Vol. 3 No. (1) (June 2025) 42-47

Journal of Data Insights

http://journalnew.unimus.ac.id/index.php/jodi

Implementation of K-Means Algorithm to Group Age of Cardiovascular Disease Patients

Mulya Asy-syifa Rahmi [1], **Prizka Rismawati Arum**[2], **Tiani Wahyu Utami**[3]
[123]Universitas Muhammadiyah Semarang, Indonesia

DOI: https://doi.org/10.26714/jodi.v3i1.216

Article Information

Article History:
Submitted: 24th Dec 2024
Revised: 23th June 2025
Accepted: 27th June 2025

Keywords:
Cardiovascular; Clustering; Elbow Method; K-Means.

Abstract

Cardiovascular disease, including coronary heart disease, peripheral arteries and heart failure, is a serious disease that is the leading cause of death globally. Risk factors such as high blood pressure, dyslipidemia, smoking, diabetes, and obesity contribute to the development of this disease. This study aims to group cardiovascular disease sufferers based on age using the k-means clustering method with optimization of the k value using the elbow method. The data used comes from more than 35,000 preprocessed observations. The analysis results show that the optimal number of clusters is five. Data preprocessing succeeded in cleaning the data from missing values, and the elbow method helped determine the number of clusters that were relevant for age grouping of cardiovascular disease sufferers. The results of this grouping can be used for further analysis in efforts to prevent and manage cardiovascular disease.

✉ Correspondence Author:
E-mail: rarasyifaa16@gmail.com

e-ISSN: 2988 - 2109

INTRODUCTION

Diseases involving the heart and blood vessels, such as coronary heart disease, peripheral artery disease, heart valve disease, and heart failure, are known as cardiovascular diseases. These types of diseases can affect people of all ages. These diseases should not be left untreated or prevented. They can cause serious problems in other parts of the body. Worldwide, cardiovascular diseases are the leading cause of death (Pane et al., 2022), and risk factors such as hypertension, dyslipidemia, smoking, diabetes, and obesity contribute greatly to the development of these diseases.

Heart attack, stroke, heart failure, and heart rhythm disturbances are some of the serious complications that can be caused by cardiovascular disease. Risk factors include genetics, unhealthy lifestyle, and the environment. Controlling these risk factors is part of preventing and managing cardiovascular disease. This includes maintaining normal blood pressure, controlling cholesterol levels, quitting smoking, maintaining a healthy weight, and living an active lifestyle (Halodoc, 2021). The use of drugs, surgical procedures, and cardiac rehabilitation are some examples of treatments.

To reduce the mortality and morbidity caused by cardiovascular diseases, prevention and management of these diseases are very important. Morbidity is the incidence or prevalence of a disease or health disorder in a population over a certain period of time. It can be measured in various ways, such as the number of new cases diagnosed, the number of visits to health facilities, or the number of people experiencing certain symptoms or conditions. It is possible to reduce the burden of cardiovascular disease in the community by increasing awareness of risk factors and the importance of implementing a healthy lifestyle (Ministry of Health of the Republic of Indonesia, 2021).

One of the data mining processes is clustering. Clustering is a way to divide a large amount of data into small parts that have the same properties, locations, characteristics or other filters. There are two main categories of clustering techniques, namely hierarchical and non-hierarchical methods (Pandjaitan, 2017). Therefore, in this study, age grouping was carried out based on several factors.

Previous research by (Lashiyanti et al., 2023) stated that optimizing the k value using the elbow method in k-means clustering gave results that the use of the elbow method provided an indication of the most appropriate number of value groups. This study uses the k-means algorithm, which can group data based on the centroid or the closest cluster center point (Azwanti, 2018). Therefore, this study aims to determine the age grouping of cardiovascular disease sufferers using the elbow method optimization in the k-means clustering algorithm.

METHOD

2.1. Data Mining

Data mining, also known as Knowledge Discovery in Database (KDD), includes collecting, using historical data to find matches, and designing relationships in large data sets. Data mining produces results that can be used to make future decisions. One of the data mining techniques for grouping is also called clustering (Fauzi & Dana, 2023).

2.2. Clustering

Clustering, is one of the techniques in data mining that is known to be unsupervised learning. This shows that the characteristics of each cluster are not predetermined, but are based on the similarity of the characteristics of the group or cluster. In other words, clusters organize different data sets into groups or clusters based on the similarity of these characteristics. These comparable characteristics are displayed as points in a multidimensional space. In data mining, clustering consists of hierarchical and non-hierarchical (Kaamilah, 2023).

2.3. *K-means Clustering*

K-means is a non-hierarchical data grouping technique that aims to divide data into two or more groups, so that data that has different characteristics is grouped with others. The purpose of this data

grouping is to reduce the objective function set in the grouping, which usually aims to reduce variation in one group (Pandjaitan, 2017).

2.4. Metode *Elbow*

Determining the optimal k value can be done using the elbow method. The Elbow method focuses on the percentage of variance as a function of the number of clusters. The k value will be checked one by one and the *Sum Square Error* (SSE) value will be recorded to find the optimal k value (Kiat et al., 2020). The SSE equation is as follows

$$SSE = \sum_{i=1}^{n} (y_i - \mu_i)^2$$

Where SSE is the sum of square errors for the number of clusters k. N is the number of data, y_i is the i-th data point and μ_i is the cluster center closest to the i-th data point. The optimal number of clusters is usually considered when the decline in SSE begins to slow down. This is because data division becomes less important for minimizing SSE.

2.5. Data Source

The data used is secondary data obtained from the website www.kaggle.com. The number of observations used is 35,021 observations.

2.6. Research Variables

In the study, there are several variables used. Each of these variables is presented in Table 1.

Tabel 1. 2.6. Research Variables

No.	Variables	Information
1.	*Id*	-
2.	*Age*	Tahun
3.	*Height*	Cm
4.	*Weight*	Kg
5.	*Gender*	1: Perempuan
		2: Laki-Laki
6.	*Systolic Blood Pressure*	-
7.	*Diastolic Blood Pressure*	-
8.	*Cholesterol*	1: Normal
		2: Diatas Normal
		3: Jauh diatas Normal
9.	*Glucose*	1: Normal
		2: Diatas Normal
		3: Jauh diatas Normal
10.	*Smoking*	0: Tidak Merokok
		1: Merokok
11.	*Alcohol Intake*	0: Tidak Pengguna Alkohol
		1: Pengguna Alkohol
12	*Physical Activity*	0: Aktif
		1: Tidak Aktif

2.7. Research Steps

a. Conduct data collection
b. Perform data pre-processing
c. Determining the number of clusters using the elbow method
d. Determine the average age based on the clusters that have been formed
e. Visualizing the average age data

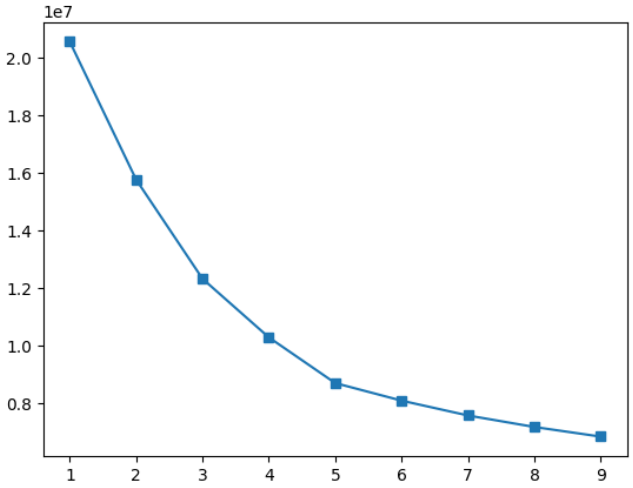
RESULTS AND DISCUSSION

The stage in grouping after collecting data is data pre-processing. Pre-processing is a way to change the raw data format into a useful format. This aims to make the data more ready for use in knowledge extraction..

```
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   id           35021 non-null  int64
 1   age          35021 non-null  int64
 2   gender       35021 non-null  int64
 3   height       35021 non-null  int64
 4   weight       35021 non-null  int64
 5   ap_hi        35021 non-null  int64
 6   ap_lo        35021 non-null  int64
 7   cholesterol  35021 non-null  int64
 8   gluc         35021 non-null  int64
 9   smoke        35021 non-null  int64
 10  alco         35021 non-null  int64
 11  active       35021 non-null  int64
```

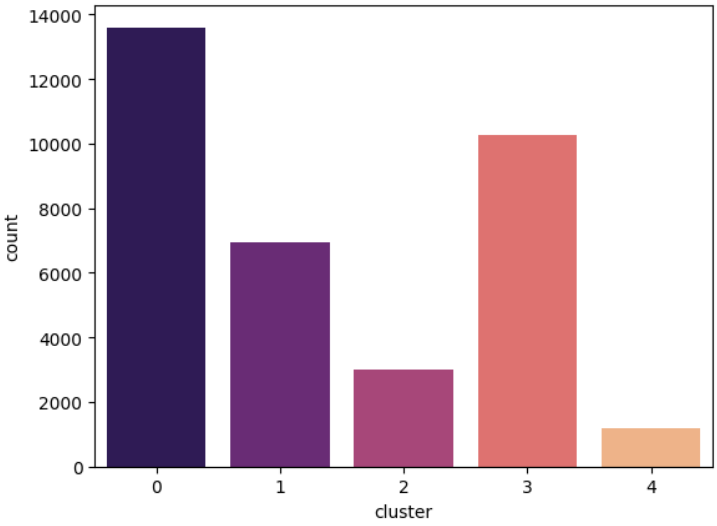
Picture 1. Data Pre-Processing Results

Figure 1 shows that all variables do not have missing values. The type of data used is an integer that represents whole numbers. In addition, there is no duplicate data so that the data is clean and the data format can be continued to the next stage. The next step is to determine the value of k using the elbow method.



Picture 2. Visualisasi Metode *Elbow*

Determination of the value of k by looking at the slope of the perfect SSE value. Figure 2 shows that at point 5 is the perfect SSE value. This is the optimal value for the number of clusters of 5. Therefore, in this study, the value of k = 5 is used as the number of clusters.



Picture 3. *Cluster Result*

Picture 3 shows that there are 5 clusters for age grouping. There are 13,602 observations grouped in cluster 0. There are 6,943 observations included in cluster 1. In cluster 2 there are 2,996 observations, 10,278 observations in cluster 3 and 1,202 observations in cluster 4.

CONCLUSION

Data pre-processing is an important initial step in the data grouping stage after collection which aims to change the raw data format to be more practical and ready to be used in the knowledge extraction process. The data uses integer and clean data types as shown in Figure 1. After pre-processing, the next step is to calculate the ideal number of clusters using the elbow method as shown in Figure 2. The optimal k value is 5, which indicates the right number of clusters for age grouping. The clustering results are depicted in Figure 3, where five clusters with different numbers of observations in each cluster. So the elbow method can help group age data into five relevant groups for further analysis.

REFERENCES

Azwanti, N. (2018). Segmentasi Tingkat Pemakaian Material dengan Data Mining *Cluster*ing. *Jurnal Komputer Terapan*, *4*(2), 16–27. http://jurnal.pcr.ac.id

Fauzi, I. A., & Dana, R. D. (2023). Implementasi Data Mining *Cluster*ing Dalam Mengelompokan Kasus Perceraian Yang Terjadi Di Provinsi Jawa Barat Menggunakan Algoritma K-Means. *Jurnal Riset Ilmu Manajemen Dan Kewirausahaan*, *1*(4), 58–72.

Halodoc. (2021, November 3). *Cara Menjaga Kesehatan Sistem Kardiovaskuler Tubuh*. Https://Www.Halodoc.Com/Artikel/Cara-Menjaga-Kesehatan-Sistem-Kardiovaskuler-Tubuh.

Kaamilah, L. L. (2023). Analisis Kelompok Lansia Berdasarkan Kategori Usia Dengan Metode K-Means *Cluster*ing. *Jurnal Riset Ilmu Akuntansi*, *2*(2), 1–14.

Kementerian Kesehatan Republik Indonesia. (2021, September 29). *Peringatan Hari Jantung Sedunia 2021: Jaga Jantungmu untuk Hidup Lebih Sehat* . Https://Ayosehat.Kemkes.Go.Id/Peringatan-Hari-Jantung-Sedunia-2021-Jaga-Jantungmu-Untuk-Hidup-Lebih-Sehat.

Kiat, A. B. H., Azhar, Y., & Rahmayanti, V. (2020). Penerapan Metode K-Means Dengan Metode *Elbow* Untuk Segmentasi Pelanggan Menggunakan Model RFM (Recency, Frequency & Monetary). *REPOSITOR*, *2*(7), 945–952.

Lashiyanti, A. R., Munthe, I. R., & Nasution, F. A. (2023). Optimisasi Klasterisasi Nilai Ujian Nasional dengan Pendekatan Algoritma K-Means, *Elbow*, dan Silhouette. *Jurnal Ilmu Komputer Dan Sistem Informasi (JIKOMSI*, *6*(1), 14–20.

Pandjaitan, B. (2017). *Cluster*ing Data Akademik Mahasiswa Fakultas Teknik USNI dengan Algoritma K-Means. *Jurnal Satya Informatika*, *2*(2), 1–8. https://doi.org/https://doi.org/10.59134/jsk.v2i2.425

Pane, J. P., Simorangkir, L., & Saragih, P. I. S. B. S. (2022). Faktor-Faktor Risiko Penyakit Kardiovaskular Berbasis Masyarakat. *Jurnal Penelitian Perawat Profesional*, *4*(4), 1183–1192. http://jurnal.globalhealthsciencegroup.com/index.php/JPPP