pp. S381, 2019. https://doi.org/10.1016/j.healun.2019.01.971

[7] B. Chapman, A. D. DeVore, R. J. Mentz, and M. Metra, "Clinical profiles in acute heart failure: an urgent need for a new approach," *ESC Hear. Fail.*, vol. 6, no. 3, pp. 464–474, 2019. https://dx.doi.org/10.1002%2Fehf2.12439

[8] L. Chiodo, M. Casula, E. Tragni, A. Baragetti, D. Norata, and A. L. Catapano, "Profilo cardiometabolico in una coorte lombarda: lo studio PLIC. Cardio-metabolic profile in a cohort from Lombardy region: the PLIC study," *G. Ital. di Farm. e Farm.*, vol. 9, no. 2, pp. 35–53, 2017.

[9] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, pp. 16, 2020. https://dx.doi.org/10.1186%2Fs12911-020-1023-5

[10] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, "Improved automatic detection and segmentation of cell nuclei in histopathology images," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 841–852, 2009. https://doi.org/10.1109/TBME.2009.2035102

[11] P. Arora and S. Varshney, "Analysis of k-means and k-medoids algorithm for big data," *Procedia Comput. Sci.*, vol. 78, pp. 507–512, 2016. https://doi.org/10.1016/j.procs.2016.02.095

[12] T. S. Madhulatha, "Comparison between k-means and k-medoids clustering algorithms," in *International Conference on Advances in Computing and Information Technology*, 2011, pp. 472–481. https://doi.org/10.1007/978-3-642-22555-0_48

[13] R. Madhuri, M. R. Murty, J. V. R. Murthy, P. P. Reddy, and S. C. Satapathy, "Cluster analysis on different data sets using K-modes and K-prototype algorithms," in *ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India-Vol II*, 2014, pp. 137–144. https://doi.org/10.1007/978-3-319-03095-1_15

[14] J. Supranto, "Statistik Deskriptif." Jakarta: Airlangga, 1988.

[15] T. Ahmad, A. Munir, S. H. Bhatti, M. Aftab, and M. A. Raza, "Survival analysis of heart failure patients: A case study," *PloS one*, 2017.

[16] A. A. Mattjik, I. Sumertajaya, G. N. A. Wibawa, and A. F. Hadi, "Sidik peubah ganda dengan menggunakan SAS." 2011.

[17] S. Sharma and S. Sharma, "Applied multivariate techniques," 1996.

[18] S. G. Rao and A. Govardhan, "Performance validation of the modified k-means clustering algorithm clusters data," *Int. J. Sci. Eng. Res.*, vol. 6, no. 10, pp. 726–730, 2015.

[19] Z. Ansari, M. F. Azeem, W. Ahmed, and A. V. Babu, "Quantitative evaluation of performance and validity indices for clustering the web navigational sessions," *arXiv Prepr. arXiv1507.03340*, 2015.

[20] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987. https://doi.org/10.1016/0377-0427(87)90125-7

[21] N. J. Salkind, *Encyclopedia of measurement and statistics*. SAGE publications, 2006.

[22] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Min. Knowl. Discov.*, vol. 2, no. 3, pp. 283–304, 1998.

[23] R. A. Johnson and D. W. Wichern, *Applied multivariate statistical analysis*, vol. 5, no. 8. Prentice hall Upper Saddle River, NJ, 2002.

[24] G. Gan, C. Ma, and W. Jianhong, "Center-based clustering algorithms," *Data Clust. Theory, Algorithms Appl.*, 2007. https://doi.org/10.1137/1.9780898718348.ch9