ISSN Online 3025-1605
Journal of Statistical Methods and Data Science
Volume 01 Number 2 December 2023
https://jurnalmipa.unri.ac.id/jsmds

K-MEANS CLUSTERING USING ELBOW METHOD IN CASE OF DIABETES MILLETUS TYPE II IN INDONESIA

Ayu Sofia[1*]
[1] Institut Teknologi Sumatera, Lampung, 35365, Indonesia

***e-mail**: *ayu.sofia@at.itera.ac.id*

Article Info:
Received: 08-06-2023
Accepted: 25-03-2024
Available Online: 05-04-2024

Keywords:
clustering
diabetes miletus
k-means

Abstract: *Indonesia is ranked 7th out of 10 countries with the most Diabetes Miletus sufferers. BPJS Health includes First Level Health Facilities (FKTP) where FKTP health facilities provide non-specialist individual health services and Advanced Referral Health Facilities (FKRTL) which means health facility provides specialist or sub-specialty individual health services. Based on the data sample diabetes mellitus Type 2 is one of the ten most common diagnoses at FKRTL visits and ranks third after follow-up examinations after treatment for conditions other than malignant neoplasms and kidney failure with a percentage of 3.54% and a total of 62,455 for 2019 to 2020. In deciding policies related to the funding of BPJS participants who suffer from diabetes mellitus, it's necessary to have the characteristics of each region so that policymaking is more appropriate. The method used in this study uses clustering analysis using the K-means algorithm for type II diabetes mellitus in Indonesia from 2015 to 2020 by province. Based on the outcome of the clustering provinces in Indonesia using the K-Means algorithm with optimization the number of clusters using the elbow method formed 3 clusters. 1st cluster has 21 provinces, the 2nd cluster has 9 provinces and the 3rd cluster has 4 provinces.*

1. INTRODUCTION

Diabetes mellitus (DM) is interpreted as an illness or chronic metabolic disorder with multiple etiologies characterized by high blood sugar levels that go along with disturbances in carbohydrate, lipid, and protein metabolism as a result of insulin function insufficiency. Insufficiency of insulin function can be caused by interference or deficiency of insulin production by Langerhans beta cells of the pancreas gland, or caused by a lack of responsiveness of the body's cells to insulin. (WHO, 2023). Indonesia is ranked 7th out of 10 countries by 11.3% with the highest number of sufferers, 10.7 million (Kemenkes,2020).

In general, there are 2 types of diabetes mellitus, namely diabetes mellitus type 1 (DMT1) and diabetes mellitus type 2 (DMT2). Based on the explanation from the Ministry of Health of the Republic of Indonesia, DMT1 is a disease caused by the cessation of the pancreas producing insulin. The body's inefficient use of insulin causes DMT2, also known as non-insulin-dependent or adult-onset (PERKENI,2021). DMT2 affects more than 95% of those with this disease. This type of diabetes is mainly caused by the increasing body weight and the less active activities that people have. Although often not very serious, the indications of DMT2 may be comparable to those of DMT1. As a result, the condition may not be detected until DMT2 has developed into a serious complication (Kemenkes,2020). those who consume at least one type sweet fizzy drinks every day willhave twice the risk of developing T2DM bigger than that rarely consume it. No doubt that nutrition is an important factor in the emergence of T2DM. Lifestyle Westernized and relaxed living constitute factors that increase the prevalence of DM. Healthy daily eating patterns and balance needs to be considered, so can maintain ideal body weight (Nuraini, 2019)

Every Indonesian citizen is advised to become a participant so that the illness can be covered by BPJS Kesehatan shall be the Social Health Security Agency (BPJS). Benefits that BPJS Health can bear include First Level Health Facilities (FKTP) and Advanced Referral Health Facilities (FKRTL). FKTP is a health facility providing non-specialist individual health services for observation, diagnosis, treatment, treatment, and/or other health services. FKRTL is a health facility that provides specialist or sub-specialty individual health services which include advanced-level outpatient care, advanced-level inpatient care, and inpatient care in special care rooms (Kemenkes, 2014). Type 2 diabetes mellitus is one of the ten most common diagnoses at FKRTL visits and ranks third after follow-up examinations after treatment for conditions other than malignant neoplasms and kidney failure with a percentage of 3.54% and a total of 62,455 for 2019 to 2020 (BPJS,2020).

The International Diabetes Federation (IDF) states that diabetes mellitus is the 7th leading cause of death in the world where the proportion of type 2 diabetes mellitus is 95% of the world's population suffering from diabetes mellitus (IDF,2014). So, the government needs to know which provinces have high and low diabetes mellitus rates in each region. as a form of prevention which will later be applied in the form of policies or quick steps to reduce diabetes mellitus, likewise with BPJS Health. In deciding policies related to the funding of BPJS participants who suffer from diabetes mellitus, it is necessary to have the characteristics of each region so that policymaking is more appropriate. Previous research was grouping using hierarchical clustering, the groupings formed in this research resulted in 2 clusters. Cluster 2 consists of the provinces of East Java, West Java, Central Java, and Jakarta. Meanwhile, 30 other provinces are members of Cluster 1 (Sulthoni,2023). The author thinks that the cluster formed is still not optimal because the differences between the members are very large. The algorithm or method that can be used in grouping an object is the K-Means algorithm. K-Means is a method with the user determining the number of clusters that need to be grouped in the dataset and determining the centroid for each cluster. It is hoped that this research can provide an overview regarding the distribution of participants diagnosed with DMT2 entering FKRTL services and their grouping so that it can be used as a reference for developing prevention programs that are more targeted and effective for planning and managing FKRTL in each province.

2. LITERATURE REVIEW

2.1. Elbow Method

The elbow method is a method used to decide the best number of clusters through percentage comparison of the results of calculating the Sum of Square Error (SSE) value in each number of clusters between the number of clusters that will be formed elbow at a point resulting from the graph line. This method uses the percentage results with the biggest difference which is illustrated via a graph. When the value of the first cluster and the value of the second cluster show an angle or angle on the graph or have a value with the largest decrease, then at the corner point formed that is the most appropriate number of clusters or it can be said to be the optimum number of clusters (Bholowalia, 2014). To produce a comparative value, a Sum of Square (SSE) calculation is performed for each cluster value. The Sum of Square (SSE) formula is as follows (Dewi and Pramita,2019):

$$SSE = \sum_{k=1}^{k} \sum_{x_{i \in Sk}} \|x_i - C_k\|_2^2$$

x_i : attribute value of the *i-th* data
C_k : attribute value of the center *k-th cluster* cluster

2.2. K-Means Clustering

Clustering is a process of creating a group of a dataset into clusters that have similar characteristics (Han,2012). One way to create a cluster is to use K-Means clustering. K-Means Clustering is a data analysis method or Data Mining method that performs unsupervised learning modeling processes and uses methods that classify data from various partitions. K-means clustering is done to minimize the variation in data within a cluster and maximize the variation in data in other clusters (Umargono,2019). K-means clustering needs the number of *k*, where *k* is the number of clusters. The algorithm of K-means clustering divides the object or data into one cluster based on the similarity of attributes owned. The level of similarity between the object can be known by applying distance measurements. One of the methods that are often used to calculate distance measurement is Euclidean Distance. For 2 *x* and *y* data points in the d-dimension of the data, the calculation of distance using Euclidean distance is formulated with Equation (Ediyanto,2013).

$$d_{euc}(x, y) = \sqrt{\sum_{j=1}^{d} \left(x_j - y_j\right)^2}$$

Where: x_j, y_j = value of j attribute
$d_{euc}(x, y)$ = the value of distance

3. METHODOLOGY

3.1. Data and Sources

The data that used in this research is secondary data obtained from the Social Health Security Agency or BPJS. from Sample Data 2015 to 2020. The data is in the form of diabetes mellitus contextual sample data at Advanced Referral Health Facilities (FKRTL) services in BPJS. The data that used is in monthly period for five years started from January 2015 to December 2020 in 34 provinces in Indonesia.

3.2. Method

The method used in this study uses clustering analysis using the k-means algorithm for type II diabetes mellitus in Indonesia from 2015-2020 by province. After the implementation of the k-means algorithm, testing will be carried out using the elbow method in decide the best number of clusters.

3.3. Data Analysis

The basic process of the k-means algorithm can be seen as follows (Rohmawati et.al 2015):

i. Determine the number of clusters you want to form (k).

ii. Identify the centroid value. determining the initial centroid value taken from existing data taken randomly for the initial iteration taken from existing data. When calculating the centroid value, which is the iteration stage, use the following formula:

$$V_{ij} = \frac{1}{Ni}\sum_{K=0}^{Ni} X_{kj}$$

V_{ij} :the i-th cluster centroid/average for the j-variable

Ni :the amount of the data that is a member of the i-th cluster

k :the index of clusters

j :the index of the variable X

X_{kj} :the k-th data value in the cluster for the j-variables

iii. Calculating the distance between the centroid and the point of each object so that the closest distance is found for each data using the Euclidean Distance equation:

$$d_{euc}(x,y) = \sqrt{\sum_{j=1}^{d} (x_j - y_j)^2}$$

Where: x_j, y_j is the value of j attribute

iv. Calculate the minimum distance objects to form cluster members. The value obtained in the distance matrix is 0 or 1, where the value is 1 for data allocated to clusters, and the value of 0 is for data allocated to other clusters. Grouping objects based on their proximity to the centroid (smallest distance).

v. Determination of the new cluster center, namely determining the centroid value at the iteration stage so that a new centroid value is obtained. Repeat until got the centroid value does not change and the cluster members do not move to another cluster, then the process is complete.

Elbow Method Algorithm in determining the value of *k* on K-Means Clustering (Febrianti, 2018):

i. Initial initialization of the value of K

ii. Increase the value of K

iii. Calculate the sum of square error results for each value of K

iv. Look at the results of the sum of square error from the value of K which drops drastically

v. Set an angled K value

4. RESULTS AND DISCUSSION

Diabetes does not only cause premature death in Indonesia. This disease is also the main cause of various other diseases such as blindness, heart disease and kidney failure. Indonesia is the only country in Southeast Asia that has the highest number of diabetes patients. From 2015 to 2020, the number of people who suffers diabetes mellitus continues to increase but has experienced a decline in 2020.

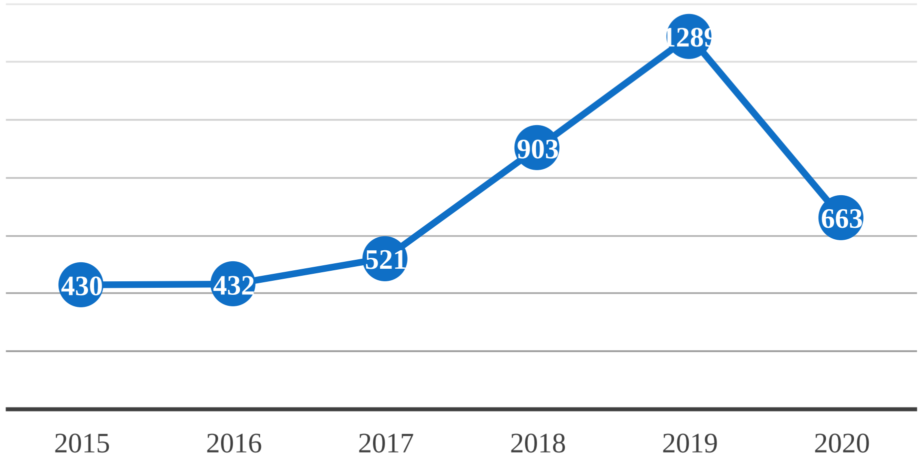

Figure 1. The Average Number of Participants Diagnosed with DM II In Indonesia

Participants diagnosed in Indonesia based on 34 existing provinces from 2015 to 2020 gradually increased and decreased in 2020. From 2015 to 2017 the average number of participant DM II do not have much different. In 2018 and 2019 the participant that diagnosed DM II increased dramatically by 903 participants in 2018 and 1289 participant in 2019. In 2020 it fell to 663 participant that suffered DM II. There are 3 top provinces based on the number of participants diagnosed with DM II for each year, namely West Java, Central Java and East Java . The resulting pattern is the same for each year, which has increased from 2015 to 2019 and has decreased in 2020 as can be seen in figure 2.

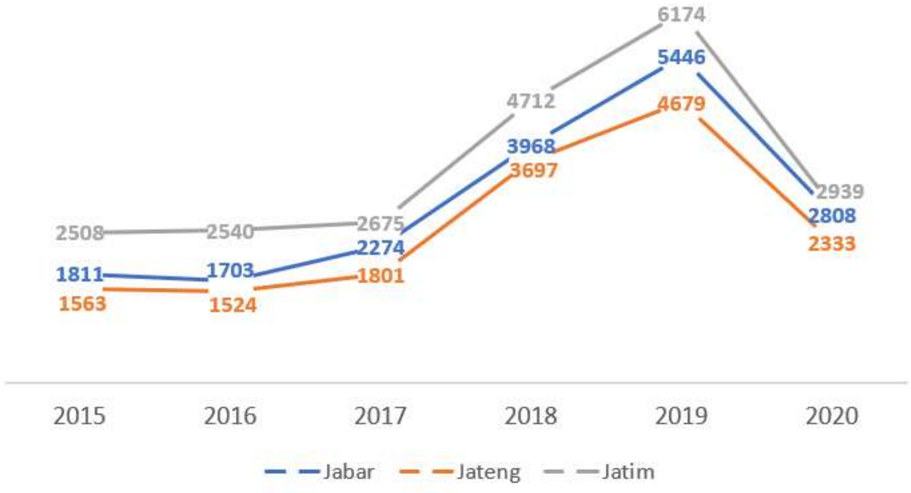

Figure 2. Three Top Province based on The Number of Participants Diagnosed with DM II

Based on Figure 2, East Java had the highest number of participants diagnosed with DM II starting with 2508 participants in 2015 and increasing to 6174 participants in 2019. In 2020, the number of participants diagnosed with DM II is much reduced from the previous year, namely 2939 participants. The same pattern occurs in Central Java and West Java. West Java has 1811 participants in 2015, increasing to 5446 participants in 2019. In 2020, decreased from the previous year, with 2808 participants. Central Java is the third-ranked with the number of participants diagnosed with DM II, 1563 participants in 2015 and increased to 4679 participants in 2019. In 2020, drop dramatically from the previous year,2333 participants.

K-means clustering shows that three clusters were formed based on Figure 3. The analysis in finding the number of clusters to be used in this case is to use the elbow method which can be seen from the largest decrease in SSE values that form the corner angles, so that later the evaluation of the results is to determine the best number of clusters in the study can be seen in Figure 3.

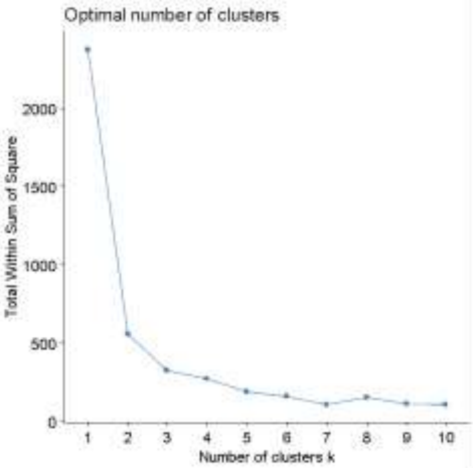

Figure 3. Optimal Cluster based on Elbow Method

As shown in Figure 3, the graph shows the most drastic decrease in the Sum of Square (SSE) value referring to the number of clusters formed that can be used, namely at point 3 with the difference in SSE at the previous point, namely point 2. then SSE gradually decreases without a drastic spike towards the lowest point so as to make the angle between the number of clusters 2 and 3 which from the inspection in the graph forms a right angle, so that the number of clusters 3 is the best number of clusters to use used in this study.

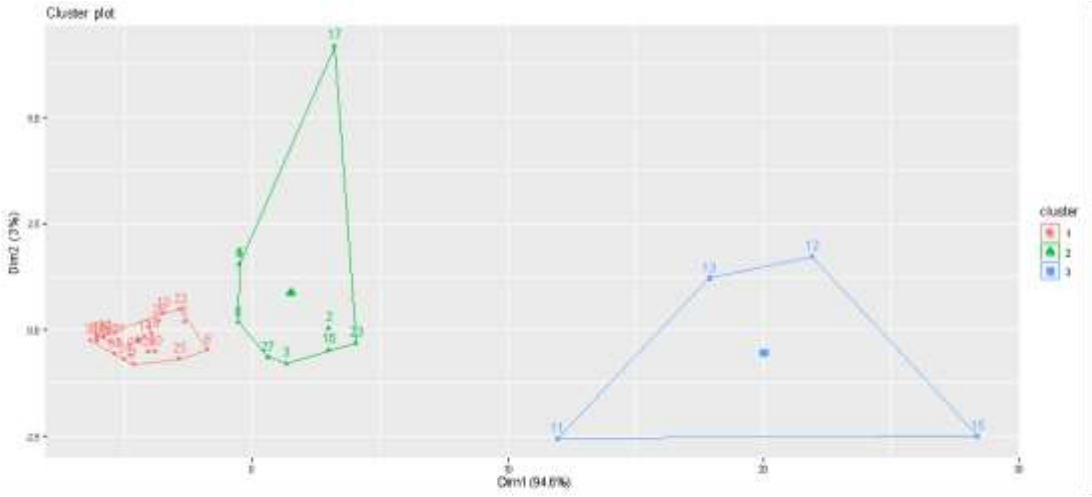

Figure 4 . Visualization of K-Means Clustering

On figure 4, the color indicates the cluster, performs a number of clusters is 3 clusters, the red point is 1st cluster has 21 provinces, the green point is the 2nd cluster has 9 provinces and the blue point is the 3rd cluster has 4 provinces. Based on the k-means clustering, the member of the 3rd cluster is the province that has the largest number of participants diagnosed with DM II in Indonesia.

Table 1. Member of Each Cluster

Cluster	n	Details
Cluster 1	21 provinces	1(Aceh), 6(Sumatera Selatan), 7(Bengkulu), 9(Kepulauan Bangka Belitung), 10(Kepulauan Riau), 14(DIY), 18 (Nusa Tenggara Barat), 19 (Nusa Tenggara Timur), 20(Kalimantan Barat), 21(Kalimantan Tengah), 22(Kalimantan Selatan), 25 (Sulawesi Utara), 28(Sulawesi Tenggara), 29(Gorontalo), 30(Sulawesi Barat), 31(Maluku), 32(Maluku Utara), 33(Papua Barat), 26 (Sulawesi Tengah), 24(Kalimantan Utara), And 34(Papua)
Cluster 2	9 provinces	3(Sumatera Barat), 2 (Sumatera Utara), 4(Riau), 8(Lampung), 16(Banten), 17(Bali), 23(Kalimantan Timur), 27(Sulawesi Selatan), And 5(Jambi)
Cluster 3	4 provinces	13(Jawa Tengah) 12(Jawa Barat) 15(Jawa Timur) Dan 11 (Jakarta)

As can be seen in the Table 1 member of each cluster, Cluster 1 has 21 provinces, which almost that province come from east of Indonesia which means that province has the smallest number of participants diagnosed with DM II if we compare with another cluster. Cluster 2 has 9 provinces, that almost that province come from west of Indonesia, like West Sumatera, North Sumatera, Riau, Lampung, Banten, and Jambi. We can conclude that the 3rd cluster is dominated by the province with the highest number of participants diagnosed with DM II. West Java, Central Java and East Java is the top 3 province that has the highest number of participants in Indonesia.

5. CONCLUSION

Based on the result of the clustering data experiment using the K-Means algorithm with optimization of the determination of the best number of clusters using elbow method is 3clustersr. 1st cluster has 21 province, the 2nd cluster has 9 province and the 3rd cluster has 4 province. With the formation of these 3 clusters, it is hoped that they will provide benefits in policy making in dealing with diabetes mellitus both in terms of treatment and financing that will be issued by BPJS Kesehatan for each cluster that is formed. Because it is expected that each cluster has the same criteria so that it is easier in terms of concluding which can represent each cluster.

REFERENCES

BPJS.2020. Data Sampel BPJS Kesehatan 2015-2020. Jakarta.

Bholowalia, Purnima & Kumar, Arvind. (2014). EBK-Means: A Clustering Techiniques based on Elbow Method and K-Means in WSN. *International Journal of Computer Application*. IX(105), 17-24. https://doi.org/10.5120/18405-9674

Dewi, D. A. I. C. and Pramita, D. A. K. (2019). Analisis Perbandingan Metode Elbow dan Sillhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali. *Jurnal Matrix*. 9(3). http://dx.doi.org/10.31940/matrix.v9i3.1662.

Ediyanto, Mara, M.N. & Satyahadewi, N. (2013). Pengklasifikasian Karakteristik Dengan Metode K-Means Cluster Analysis. *Buletin Ilmiah Mat. Stat. dan Terapannya*. II(2),133-36. ISSN: 2302-9854.

Febrianti, A. F., Cabral, A. H., Anuraga, G. (2018). K-means Clustering Dengan Metode Elbow Untuk Pengelompokkan Kabupaten dan Kota Di Jawa Timur Berdasarkan Indikator Kemiskinan. 863- 870. http://dx.doi.org/10.46984/sebatik.v26i2.2134.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques (Third Edition)*. Waltham. MA: Morgan Kaufmann Publishers.

Infodatin Pusat Data dan Informasi Kementerian Kesehatan RI. (2020). *Tetap Produktif, Cegah, dan Atasi Diabetes Melitus*. Jakarta. ISSN 2442-7659.

International Diabetes Federation (IDF). 2014. IDF Diabetes Atlas. Available from: https://idf.org/e-library/epidemiologyresearch/diabetes-atlas.html.

Kementerian Kesehatan RI. (2014). *Peraturan Menteri Kesehatan Republik Indonesia No. 28 Tahun 2014 Tentang Pedoman Pelaksanaan Jaminan Kesehatan Nasional*. Jakarta : Kementerian Kesehatan RI.

Nuraini HY, Supriatna R. (2019) Hubungan Pola Makan, Aktivitas Fisik Dan Riwayat Penyakit Keluarga Terhadap Diabetes Melitus Tipe 2. *Jurnal Ilmu Kesehatan Masyarakat*. 5(1):5-13.

PERKENI. (2021). *Petunjuk Praktis Terapi Insulin Pada Pasien Diabeter Melitus*. Jakarta - PB. PERKENI.

Rohmawati, N., Defiyanti, S., & Jajuli, M. (2015). Implementasi Algoritma K-MEANS dalam Pengklasteran Mahasiswa Pelamar Beasiswa. *Jurnal Ilmiah Teknologi Informasi Terapan*. 1(2), 62-67.

Sulthoni, H.S. and Sofia, Ayu. 2023. Prediction of FKRTL service Diagnosed with Type 2 Diabetes Mellitus Using the Hierarchial Agglomerative Clustering Time Series Method. *International Journal of Scientific Research in Science, Engineering and Technology*. 10(6):144-156. http://dx.doi.org/10.32628/IJSRSET231064.

Umargono, E, Suseno,J.E & Gunawan, S.K.V. 2019. Kmeans Clustering Optimization Using the Elbow Method and Eraly Centroid Determination Based on Mean and Median Formula. *Advances in Social Science, Education and Humanities Research*. 474. http://dx.doi.org/10.2991/assehr.k.201010.019.

WHO. (2023). Available: https://www.who.int/news-room/fact-sheets/detail/diabetes. [Accessed 07 Juli 2023].