International Journal of Scientific Research in Science, Engineering and Technology

Print ISSN - 2395-1990
Online ISSN : 2394-4099

OPEN ACCESS

Available Online at : www.ijsrset.com
doi : https://doi.org/10.32628/IJSRSET231064

Prediction of FKRTL Services Diagnosed with Type 2 Diabetes Mellitus Using the Hierarchical Agglomerative Clustering Time Series Method

Hana Sabrina Sulthoni*, Ayu Sofia

Program Studi Sains Aktuaria, Jurusan Sains, Institut Teknologi Sumatera, Lampung Selatan, Indonesia

ARTICLEINFO

Article History :

Accepted: 02 Nov 2023
Published: 20 Nov 2023

Publication Issue :
Volume 10, Issue 6
November-December-2023
Page Number :
144-156

ABSTRACT

Diabetes Mellitus (DM) is a disease characterized by hyperglycemia or increased blood sugar levels and metabolic disorders. DM is a disease with a high prevalence rate in Indonesia. DM is included in the 10 most frequent Advanced Referral Health Facility (FKRTL) visits diagnoses. This research aims to find out provincial groupings based on data from FKRTL visits of participants diagnosed with T2DM using the Hierarchical Agglomerative Clustering method (single linkage, complete linkage, average linkage) and to find out predictions of participants diagnosed with T2DM at FKRTL visits using the Autoregressive Integrated Moving Average (ARIMA) method. The method used is Hierarchical Agglomerative Clustering Time Series. This research shows that the best algorithm is average linkage with a number of clusters of two. Forecasting provides a pattern that tends to decline.

Keywords: Diabetes Mellitus, *Hierarchical Agglomerative Clustering Time Series,* Advanced Referral Health Facility (FKRTL)

I. INTRODUCTION

Technological developments and social life impact all aspects of life, one of which is health-related issues. The most common health-related case in Indonesia is diabetes mellitus. Diabetes is a serious condition caused by insufficient insulin production by the pancreas. Diabetes Mellitus (DM) is a disease characterized by hyperglycemia or increased blood sugar levels as well as disorders of protein, carbohydrate and fat metabolism due to abnormalities in insulin secretion, insulin action or both [1]. Diabetes is divided into several types, including type 1 diabetes mellitus (T1DM), type 2 diabetes mellitus (2DM), gestational diabetes, and other diabetes caused by drug use and other diseases. Based on an explanation from the Ministry of Health of the Republic of Indonesia, T2DM is DM caused by ineffective insulin action. Based on the International Diabetes Federation in 2019, Indonesia entered the 10 countries with the highest number of diabetes sufferers in the world with a total of 10,700,000 sufferers. Seeing the large number of diabetes cases in Indonesia, the Ministry of Health of the Republic of Indonesia prioritizes treating diabetes mellitus among other metabolic disorders. Currently, diabetes mellitus services are taking place at

Community Health Centers with medication provided according to the capabilities of each region. In accordance with the hospital doctor's recommendations, patients with diabetes mellitus who are given a referral to a hospital that is a health insurance participant can be given oral medication or injections. Every Indonesian citizen (WNI) is advised to become a participant so that the illness they suffer can be covered by BPJS Health. Benefits that can be covered by BPJS Health include First Level Health Facilities (FKTP) and Advanced Referral Health Facilities (FKRTL). T2DM is one of the ten most frequent diagnoses received in visits from 2019 to 2020.

An actuary can obtain important information regarding risk modeling and health insurance claims management through grouping analysis of diabetes mellitus data by province. Actuaries can identify provinces with higher risk and develop more effective strategies for setting premium rates or managing health insurance portfolios by analyzing patterns and variations in diabetes mellitus prevalence rates between provinces. To find out the grouping in the provincial distribution of participants diagnosed with diabetes mellitus entering the FKRTL visit, the Hierarchical Agglomerative Clustering (HAC) method can be used. The purpose of clustering is to group data that is the same for each pair of data. HAC can be used to combine data by arranging it into hierarchies. If the data is similar it will be placed in an adjacent hierarchy and vice versa [2]. The results of the clustering were then modeled to see the predictions for patients diagnosed with diabetes mellitus entering the FKRTL visit. Modeling is carried out using the Autoregressive Integrated Moving Average (ARIMA) method.

Based on this background description, the author is interested in conducting research related to the application of hierarchical agglomerative clustering to group provinces in Indonesia based on the number of FKRTL admission services for patients diagnosed with T2DM, the results of which are then used to create a prediction model. It is hoped that this research will provide an overview of the distribution of participants diagnosed with T2DM 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.

II. METHODS

Cluster Analysis

Cluster analysis is a statistical analysis technique used to place a collection of objects into two groups, even more so, based on the characteristics of the objects in common [3]. The cluster formed is said to be good if it has large homogeneity between objects in the same cluster and has large heterogeneity between one cluster and another [4].

Euclidean *Distance*

Distance determination can use the Euclidean distance method. The equation of Euclidean distance is as follows [5]:

$$d_i = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

with:

d_i : ith Euclidean distance

x_i : Coordinates of point x in-i dimension

y_i : Coordinates of point y in-i dimension

After obtaining the distance from the Euclidean distance calculation, clustering is carried out using the complete linkage, single linkage and average linkage distance parameter approaches.

Agglomerative Hierarcical Clustering

The Hierarchical Agglomerative Clustering method is a method with the concept of merging two small clusters with the closest distance into one larger cluster [6]. Using this method, clusters will be formed with a bottom-up approach starting from grouping singleton data which is then combined repeatedly.

1. *Complete Linkage* is a merging grouping based on the maximum distance or furthest distance between objects in a cluster [7]. Minimum value in D = d_{ij} searched first and combined with corresponding objects such as i and k to obtain

cluster (ij). The distance between other clusters k and (ij) is calculated using the method [3]:

$d_{(ij)k}=max\{d_{ik}d_{jk}\}$.....(2)

with:

$d_{(ij)k}$: distance between cluster i and j to k

d_{jk} : distance between cluster j and k

d_{ik} : distance between cluster i and k

2. *Single linkage* is a merging grouping based on the minimum distance or furthest distance between objects in a cluster. The distance between clusters is determined based on the closest distance [8]. The distance between other clusters k and (ij) is calculated using the following method:

$d_{(ij)k}=min\{d_{ik}d_{jk}\}$.....(3)

with:

$d_{(ij)k}$: distance between cluster i and j to k

d_{jk} : distance between cluster j and k

d_{ik} : distance between cluster i and k

3. *Average linkage* is a grouping of mergers based on average distance. Average linkage aims to minimize the average distance of all pairs of observations from two groups combined. The distance between other clusters k and (ij) is calculated using the method [8]:

$d_{(ij)k}= \overline{d_{ik}d_{jk}}$.....(4)

Keterangan:

$d_{(ij)k}$: distance between cluster i and j to k

d_{jk} : distance between cluster j and k

d_{ik} : distance between cluster i and k

Cophenetic correlation

Determining the best algorithm can be done using cophenetic correlation. Cophenetic correlation is the relationship between the elements produced by the dendogram and the original elements of the dissimilarity matrix. The equation of cophenetic correlation is as follows.

$$rcoph = \frac{\sum_{i<j}^{n}(d_{ij}-\bar{d})(d_{c_{ij}}-\bar{d}c)}{\sqrt{[\sum_{i<j}^{n}(d_{ij}-\bar{d})^2][\sum_{i<j}^{n}(d_{c_{ij}}-\bar{d}c)^2]}}$$.....(5)

Keterangan:

rcoph : coefficient of cophenetic correlation

d_{ij} *: euclidean distance* between the ith individual and the jth individual

$\bar{d}$: mean d_{ij}

$d_{c_{ij}}$: *cophenetic distance* between the ith individual and the jth individual

$\bar{d}c$: mean $d_{c_{ij}}$

Silhouette coefficient

Validity testing can be done by looking at the silhouette coefficient so that the best cluster can be determined. Calculating the silhouette coefficient requires components a_i and components b_i. The value of a_i can be found with the following equation [9]:

$$a_i = \frac{1}{m_j-1}\sum_{\substack{r=1\\r\neq 1}}^{m_j} d(x_i^j, x_r^j)$$.....(6)

with:

j : *cluster*

i : data index (i : 1,2,…, m_j)

a_i : the average of the i-th distance to all data in one cluster

m_j : amount of data in *cluster-j*

$d(x_i^j, x_r^j)$: distance of the ith data to the rth data in one cluster j

Meanwhile, the b_i value can be found using the following equation [9]:

$$b_i = min_{\substack{n=1,\dots k\\n\neq j}}\left\{\frac{1}{m_n}\sum_{\substack{r=1\\r\neq 1}}^{m_n} d\left(x_i^j, x_r^n\right)\right\}$$.....(7)

with:

j : *cluster*

i : data index (i : 1,2,…, m_j)

b_i : the average of the i-th distance to all data in one cluster

m_n : amount of data in *cluster-n*

$d(x_i^j, x_r^j)$: distance of the ith data to the rth data in one cluster j

So we get the formula for finding silhouettes with the following equation [9]:

$$\boldsymbol{Si} = \frac{b_i - a_i}{\mathbf{max}\{b_i, a_i\}}$$.....(8)

With:

$\boldsymbol{Si}$: *silhouette index* i-th data in one cluster

The criteria for cluster accuracy and quality based on the silhouette coefficient value can be stated as in the following table [9]:

Table 2.1 Criteria *coefficient silhouette*

Nilai *Coefficient Silhouette*	Criteria
0,71 – 1,00	Strong
0,51 – 0,70	Good
0,26 – 0,50	Weak
≤ 0,25	Poor

Autoregressive Integrated Moving Average (ARIMA)

Data and Data Sources

The data used in this final project research is secondary data obtained from the Social Security Administering Body (BPJS) Health Sample Data from 2015 to 2020. This data is in the form of contextual sample data on diabetes mellitus at the Advanced Health Referral Facility (FKRTL) service. The data used is monthly period data for five years starting from January 2015 to December 2020.

Analysis Stages

1. Conduct data exploration from time series data.
2. Grouping using the HAC method with the following process.

a. Standardize values on time series data using Z-Score for each time series.
b. Form a distance matrix with Euclidean distance.
c. Grouping time series data using the HAC method (*single linkage, complete linkage,* dan *average linkage*.
d. See the most optimal size by calculating cophenetic correlation.
e. Determining the optimal number of groups with silhouette coefficients.

3. Create a group prototype based on the average value of each province each month.
4. Determine the ARIMA model on group level time series data with a prototype representation, namely the average data in each group.
5. Evaluate the forecast using the Mean Absolute Percentage Error value.

III. RESULTS AND DISCUSSION

Preprocesing Data

In this research, clustering was carried out and then continued with data projections on the number of FKRTL visits for BPJS Health DMT2 participants. The data used in this research is BPJS Health FKRTL data. There are 5 BPJS Health FKRTL data variables used in this research.

Individual data recorded during FKRTL visits is taken once every month to obtain the accumulative number of participants who visit. Next, the accumulative data is multiplied by weighted data on visits by DMT2 FKRTL BPJS Health participants for each individual. Weighting data has different values for each individual.

Descriptive Statistics

Data exploration aims to determine the description and characteristics of research data in the form of time series data. Each province has different prevalence rates for type 2 diabetes mellitus. Identification of patterns formed in the data can be seen through time series plots. The following is a time series plot of FKRTL service visit data for DMT2 patients.

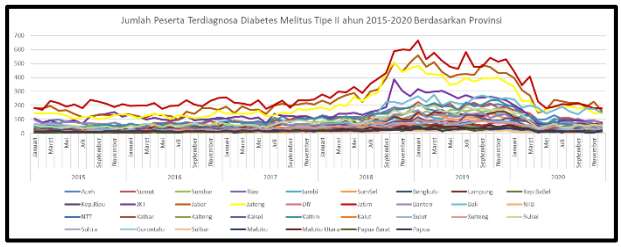

Figure 3. 1 *Grafik Kunjungan FKRTL Terdiagnosis DMT2*

Based on Figure 3.1, it can be seen that the condition of FKRTL service visits for T2DM patients in several provinces in Indonesia shows a fluctuating pattern. The provinces with the highest visits include East Java, West Java, Central Java, Bali and Jakarta.

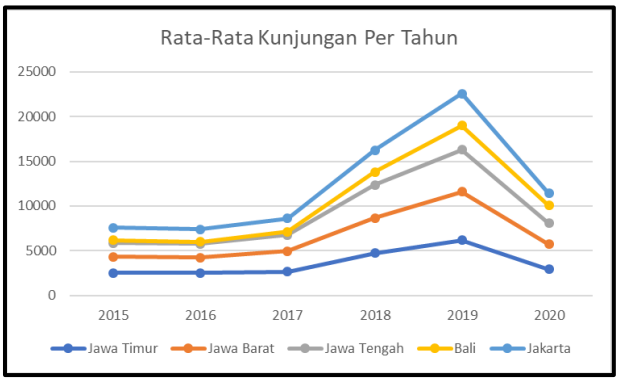

Figure 3. 2 Plot Rata-Rata Data 5 Provinsi Tertinggi Kunjungan Per Tahun

From Figure 3.2, it can be seen that the data shows an increase from 2015 to 2018. Furthermore, the data shows a quite drastic increase in 2019, but in 2020 the number of FKRTL service visits has decreased again.

Analisis *Hierarchical Agglomerative Clustering*

Grouping using the HAC method begins by calculating a distance matrix by calculating a similarity measure using Euclidean distance. Before proceeding to the next stage, time series data needs to be standardized using Z-score rules because the values in the data have different scales. The data that has been standardized is then searched for the Euclidian distance and formed in a distance matrix. The HAC algorithm is then carried out which consists of single linkage, average linkage, and complete linkage.

1. Metode *Single Linkage*

The grouping process uses the single linkage method, namely grouping is carried out based on the minimum distance or closest distance and is carried out repeatedly until all objects become one cluster. The cluster results of the single linkage method can be seen based on the following dendogram.

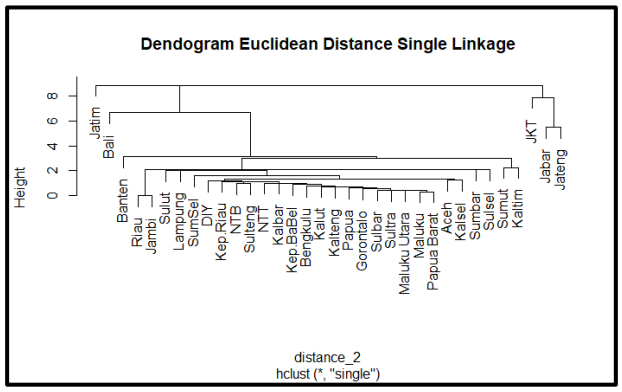

Figure 3. 3 Dendogram *Single Linkage*

2. Metode *Complete Linkage*

The grouping process uses the complete linkage method, namely grouping is carried out based on the maximum distance or furthest distance and is carried out repeatedly until all objects become one cluster. The cluster results of the complete linkage method can be seen based on the following dendogram.

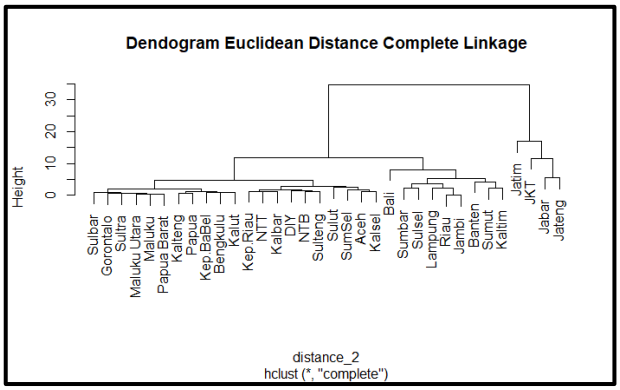

Figure 3. 4 Dendogram *Complete Linkage*

3. Metode *Average Linkage*

The grouping process uses the average linkage method, namely grouping is done based on average distance. Average linkage aims to minimize the average distance of all pairs of observations from two groups combined. The cluster results of the average linkage method can be seen based on the following dendogram..

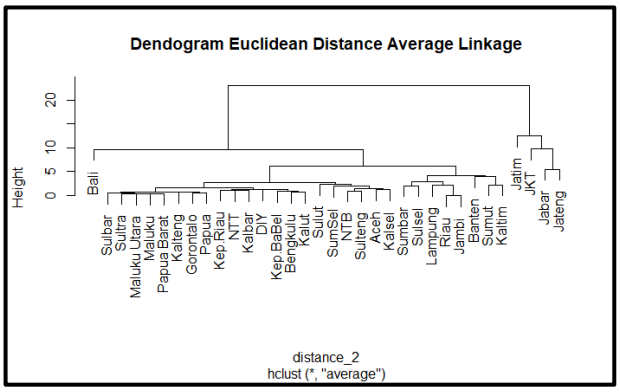

Figure 3. 5 Dendogram *Average Linkage*

Calculating the cophenetic correlation of each Euclidean distance on single linkage, complete linkage, and average linkage aims to see the most optimal group size. The most optimal group size will then be made into a dendogram. The following are the cophenetic correlation coefficient values for each HAC algorithm with the Euclidean similarity measure:

Table 3. 1 Cophenetic correlation coefficient based on Euclidean distance with the HAC algorithm

Algorithm	*Cophenetic Correlation*
Single Linkage	0.9024302
Complete Linkage	0.91824274
Average Linkage	0.92652957

Table 4.1 shows a comparison of cophenetic correlation based on Euclidean distance with the HAC algorithm. The higher the cophenetic correlation coefficient, the better the algorithm is at grouping objects. Based on Table 4.1, it can be concluded that the average linkage algorithm is the best algorithm for grouping provinces in Indonesia based on visit data from DMT2 FKRTL service participants..

Determination of Group Number

After obtaining the most optimal grouping algorithm, namely average linkage, then determine the optimal number of groups in grouping provinces in Indonesia. Calculation and determination of the optimal number of groups in grouping provinces in Indonesia based on visit data from DM2 FKRTL service participants using the silhouette coefficient. The following are the silhouette coefficient values.

Table 3. 2 Silhouette Coefficient Calculation Results

Number of Clusters	Coefficient *Silhouette*
2	0.7764
3	0.5524
4	0.3586
5	0.3307

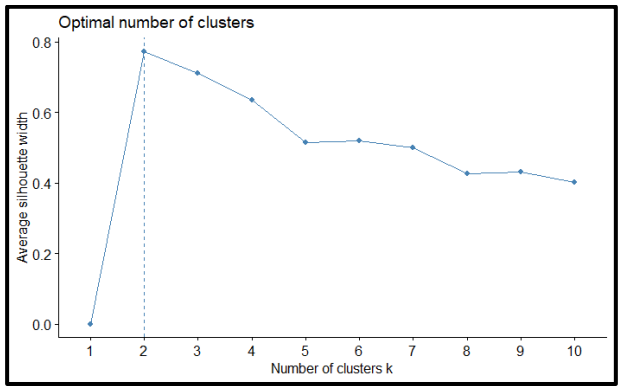

Figure 3. 6 Coefficient *Silhouette*

Determining the number of clusters is said to be good if the silhouette coefficient value is 1 [4]. Cluster results are said to be good if the silhouette coefficient value is closer to 1. Based on Table 4.2, the silhouette coefficient is highest in the number of clusters of 2 with a coefficient value of 0.7764. Figure 4.5 shows that the number of clusters that are good for grouping provinces in Indonesia based on data on the number of monthly visits by DM2 patients for FKRTL services is 2.

Group Profiling

After determining the best number of clusters using the silhouette coefficient, namely 2 clusters, we then look at the dendogram results of the average linkage algorithm, so we get the dendogram in Figure 4.6.

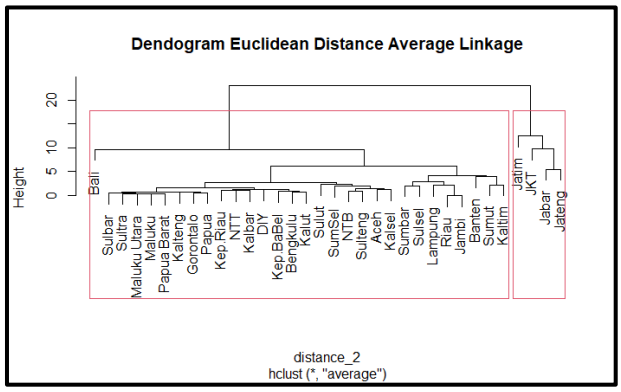

Figure 3. 7 Dendogram *Average Linkage* Provinsi in Indonesia

Based on Figure 3.7, the grouping results for cluster 1 consist of the provinces of East Java, Central Java, West Java and Jakarta. Meanwhile, cluster 2 consists of the provinces of Bali, West Sulawesi, North Sulawesi, North Maluku, Maluku, West Papua, Central Kalimantan, Gorontalo, Papua, Riau Islands, West Kalimantan, Bangka Belitung Islands, Bengkulu, North Kalimantan, North Sulawesi, West Nusa Tenggara , South Sumatra, Central Sulawesi, Aceh, South Kalimantan, West Sumatra, South Sulawesi, Lampung, Riau, Jambi, Banten, North Sumatra and East Kalimantan.

Group Level Forecasting

Group level forecasting is carried out by creating provincial prototypes. Prototype formation is based on the average value each month in each province. The group prototype plot for each cluster can be seen as follows.

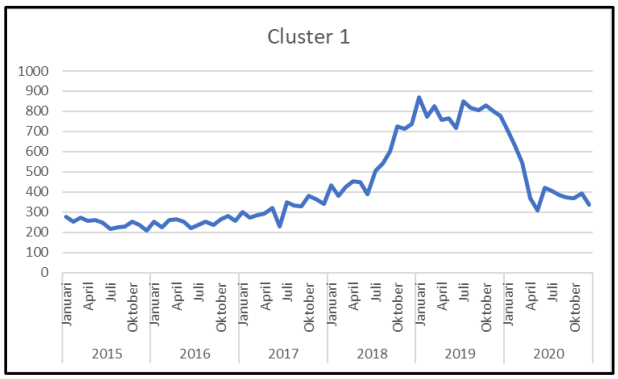

Figure 3. 8 *Prototipe Data* Province *in Cluster* 1

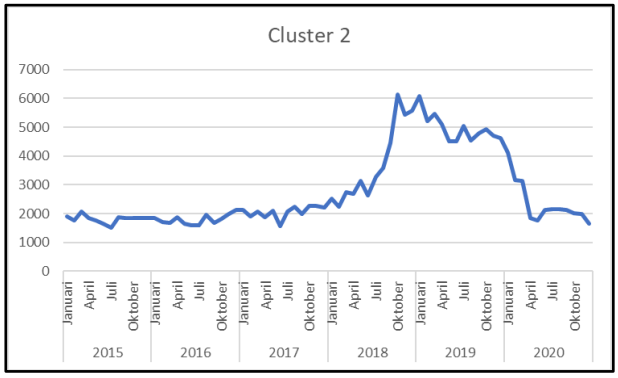

Figure 3. 9 *Prototipe Data* Province in *Cluster* 2

Based on the time series data plot in the figure, it can be concluded that the non-stationary prototype data tends to increase in late 2018 and then decrease drastically in 2020. The group level forecasting process first tests the stationarity of the data.

Data Stationarity Test

Stationarity testing can be carried out on the mean and stationarity testing on variance. Stationarity testing of the average can be done using the Augmented Dickey-Fuller test (ADF Test). Stationarity testing of variance can be done by carrying out the Boxcox Lambda test.

1. Stasionerity for Mean

The stationarity test for the mean uses the Augmented Dickey-Fuller (ADF test) with a significance level of 5%. If the test results are not stationary, differencing is necessary. Following are the results of the stationarity test against the mean.

Table 3.3 ADF *Test Prototipe* Data

Cluster	*P-Value*	Decision
Cluster 1	0.76351	Not Stationary
Cluster 2	0.89919	Not Stationary

As can be seen from Table 3.4, cluster 1 and cluster 2 have a p-value that is greater than the significance level of 0.05. It can be concluded that the time series data is still not stationary so a differencing process is needed.

Table 3.4 ADF First Differentiation Test Prototype Data

Cluster	*P-Value*	Decision
Cluster 1	0.07539	Not Stationary
Cluster 2	0.04177	Stasioner

As can be seen from Table 3.5, cluster 1 has a value greater than 0.05 so it can be concluded that the data is not stationary so a second level of differencing is necessary. Cluster 2 has a p-value that is smaller than the significance level of 0.05. Therefore, H1 is accepted so it can be concluded that the data is stationary with respect to the stationary mean in the first level differencing process for cluster 2. The results of the second level differencing in cluster 1 are as follows.

Table 3.5 ADF Second Differentiation Test Results Prototype Data

Cluster	*P-Value*	Decision
Cluster 1	0.01	Stasioner

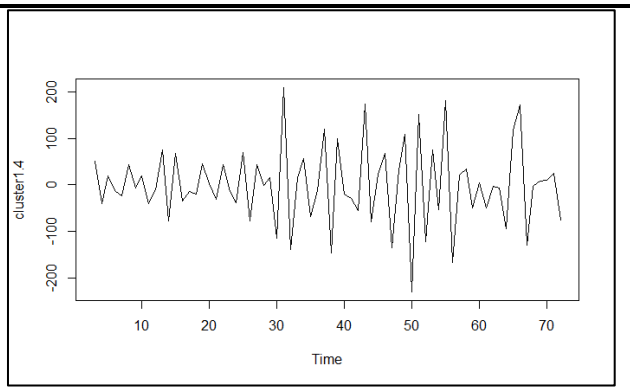

Figure 3. 10 Plot of Differentiation Data for Both Clusters 1

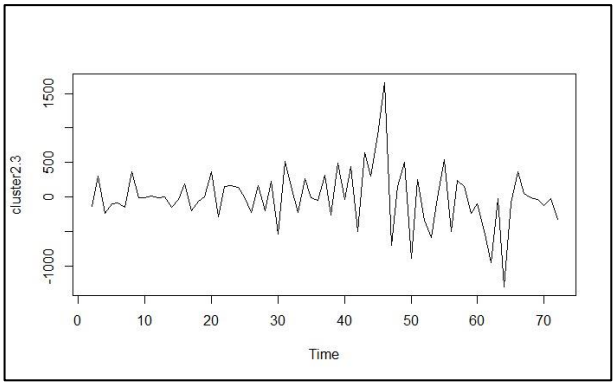

Figure 3. 11 Cluster 2 First Differentiation Data Plot

After carrying out second level differencing on cluster 1 data, a p-value of 0.01 was obtained. Therefore, H1 is accepted so that it can be concluded that the data is stationary with respect to the stationary mean in the first level differencing process for cluster 2. Checking the stationarity of the mean can be done by looking at the differentiated data plot. Based on Figure 4.9 and Figure 3.11, it can be seen that the average of the data moves constantly over time so it can be said that the data is stationary.

2. Stasionerity for the Variants

The stationarity test for variance can be carried out using the Boxcox Lambda test. Boxcox test data processing was carried out with the help of Rstudio. Based on the test results listed in Appendix B, the largest Pearson Product-Moment Correlation Coefficient (PPCC) value was obtained at lambda 0.5. This shows that the √Zt transformation is needed for cluster 1.

Table 3. 6 Boxcox Lambda Test Results Prototype Data

Cluster	PPCC Value	Lambda	Decision
Cluster 1	0.9916	1.0	Stasioner
Cluster 2	0.9620	0.5	Not Stasioner

Stationary checking of variance is carried out after the data has been transformed. Based on the test results, the largest PPCC value was obtained at lambda 0.5 with a value of 0.9935027. The following are the results of checking the stationarity of the variance after transformation:

Table 3.7 Boxcox Lambda Test Results Prototype Data

Cluster	PPCC Value	Lambda	Decision
Cluster 2	0.9620	1.0	Stasioner

Model Identification

After carrying out the stationarity test, a lag check can be carried out based on the ACF and PACF plots to determine the ARIMA model that allows.

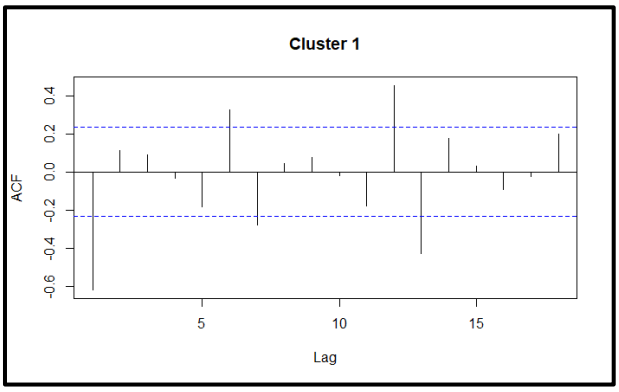

Figure 3. 12 Plot ACF *Cluster* 1

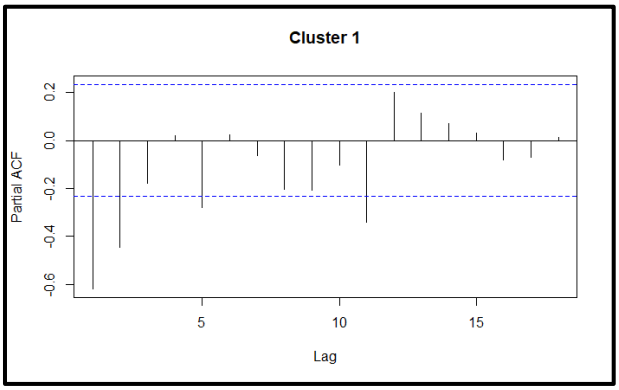

Figure 3. 13 Plot PACF *Cluster* 1

Based on the ACF plot image, cut-off occurs at the 1st and 2nd lags. In the PACF plot, cut-off occurs at the 1st, 2nd and 3rd lags, so that the possible ARIMA model is formed as follows.

Table 3. 8 Cluster 1 ARIMA model

No	Model
1	ARIMA(1,2,0)
2	ARIMA(1,2,1)
3	ARIMA(1,2,2)
4	ARIMA(2,2,0)
5	ARIMA(2,2,1)
6	ARIMA(2,2,2)
7	ARIMA(3,2,0)
8	ARIMA(3,2,1)

Meanwhile for cluster 2 the results of the ACF and PACF plots are as follows.

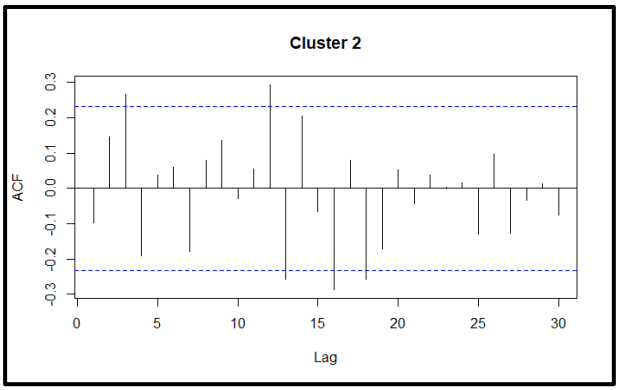

Figure 3. 14 Plot ACF *Cluster* 2

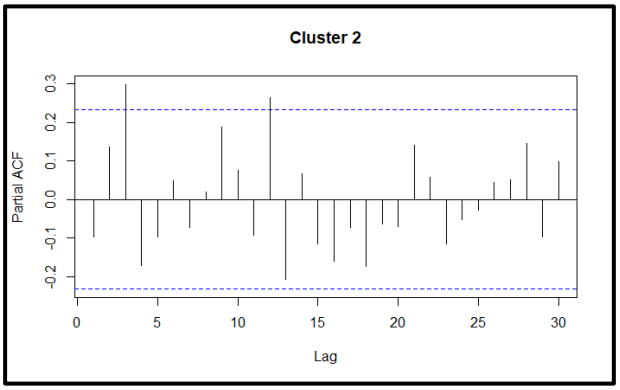

Figure 3. 15 Plot PACF Cluster 2

Based on the ACF plot, cut-off occurs at the 1st and 3rd lags. In the PACF plot, a cut-off occurs at the 1st and 3rd lags so that the possible ARIMA model is formed as follows.

Table 3. 9 Cluster 2 ARIMA model

No	Model
1	ARIMA(3,1,3)
2	ARIMA(3,1,0)
3	ARIMA(0,1,3)
4	ARIMA(3,1,1)
5	ARIMA(1,1,3)

Parameter Estimation

The ARIMA model parameter significance test requires information regarding the coefficients of the parameters obtained as well as the standard errors of the estimated parameters. The hypothesis stated in this test is as follows:

H0 : parameter is not significant in the model

H1: significant parameter in the model

with a significance level α of 5%. H_0 is rejected if the p-value is less than the α value. A summary of the test results can be seen in the following table. Based on the test results attached in Appendix 5 using the help of Rstudio, a summary of the parameter significance test results for cluster 1 and cluster 2 is obtained as follows..

Table 3. 10 Cluster 1 ARIMA Model Significance Test Results

No	Model	Keterangan
1	ARIMA(1,2,0)	Signifikan
2	ARIMA(1,2,1)	Signifikan
3	ARIMA(1,2,2)	Signifikan
4	ARIMA(2,2,0)	Signifikan
5	ARIMA(2,2,1)	Signifikan
6	ARIMA(2,2,2)	Signifikan
7	ARIMA(3,2,0)	Signifikan
8	ARIMA(3,2,1)	Signifikan

Table 3. 11 Cluster 1 ARIMA Model Significance Test Results

No	Model	Keterangan
1	ARIMA(3,1,3)	Tidak Signifikan
2	ARIMA(3,1,0)	Tidak Signifikan
3	ARIMA(0,1,3)	Signifikan
4	ARIMA(3,1,1)	Tidak Signifikan
5	ARIMA(1,1,3)	Tidak Signifikan

In cluster 1 the significant models are the ARIMA(1,2,0), ARIMA(1,2,1), ARIMA(1,2,2), ARIMA(2,2,0), ARIMA(2,2, 1), ARIMA(2,2,2), ARIMA(3,2,0), and ARIMA(3,2,1). In cluster 2 the significant model is the ARIMA(0,1,3) model. Models that have passed and are said to be significant are then processed for parameter diagnosis.

Diagnose Parameter

Normality *Residual*

Normality testing can use the Shapiro-Wilk test. The following are the results of normality testing:

Table 3. 12 Summary of Shapiro-Wilk Test Results

No.	*Cluster* 1		
	Model ARIMA	***P-Value***	**Information**
1	ARIMA(1,2,0)	0.14558	Normally distributed
2	ARIMA(1,2,1)	0.28614	Normally distributed
3	ARIMA(1,2,2)	0.015139	Tidak Normally distributed
4	ARIMA(2,2,0)	0.19289	Normally distributed
5	ARIMA(2,2,1)	0.011137	Tidak Normally distributed
6	ARIMA(2,2,2)	0.000169	Tidak Normally distributed
7	ARIMA(3,2,0)	0.83654	Normally distributed
8	ARIMA(3,2,1)	0.059172	Normally distributed
No.	***Cluster* 2**		
1	ARIMA(0,1,3)	0.1979	Normally distributed

Based on Table 3.13, the ARIMA(3,2,1) and ARIMA(0,1,3) models have a normal distribution so they can be continued for the next test, namely the Ljung-Box Test.

Ljung-Box Test

The L-Jung Box test can be carried out to determine whether the residual model is white noise [12]. The hypothesis stated in this test is as follows:

H0: the residue satisfies the white noise process

H1: the residue does not fulfill the white noise process.

with a significance level α of 5%. H0 is rejected if the p-value is less than the α value. A summary of test results can be seen in the following table.

Table 3. 13 Summary of Ljung-Box Test Results

No.	*Cluster* 1		
	Model ARIMA	*P-Value*	Information
1	ARIMA(3,2,1)	0.96567	*White Noise*
No.	*Cluster* 2		
1	ARIMA(0,1,3)	0.99548	*White Noise*

Mean Absolute Percentage Error **(MAPE)**

From the models obtained, the best model was selected by looking at the smallest MAPE value. With Rstudio software you can get it:

Table 3. 14 MAPE Value for ARIMA Model

Cluster	Model	MAPE(%)
1	ARIMA(3,2,1)	5.04
2	ARIMA(0,1,3)	2.91

The MAPE value for the ARIMA(3,2,1) model is 5.04% in cluster 1. The MAPE value for the ARIMA(0,1,3) model is 2.91% in cluster 2. The ARIMA(3,2,1) model equation is as follows following.

$$Y_t = 2Y_{t-1} - Y_{t-2} + (-1.48850)Y_{t-1} + (0.96676)e_{t-1} + (-2.42276)e_{t-2} + (0.47653)e_{t-3} + e_t$$

The ARIMA(0,1,3) model equation is as follows.

$$Y_t = Y_{t-1} + (-1.14471)e_{t-1} + (0.50372)e_{t-2} + (-0.30147)e_{t-3} + e_t$$

Prediction of FKRTL Service Visits for DMT2 Participants

Based on the MAPE values obtained previously, the ARIMA(3,2,1) model for cluster 1 and the ARIMA(3,1,0) model for cluster 2 are adequate for predictions. Predictions for FKRTL service visits for DMT2 participants are as follows:

Table 3. 15 Cluster 1 and Cluster 2 Prediction Results

Periode	*Cluster* 1	*Cluster* 2
1	337	1641
2	318	1590
3	298	1495
4	285	1518
5	266	1478
6	250	1463
7	233	1471
8	217	1452
9	201	1455
10	184	1454

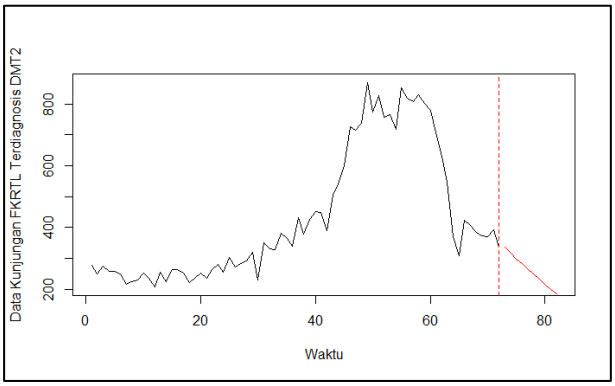

Figure 3. 16 Cluster 1 Prediction Time Series Plot

Based on Figure 3.16, the plot shows that the predicted data has a value that continues to decline from the previous period, namely in 2020.

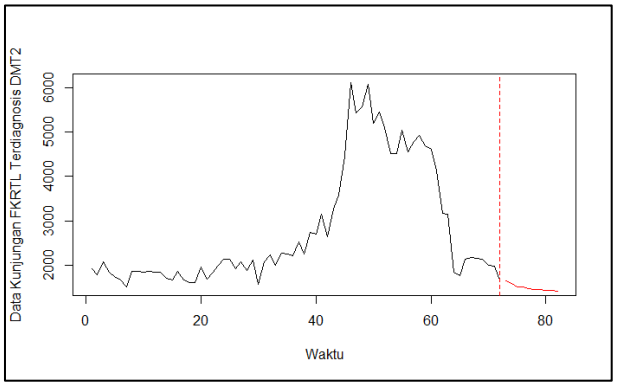

Figure 3. 17 Cluster 2 Prediction Time Series Plot

Based on Table 3.16, it can be seen that the prediction results that have been obtained can be said to tend to decrease every year. This could be caused by the data in 2020 which decreased drastically. Based on Figure 3.17, the plot shows that the predicted data has a value that tends to decrease but does not decrease much in each period.

IV.KESIMPULAN

Based on the results of research and discussion of the Hierarchical Agglomerative Clustering Time Series method for predicting the number of FKRTL BPJS Health services for T2DM sufferers in Indonesia in 2015-2020, the conclusion of this research is:

1. 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.
2. The prediction results show that the predicted data has a value that continues to decrease from the previous period, namely in 2020 for the cluster. The prediction results in cluster 2 have values that tend to decrease but do not decrease much in each period.

The clustering results that have been obtained are only limited to the best number of 2 clusters, so it is recommended to use other clustering methods to see the comparison of the number of best clusters formed. Using more and latest data, especially above 2020, to see whether there is interference such as increasing and decreasing data that is too significant so that other methods can be used and the resulting model is more accurate. Suggestions from researchers for the government and BPJS Health to immediately create and decide on a new, more efficient policy for handling T2DM participants in FKRTL services in the future.

V. REFERENCES

[1]. D. S. Prawitasari, "Diabetes Melitus dan Antioksidan," Jurnal Kesehatan dan Kedokteran, vol. 1, pp. 48-52, 2019.

[2]. R. P. Justitia, N. Hidayat and E. Santoso, "Implementasi Metode Agglomerative Hierarchical Clustering Pada Segmentasi Pelanggan Barbershop (Studi Kasus : RichDjoe Barbershop Malang)," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. V, pp. 1048-1054, 2021.

[3]. N. Ulinnuh and R. Veriani, "Analisis Cluster dalam Pengelompokan Provinsi di Indonesia Berdasarkan Variabel Penyakit Menular Menggunakan Metode Complete Linkage, Average Linkage dan Ward," Jurnal Nasional Informatika dan Teknologi Jaringan, vol. 5, 2020.

[4]. A. T. Dani, S. Wahyuningsih and N. A. Rizki, "Penerapan Hierarchical Clustering Metode Agglomerative pada Data Runtun Waktu," Jambura Journal of Mathematics, vol. I, pp. 64-78, 2019.

[5]. R. A. Rahman, F. M. Afendi, W. Nugraheni, K. Sadik and A. Rizki, "Pengelompokan dan Peramalan Deret Waktu pada Produksi Bawang Merah Tingkat Provinsi di Indonesia," Seminar Nasional Official Statistics , pp. 457-464, 2021.

[6]. A. Septianingsih, "Pemetaan Kabupaten Kota di Provinsi Jawa Timur Berdasarkan Tingkat Kasus Penyakit Menggunakan Pendekatan Tingkat Kasus Penyakit Menggunakan Pendekatan Agglomeratif Hierarchical Clustering," Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika, vol. III, pp. 367-386, 2022.

[7]. Mustika, Y. Ardilla, A. Manuhutu, N. Ahmad, I. Hasbi, Guntoro, M. A. Manuhutu, M. Ridwan, Hozairi, A. K. Wardhani, S. Alim, I. Romli, Y. Religia, D. T. Octafian, U. U. Sufandi and I. Ernawati, Data Mining dan Aplikasinya, Bandung: Widina Bhakti Persada , 2021.

[8]. Iis, I. Yahya , G. N. A. Wibawa, Baharuddin, Ruslan and L. Laome, "Penggunaan Korelai Cophenetic Untuk Pemilihan Metode Cluster Berhierarki Pada Pengelompokkan Kabupaten/Kota Berdasarkan Jenis Penyakit Di Provinsi Sulawesi Tenggara," SINTA, vol. VI, 2022.

[9]. D. A. I. C. Dewi and D. A. K. Pramita, "Analisis Perbandingan Metode Elbow dan Sillhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Porduksi Kerajinan Bali," MATRIX, vol. IX, pp. 102-109, 2019.

[10]. M. A. Zen, S. Wahyuningsih and A. T. R. Dani, "Aplikasi Pendekatan Agglomerative Hierarchical Time Series Clustering untuk Peramalan Data Harga Minyak Goreng di Indonesia," Seminar Nasional Statistics, pp. 293-302, 2022.

[11]. D. Andiani, S. D. r. Septiani and A. Riana, "Analisis Teknik non-Hierarki untuk Pengelompokkn Kabupatn/Kota di Provinsi Jawa Barat Berdasarkan Indikator Kesejahteraan Rakyat 2020," Jurnal Riset Matematika dan Sains Terapan, pp. 21-28, 2022.

Cite this article as :

Hana Sabrina Sulthoni, Ayu Sofia, "Prediction of FKRTL Services Diagnosed with Type 2 Diabetes Mellitus Using the Hierarchical Agglomerative Clustering Time Series Method", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 10 Issue 6, pp. 144-156, November-December 2023. Available at doi : https://doi.org/10.32628/IJSRSET231064
Journal URL : https://ijsrset.com/IJSRSET231064