Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2025, 2025:28
https://doi.org/10.28919/cmbn/9079
ISSN: 2052-2541

MODEL-BASED CLUSTERING APPROACH FOR CLUSTERING OF HEART DISEASE PATIENTS BASED ON RISK FACTORS

AYU SANGRILA[1], BUDHI HANDOKO[2,*], DEFI YUSTI FAIDAH[2]

[1]Post-Graduate Program of Applied Statistics, Universitas Padjadjaran, Bandung 45363, Indonesia

[2]Department of Statistics, Universitas Padjadjaran, Bandung 45363, Indonesia

Abstract: Cardiovascular disease causes the most non-communicable disease deaths or 17.9 million people each year. One of the efforts that can be made to reduce the threat of heart disease death is the grouping of heart disease patients based on their risk factors. The risk factors are age, blood pressure, cholesterol, and maximum heart rate. The data in this study is continuous data with mixed distribution and cluster uncertainty. This research suggests a model-based clustering approach based on finite mixture models. This approach assumes that the data is generated by a mixture of probability distributions, where each distribution represents different clusters with different parameters. Model-based clustering allows each individual to have a probability value to enter another cluster, making it suitable when there is cluster uncertainty and determining the number of clusters can be done automatically. Therefore, clustering is performed using model-based clustering to automatically determine the optimal number of clusters and identify cluster uncertainty. Parameters of the model are estimated using the Expectation Maximization (EM) algorithm. The best model selection is determined based on the Bayesian Information Criterion (BIC) values. The clustering results obtained the best model EEI with the optimal number of clusters as many as two Clusters.

Keywords: model-based clustering; finite mixture models; EM algorithm; heart disease.

2020 AMS Subject Classification: 62P10.

[*]Corresponding author
E-mail address: budhi.handoko@unpad.ac.id

Received December 18, 2024

1. Introduction

The development of science and technology in the health sector has encouraged experts to always conduct research on various diseases. Based on the nature of transmission, diseases can be divided into communicable diseases and non-communicable diseases [1], [2]. According to the World Health Organization, non-communicable diseases are known as chronic diseases that tend to last a long time and are caused by a combination of genetic, physiological, environmental, and behavioral factors [3]. Types of non-communicable diseases are cardiovascular diseases (such as heart attack and stroke), cancer, chronic respiratory diseases and diabetes [4], [5], [6]. Based on WHO data, non-communicable diseases kill 41 million people each year, equivalent to 74% of all deaths globally. Every year, more than 17 million people die because non-communicable diseases before the age of 70. Cardiovascular disease causes the most non-communicable disease deaths at 17.9 million people each year, followed by cancer (9.3 million), respiratory diseases (4.1 million), and diabetes (2.0 million) [7].

According to the National Center for Health Statistics, the leading cause of death in America is heart disease, with approximately one person dying every 33 seconds [8]. Heart disease in America on 2022 has killed around 702,880 people and on 2019 to 2022 has spend around $252.2 billion for medical services and lost productivity due to deaths [8], [9], [10]. The facts above indicate that efforts are needed to reduce the threat of death due to heart disease. One of strategy that can be done is to provide optimal and efficient treatment to heart disease patients based on their risk factors. Before treatment, identification of specific risk factors each patient is crucial because differences in patient characteristics can affect the treatment approach. The differences in risk factors for each patient's heart disease indicate that identifying the homogeneity of each patient's heart disease risk factors requires clustering.

Clustering of heart disease patients based on disease risk factors is expected to reduce medication errors that often occur, where based on a survey from the Institute for Healthcare Improvement, it was found that 28% of patients received the wrong dose of medication, 9% received the wrong medication, 18% received the wrong prescription, and 5% of patients accidentally took medication too much [11], [12]. These errors can certainly aggravate the patient's condition or cause serious complications, so it is important to do the right patient clustering. Based on the above, this study aims to cluster heart disease patients who have similar characteristics using multivariate normal model-based clustering.

2. DATA AND METHODS

2.1 Data

The data used in this study consist of 132 heart disease patients, with 4 variables, namely patient age, blood pressure, cholesterol, and maximum heart rate. The data is secondary data accessed from the Kaggle [13]. https://www.kaggle.com/datasets/nezahatkk/heart-disease-data.

2.2 Model-Based Clustering

Model-based clustering assumes that data comes from a mixture of several subpopulations represented by probability distributions [14], [15]. This assumption leads to a mathematical probability model for the data, the finite mixture model, where each component of the mixture model represents a different probabilistic distribution, and each distribution is assumed to represent a cluster in the data [16], [17]. The model-based clustering framework is developed based on the eigenvalue decomposition of the covariance matrix $\sum_g$.

$$\sum_g = \lambda_g \boldsymbol{D}_g \boldsymbol{A}_g \boldsymbol{D}_g^T \tag{2.1}$$

where:

λ_g: a scalar value that indicates the volume of the ellipse

$\boldsymbol{D}_g$: the eigenvector matrix of $\sum_g$, and denotes the orientation of the corresponding ellipse

$\boldsymbol{A}_g$: a diagonal matrix with elements that are proportional to the eigenvalues of the $\sum_g$ in descending order, it also shows the contours of the density function.

Table 1 represents a geometric interpretation of various parameterizations of the covariance matrix $\boldsymbol{\sum_g}$ in model-based clustering.

Table 1 Geometric Interpretation of Various Parameterizations of the Covariance Matrix in Model-Based Clustering (Source: [18]).

Identifier	Model	Distribution	Volume	Shape	Orientation
E		Univariate	Same	-	-
V		Univariate	Variable	-	-
EII	λI	*Spherical*	Same	Same	NA
VII	$\lambda_g I$	*Spherical*	Variable	Same	NA
EEI	λA	Diagonal	Same	Same	*Coordinate axes*
VEI	$\lambda_g A$	Diagonal	Variable	Same	*Coordinate axes*
EVI	λA_g	Diagonal	Same	Variable	*Coordinate axes*

Identifier	**Model**	**Distribution**	**Volume**	**Shape**	**Orientation**
VVI	$\lambda_g A_g$	Diagonal	Variable	Variable	*Coordinate axes*
EEE	Σ	Ellipsoidal	Same	Same	Same
VEE	$\lambda_g DAD^T$	Ellipsoidal	Variable	Same	Same
EVE	$\lambda DA_g D^T$	Ellipsoidal	Same	Variable	Same
VVE	$\lambda_g DA_g D^T$	Ellipsoidal	Variable	Variable	Same
EVV	$\lambda D_g A_g D_g^T$	Ellipsoidal	Same	Variable	Variable
EEV	$\lambda D_g A D_g^T$	Ellipsoidal	Same	Same	Variable
VEV	$\lambda_g D_g A D_g^T$	Ellipsoidal	Variable	Same	Variable
VVV	Σ_g	Ellipsoidal	Variable	Variable	Variable

Model-based clustering, which is based on normal multivariate finite mixture, has the following probability density.

$$p(\boldsymbol{Y}) = \sum_{g=1}^{G} \pi_g f_g(\boldsymbol{Y}|\boldsymbol{\theta_g}) \tag{2.2}$$

Where:

g : index of the number of clusters-$g = 1,2, \dots, G$

$\boldsymbol{Y}$: $\boldsymbol{Y_1}, , \dots, \boldsymbol{Y_n}$

$\boldsymbol{\theta_g}$: parameters of the density function of the cluster-g, $\boldsymbol{\theta_g} = (\boldsymbol{\mu_g}, \boldsymbol{\Sigma_g})$

$f_g(\boldsymbol{Y}|\boldsymbol{\theta_g})$: density function of the cluster-g,

$$f_g(\boldsymbol{Y}|\boldsymbol{\theta_g}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma_g}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\boldsymbol{Y} - \boldsymbol{\mu_g})^T \boldsymbol{\Sigma_g}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu_g})\right] \tag{2.3}$$

$\boldsymbol{\mu_g}$ **:** Means vector $(\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \dots \boldsymbol{\mu_G})^{\boldsymbol{T}}$

$(\boldsymbol{Y} - \boldsymbol{\mu_g})^T \boldsymbol{\Sigma_g}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu_g})$: squared of Mahalonobis distance between $\boldsymbol{Y}$ and $\boldsymbol{\mu_g}$ with $\boldsymbol{\Sigma_g}$ as the covariance matrix.

2.3 Gaussian Mixture Model

Gaussian Mixture Model (GMM) is a statistical model of the probability distribution obtained from the weight value of each Gaussian distribution [19]. Gaussian Mixture Model is used to determine the probability value of each observation object that will enter a particular cluster [20],

[21]. The size of the probability is measured by hidden variables. Figure 1 illustrates the parameters graphically.

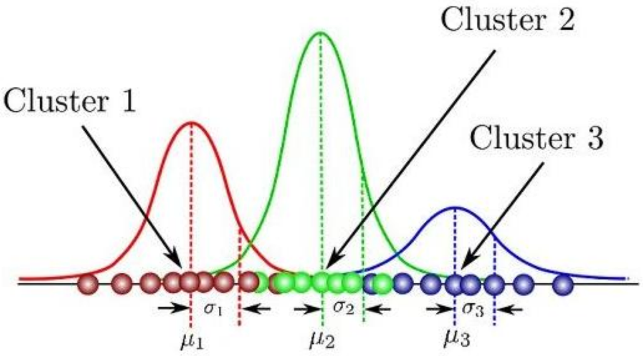

Figure 1 Three Parameters for Data Clustering [22].

The probability density function of the GMM can be written as follows.

$$p(\boldsymbol{Y}) = \sum_{g=1}^{G} \frac{\pi_g}{(2\pi)^{\frac{d}{2}} |\sum|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} \left(\boldsymbol{Y} - \boldsymbol{\mu_g} \right)^T \boldsymbol{\Sigma_g^{-1}} \left(\boldsymbol{Y} - \boldsymbol{\mu_g} \right) \right] \tag{2.4}$$

2.4 Parameter Estimator

Parameter estimator of model-based clustering generally use the Expectation-Maximization (EM) algorithm [23]. The expectation Maximization (EM) algorithm is a popular iterative algorithm that can be used to find parameter estimates by maximizing the loglikelihood function [23]. The first step before performing the EM process is to determine the initial value of the parameters to be used, namely the value of the mean μ_g, covariance $\sum_g$ and mixing coefficient π_g by applying the agglomerative hierarchical clustering method [24], [25].

The EM algorithm procedure runs in two stages, namely the Expectation (E) stage and the Maximization (M) stage [26]. The Expectation (E) stage aims to calculate the expectation value of the complete data log-likelihood function using its parameter estimator.

$$\hat{z}_{i,g}^{(s)} = (\frac{\hat{\pi}_g^{(s-1)} f_g\left(y_i \middle| \hat{\theta}_g^{(s-1)} \right)}{\sum_{h=1}^{G} \hat{\pi}_h^{(s-1)} \; f_h(\left(y_i \middle| \hat{\theta}_h^{(s-1)} \right)} \tag{2.5}$$

$$= (\frac{\hat{\pi}_g^{(s-1)} \phi_g \left(y_i \middle| \left(\widehat{\boldsymbol{\mu_g}}, \; \widehat{\boldsymbol{\Sigma_g}} \right)^{(s-1)} \right)}{\sum_{h=1}^{G} \hat{\pi}_h^{(s-1)} \; \phi_h \left(y_i \middle| \left(\widehat{\boldsymbol{\mu_h}}, \; \widehat{\boldsymbol{\Sigma_h}} \right)^{(s-1)} \right)}$$

The Maximization (M) stage aims to calculate the parameters that maximize the expected value of the log-likelihood function obtained in the E stage [27].

$$\hat{\pi}_g^s = \frac{\hat{\pi}_g^{(s-1)}}{n}, \quad \hat{\boldsymbol{\mu}}_g^s = \frac{\sum_{i=1}^{n} \hat{z}_{i,g}^{(s-1)} \boldsymbol{y_i}}{\hat{\pi}_g^{(s-1)}}, \quad \text{where} \tag{2.6}$$

$$\hat{n}_g^{(s-1)} = \sum_{i=1}^{n} \hat{z}_{i,g}^{(s-1)} \quad \text{, and}$$

$$\hat{\sum}_{\boldsymbol{g}}^{\boldsymbol{s}} = \frac{\sum_{i=1}^{n} \hat{z}_{i,g}^{(s-1)} (y_i - \hat{\boldsymbol{\mu}}_g^s)(y_i - \hat{\boldsymbol{\mu}}_g^s)^T}{\hat{n}_g^{(s-1)}}$$

The iteration of the EM algorithm will stop if the results have converged. After converging, cluster members can be grouped using the Maximum a Posteriori (MAP) classification method. Here are the MAP requirements.

$$MAP\left(\hat{z}_{i,g}^{(s)}\right) = \begin{cases} 1 & if\ max\ \hat{z}_{i,g}^{(s)}\ in\ cluster - gth \\ 0 & etc \end{cases}$$

2.5 Best Model Selection

The selection of a finite mixture model is related to determining the number of clusters to represent the clustering pattern in the data distribution. In the finite mixture model for clustering, selecting the best model that describes the data structure can be done through the likelihood approach and the Bayesian approach [28], [29]. In this study, the best model selection will be done using the Bayesian Information Criterion (BIC). BIC is derived from Bayesian principles, providing a probabilistic approach to model selection that takes into account the number of parameters [30]. The calculation of the BIC value of each model is done using the following formula:

$$BIC = 2\log p(\boldsymbol{y}|\mathcal{M}_g) \approx 2\log p\,(\boldsymbol{y}|\hat{\theta}_g\,, \mathcal{M}_g) - V_{\mathcal{M}_g}\log(n) \tag{2.7}$$

where

$p(\boldsymbol{y}|\mathcal{M}_g)$: the likelihood of the data for the model $\mathcal{M}_g$,

$p(\boldsymbol{y}|\hat{\theta}_g, \mathcal{M}_g)$: the likelihood of the data for the model $\mathcal{M}_g$

$V_{\mathcal{M}_g}$: number of free parameters estimated in the model $\mathcal{M}_g$,

$\hat{\theta}_g$: maximum likelihood estimates for parameters θ in the model $\mathcal{M}_g$

n : sample size

The best model is selected based on the smallest BIC value, which will later obtain the model shape and maximum number of clusters.

2.6 Stages of Analysis

The stages of analysis carried out in this study are as follows:

1. Collected data of heart disease patients with 4 risk factors obtained from the Kaggle
2. Data exploration, i.e. visualizing the data by looking at the scatter plot
3. Determine the model and determine the maximum number of clusters. This stage is done by looking at the BIC value
4. Perform cluster analysis
   a) Sets the initial values of the parameters used or parameter initialization
   b) Stage E, which calculates the conditional expected value of the complete data log-likelihood function using its parameter estimates
   c) Stage M, which calculates the parameters that maximize the expected value of the log-likelihood function obtained in stage E
   d) Repeat steps E and M until the iteration result converges
   e) Calculating the Maximum a Posteriori (MAP) value
5. Selection of the best model by looking at the smallest BIC value
6. Cluster heart disease patients and perform clustering visualization.

3. RESULTS

3.1 Data Exploration

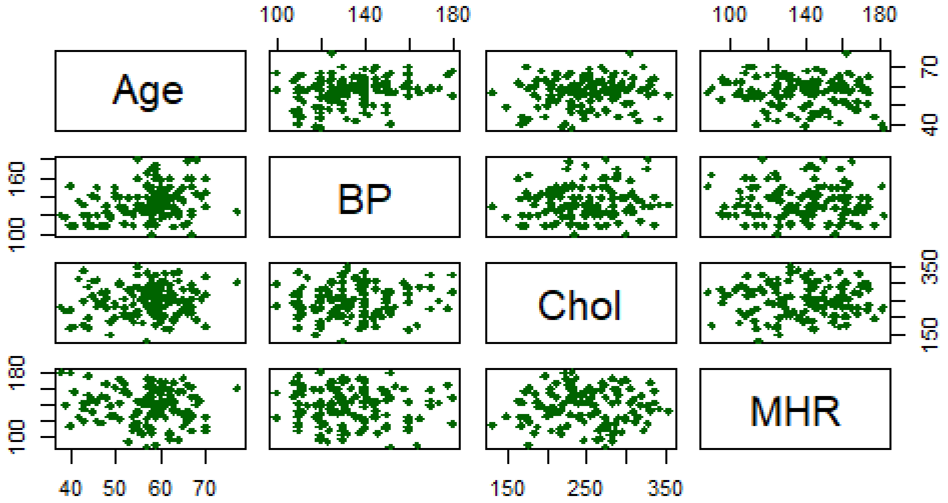

Figure 2 Scatter Plot of Heart Disease Patient Data

Figure 2 shows the distribution of data between variables. The pattern of data distribution between variables does not appear to be concentrated in one centralized pattern but rather randomly scattered without a clear center. On the diagonal distribution, some variables such as age and maximum heart rate are close to normal distribution, although not completely symmetrical. In contrast, variables such as cholesterol appear more asymmetrical, suggesting that the distribution may follow more than one distribution pattern. These patterns indicate that the data may come from several different distributions. Therefore, based on Figure 2, a multivariate normal model-based clustering approach was used in this study.

3.2 Identification of BIC Value

The identification of the Bayes Information Criterion (BIC) value is carried out to determine the best model and the number of clusters formed. The following is a figure of comparison of the BIC value for each model.

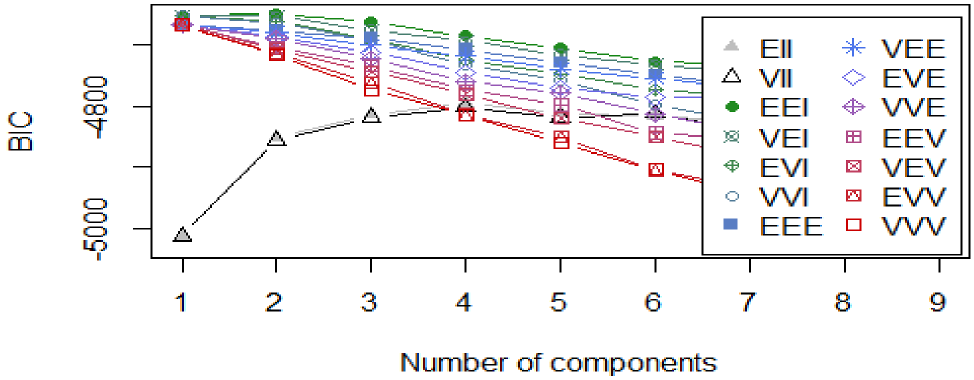

Figure 3 BIC value Each Model of GMM Result

Based on Figure 3, it is known that the smallest BIC value in the EEI model is - 4649.028, so the best model resulting from the analysis is the EEI model with a total of two Clusters. In the model-based clustering framework, the EEI model (equal volume, equal shape, Coordinate axes orientation) has a covariance matrix ($\sum_k$) which can be represented as:

$$\sum_k = \lambda \boldsymbol{A}$$

where:

λ is a scalar that represents that each cluster has the same volume.

$\boldsymbol{A}$ is a diagonal matrix that represents that each cluster has the same shape. This means that each dimension has approximately the same variance. Meanwhile, the orientation between clusters follows the main coordinate axis.

3.3 Clustering

Visualization of each cluster in the EEI model can be seen in the following figure.

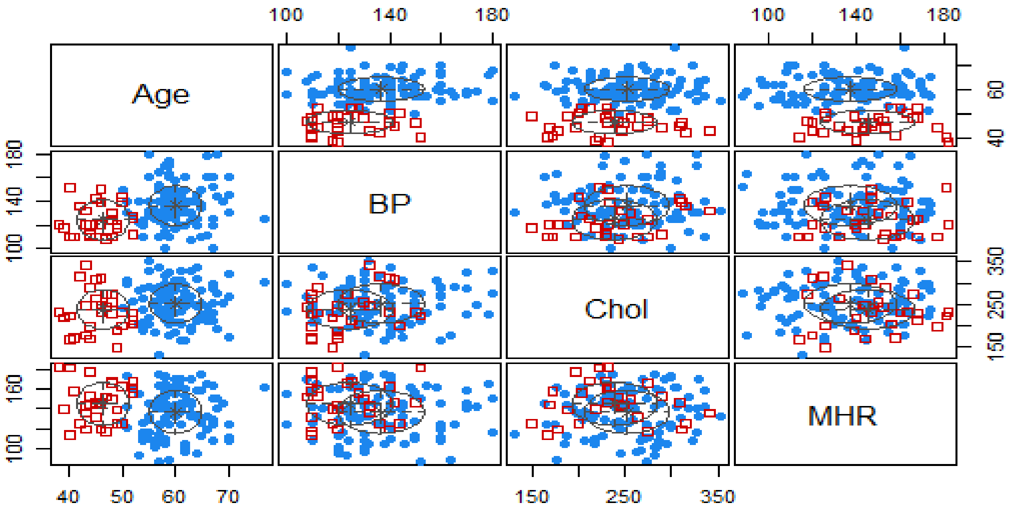

Figure 4 Visualization of Heart Disease Patient Clustering

Figure 4 shows the visualization of the clustering of heart disease patients which is divided into two Clusters. The density contour of each cluster is by the EEI model, namely each cluster is elliptical, the volume of the two Clusters is the same and the ellipses of the red and blue Clusters are parallel to the variable axis. Based on the clustering results, the members of Cluster 1 are 103 of heart disease patients and Cluster 2 is 29 of heart disease patients. Table 2 describes the Goodness of fit the EEI model.

Table 2 Components of the EEI Model

Log-likelihood	n	df	BIC
-2292.776	132	13	-4649.028

3.4 Cluster Characteristics

Once the cluster is formed, the characteristics of each cluster is known based on the mean value of each variable. The mean value of each variable in each cluster, is shown in Table 3.

Table 3 Mean Value of Each Variable

Variable	***Cluster* 1**	***Cluster* 2**
Age	60.0469	46.38628
Blood Pressure	136.6332	124.54922
Cholesterol	251.9913	238.63349
Maximum Heart Rate	137.2868	145.27006

Based on Table 3, it is obtained that the characteristics in cluster 1 are heart disease patients with risk factors of age, blood pressure and cholesterol. While the characteristics in Cluster 2 are heart disease patients with risk factors for maximum heart rate. The following is a boxplot figure of mean each variable in each Cluster.

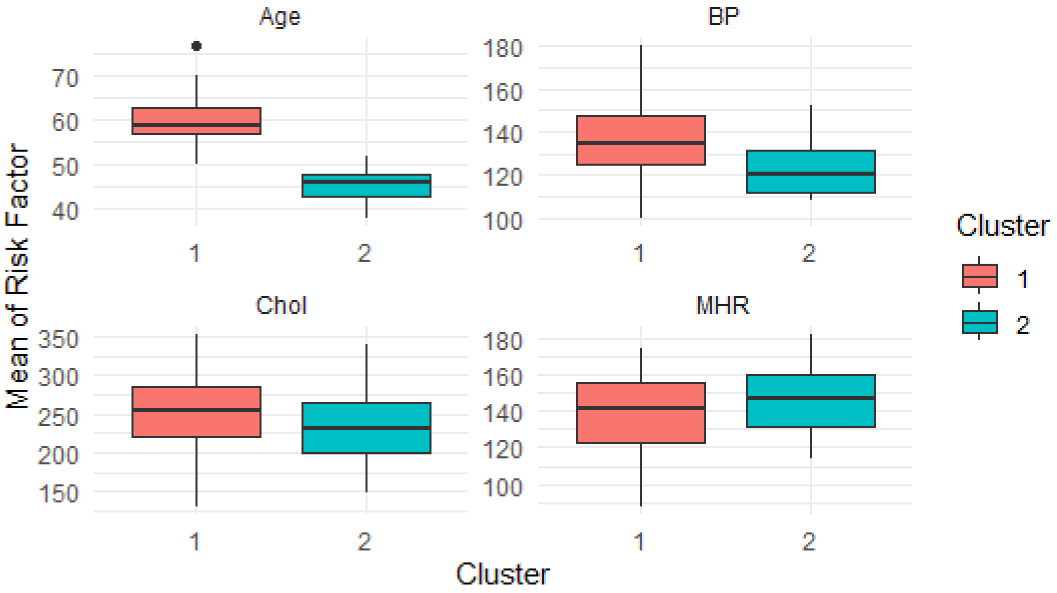

Figure 5 Boxplot of Mean Each Variable

Based on Figure 5, age, blood pressure, and cholesterol are located higher in Cluster 1, while only the maximum heart rate is high in Cluster 2. This shows that heart disease patients in Cluster 1 tend to be influenced by risk factors such as age, blood pressure and cholesterol. On the other hand, heart disease patients in Cluster 2 tend to be influenced by the risk factor of maximum heart rate. The relationship between age, blood pressure, cholesterol, and maximum heart rate to heart disease can be seen from the increasing age, the risk of developing heart disease is increasing. The higher the blood pressure, the higher the plaque buildup, which in turn increases the risk of heart disease. Higher cholesterol levels can cause fat accumulation in blood vessels, thus increasing the risk of heart disease, and an abnormal maximum heart rate can interfere with heart function and blood flow. Therefore, it is found that Cluster 1 consists of heart disease patients with characteristics of older age, high cholesterol, and high blood pressure, while Cluster 2 consists of heart disease patients with abnormal heart rate characteristics.

CONCLUSIONS

Based on the clustering results, two Clusters were formed with a Bayesian Information

Criterion (BIC) value of -4649.028 and the best model is the EEI model. Cluster 1 consists of 103 of heart disease patients with risk factors of older age, high blood pressure, and high cholesterol. Cluster 2 consists of 29 of heart disease patients with risk factors for abnormal maximum heart rate.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] S. Vadjdi, M. Farjam, Communicable Diseases and Non-Communicable Diseases: Which One Is the Priority in the Health Policies? Galen Med. J. 6 (2017), e851. https://doi.org/10.31661/gmj.v6i1.851.

[2] T.H. Tulchinsky, E.A. Varavikova, Communicable Diseases, in: The New Public Health, Elsevier, 2014: pp. 149–236. https://doi.org/10.1016/B978-0-12-415766-8.00004-5.

[3] W.G.J. Hol, C. Verlinde, Non-Communicable Diseases, in: International Tables for Crystallography, Wiley, 2006.

[4] WHO, Communicable and Noncommunicable Diseases, and Mental Health, Accessed: Dec. 04, 2024. https://www.who.int/our-work/communicable-and-noncommunicable-diseases-and-mental-health.

[5] D.J. Hunter, K.S. Reddy, Noncommunicable Diseases, N. Engl. J. Med. 369 (2013), 1336–1343. https://doi.org/10.1056/NEJMra1109345.

[6] R. Zhang, Communicable Diseases Epidemiology, in: C. Wang, F. Liu (Eds.), Textbook of Clinical Epidemiology, Springer, Singapore, 2023: pp. 197–211. https://doi.org/10.1007/978-981-99-3622-9_11.

[7] WHO, The Top 10 Causes of Death, Accessed: Dec. 04, 2024. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

[8] CDC, Heart Disease Facts, Accessed: Dec. 04, 2024. https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html.

[9] C.W. Tsao, A.W. Aday, Z.I. Almarzooq, et al. Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association, Circulation 147 (2023), e93–e621. https://doi.org/10.1161/CIR.0000000000001123.

[10] S.S. Martin, A.W. Aday, Z.I. Almarzooq, et al. 2024 Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association, Circulation 149 (2024), e347–e913. https://doi.org/10.1161/CIR.0000000000001209.

[11] NORC at the University of Chicago and IHI/NPSF Lucian Leape Institute. Americans’ Experiences with Medical Errors and Views on Patient Safety. Cambridge, MA: Institute for Healthcare Improvement and NORC at the University of Chicago; 2017.

[12] P. Anderson, T. Townsend, Medication Errors: Don’t Let Them Happen to You, Amer. Nurse Today, 5 (2010), 23–28.

[13] kaggle, Heart Disease Dataset, Accessed: Dec. 04, 2024. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset.

[14] C. Fraley, How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis, Computer J. 41 (1998), 578–588. https://doi.org/10.1093/comjnl/41.8.578.

[15] B. Grün, Model-Based Clustering, arXiv:1807.01987, (2018). https://doi.org/10.48550/arxiv.1807.01987.

[16] C. Fraley, A.E. Raftery, T.B. Murphy, et al. mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation, Technical Report No. 597, Department of Statistics, University of Washington, 2012.

[17] P.D. McNicholas, Model-Based Clustering, J. Classif. 33 (2016), 331–373. https://doi.org/10.1007/s00357-016-9211-9.

[18] C. Bouveyron, G. Celeux, T.B. Murphy, A.E. Raftery, Model-Based Clustering and Classification for Data Science: With Applications in R, Cambridge University Press, 2019. https://doi.org/10.1017/9781108644181.

[19] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum Likelihood from Incomplete Data Via the EM Algorithm, J. R. Stat. Soc. Ser. B Stat. Methodol. 39 (1977), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

[20] E. Patel, D.S. Kushwaha, Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model, Procedia Computer Sci. 171 (2020), 158–167. https://doi.org/10.1016/j.procs.2020.04.017.

[21] D.Y. Faidah, A.M. Hudzaifa, R.S. Pontoh, Clustering of Childhood Diarrhea Diseases Using Gaussian Mixture Model, Commun Math Biol Neurosci, 2024 (2024), 10. https://doi.org/10.28919/cmbn/8365.

[22] Built In, Gaussian Mixture Model Explained, Accessed: Dec. 04, 2024. https://builtin.com/articles/gaussian-mixture-model.

[23] S.K. Ng, T. Krishnan, G.J. McLachlan, The EM Algorithm, in: J.E. Gentle, W.K. Härdle, Y. Mori (Eds.), Handbook of Computational Statistics, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012: pp. 139–172. https://doi.org/10.1007/978-3-642-21551-3_6.

[24] L. Scrucca, A.E. Raftery, Improved Initialisation of Model-Based Clustering Using Gaussian Hierarchical Partitions, Adv. Data Anal. Classif. 9 (2015), 447–460. https://doi.org/10.1007/s11634-015-0220-z.

[25] L. Scrucca, C. Fraley, T.B. Murphy, et al. Model-Based Clustering, Classification, and Density Estimation Using Mclust in R, 1st ed., Chapman and Hall/CRC, Boca Raton, 2023. https://doi.org/10.1201/9781003277965.

[26] M.R. Gupta, Theory and Use of the EM Algorithm, Found. Trends Signal Process. 4 (2010), 223–296. https://doi.org/10.1561/2000000034.

[27] C. Fraley, A.E. Raftery, Model-Based Clustering, Discriminant Analysis, and Density Estimation, J. Amer. Stat. Assoc. 97 (2002), 611–631. https://doi.org/10.1198/016214502760047131.

[28] B. Zhou, J.H.L. Hansen, Unsupervised Audio Stream Segmentation and Clustering via the Bayesian Information Criterion, in: 6th International Conference on Spoken Language Processing (ICSLP 2000), ISCA, 2000: vols. vols. 3, 714-717–0. https://doi.org/10.21437/ICSLP.2000-635.

[29] J. Chen, Z. Chen, Extended Bayesian Information Criteria for Model Selection with Large Model Spaces, Biometrika 95 (2008), 759–771. https://doi.org/10.1093/biomet/asn034.

[30] I.A. Kieseppä, Statistical Model Selection Criteria and Bayesianism, Philos. Sci. 68 (2001), S141–S152. https://doi.org/10.1086/392904.