Jurnal Pijar MIPA
https://jurnalfkip.unram.ac.id/index.php/JPM

ISSN 1907-1744 (Print)
ISSN 2460-1500 (Online)

Model Based Clustering for Regency/City Grouping Based on Community Welfare Indicators in North Sumatra

Muhammad Afif Fauzi Hasibuan, Hendra Cipta, Sajaratud Dur

Department of Mathematics, Faculty of Science and Technology, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia
*E-mail: mafauzihsb@gmail.com

Received: February 27, 2024. Accepted: April 16, 2024. Published: May 19, 2024

Abstract: This thesis aims to apply model-based clustering in grouping regencies/cities in North Sumatra based on Community Welfare Indicators to determine the number of groups (clusters) formed based on community welfare indicators in regencies/cities in North Sumatra and to understand the level of community welfare from this grouping in planning and managing community welfare in regencies/cities in North Sumatra, with the hope of achieving equal welfare in every region. The research method used is Model-Based Clustering, which uses 5 research data variables: HDI, Poor Population, Unemployment Rate, GDP, and Health. In this millennial era, the assessment of community welfare requires more attention. Rapid social, technological, and environmental changes have created new dynamics that can affect community welfare. The evaluation of community welfare is not only limited to economic parameters but also considers health, unemployment, and other factors. By using Model-Based Clustering, it is possible to determine the optimal number of groups (clusters) from various and possibly correlated variables, and the results are easier to understand, making the analysis and understanding of the results easier. Readers can learn about the level of community welfare, and the community and government can evaluate their welfare for future improvements. The research results show that among the groups of regencies/cities formed, five cities consistently show lower Gross Regional Domestic Product (GDP) and Human Development Index (HDI) than other cities. Therefore, a sustainable approach is needed to improve these cities' economic conditions and social welfare.

Keywords: Categorization; Model-Based Clustering; Societal Well-being.

Introduction

The desire of every individual in Indonesia or the world is to achieve a prosperous life, both in urban and rural areas, as well as in physical and spiritual aspects. Prosperity includes social, material, and spiritual elements accompanied by security, moral values, and inner peace [1].

Essentially, achieving community welfare is the main goal of every economic development effort. In Indonesia, welfare is also one of the state's goals, as expressed in the preamble of the 1945 Constitution, which aims to "protect all Indonesian people and all descendants of Indonesia and to improve the general welfare, improve the nation's knowledge of life."

Social change in society will occur due to development projects in a region. New potential to improve the economic welfare of residents will develop along with the positive and negative impacts of these initiatives. This indicates that development not only affects significant economic growth but also results in changes in the social and cultural lives of the community. These changes involve aspects of lifestyle and the emergence of various social issues. As a step towards improving community welfare, the development process must proceed consistently, involving the community as actively involved stakeholders. In line with the spirit of regional autonomy that emphasizes community participation, this demands seriousness from the government through thorough planning involving all interested parties [2].

Model-based clustering is a clustering method that utilizes probability models to group data. One common distribution used in MBC is the normal distribution. However, it is known that not all data may fit the normal distribution, especially when there are outliers. Therefore, in 2012, Andrews and McNicholas developed a more robust model to handle data containing outliers by adopting the t-distribution.

This model was initially used to cluster objects in a population. The basic assumption in Model-Based Clustering (MBC) is that in a population, subpopulations with a certain probability distribution can be identified, and each subpopulation has unique parameters. All subpopulations have a Mixture Distribution with different proportions for each subpopulation. This assumption leads us to the mathematical probability model of Model-Based Clustering. Finite Mixture models in clustering have rapidly developed and become one of the popular clustering methods [3].

Research Methods

Descriptive Statistics

Descriptive statistics is a statistical analysis commonly used to organize and present data. Typically, descriptive statistics are used as a preliminary step to organize data before conducting further analysis. Research results can be extrapolated if the null hypothesis (H_0) is accepted. One or more variables are used in descriptive

How to Cite:
Hasibuan, M. A. F., Cipta, H., & Dur, S. (2024). Model Based Clustering for Regency/City Grouping Based on Community Welfare Indicators in North Sumatra. *Jurnal Pijar Mipa*, *19*(3), 479–487. https://doi.org/10.29303/jpm.v19i3.6595

analysis, which is done separately. Therefore, there is no comparison or correlation between variables in this analysis. [4-5].

Cluster Analysis

Cluster analysis is a statistical analysis technique used to group objects into two or more clusters based on the similarity characteristics among the objects. Additionally, cluster analysis aims to maximize the similarity of objects within clusters while maximizing the differences between clusters [6-7].

There are two categories for cluster formation processes: hierarchical and non-hierarchical. Hierarchical methods are step-by-step methods. In this method, certain stages will be formed, such as in a tree structure, and can be generated as a dendrogram. Non-hierarchical methods are also called k-means methods. This method differs from hierarchical methods because non-hierarchical methods start by determining the desired number of clusters, and then the results of these observation objects are combined and form clusters [8-10].

Deteksi Outlier Multivariate

Multivariate outlier detection in research data is carried out to test the initial hypothesis that the data tends to contain significantly different values; thus, applying *finite mixture model-based clustering* with multivariate t distribution becomes more appropriate to obtain robust clustering results and can handle the presence of extreme values. [11].

One method used to assess the presence of multivariate outliers is by calculating the Mahalanobis Distance, which is defined as follows:

$$MD_i = \left[(x_i - \mu)^T S^{-1} (x_i - \mu) \right]^{\frac{1}{2}}, i = 1, 2, ..., n \qquad (2.1)$$

Where $\overline{x}$ is the sample mean vector, and **S** is the sample covariance matrix. An observation is considered an outlier if its Mahalanobis Distance value is greater than $\sqrt{\chi^2_{p;1-a/2}}$ where p is the degrees of freedom and α is the predetermined significance level [12].

Multivariate t Distribution

Multivariate t-distribution is an alternative distribution used when there are many outliers in the data, causing the data distribution to become flatter and not follow a normal multivariate distribution. This multivariate t-distribution is an extension of the univariate t-distribution.

If a random vector variable $x = \left[x_1 x_2 ... x_p \right]^T$ has a multivariate t-distribution with v degrees of freedom, then the mean vector $\mu = \left[\mu_1 \mu_2 ... \mu_p \right]^T$ and the covariance matrix Σ have probability density functions as follows:

$$f(x) = \frac{\Gamma\left(\frac{v+p}{2}\right)}{(\pi v)^{\frac{p}{2}} \Gamma\left(\frac{v}{2}\right) |\Sigma|^{\frac{1}{2}}} \left(1 + \frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{v}\right)^{-\left(\frac{v+p}{2}\right)}; v > 2 \qquad `(2.2)$$

This v is also known as the shape parameter because the variation in its values affects the shape of the distribution. The multivariate t-distribution is recognized for handling outliers better than the multivariate normal distribution. Therefore, the multivariate t-distribution is often used in model-based clustering. [13].

Model Based-Clustering

Although model-based clustering has advantages in characterizing groups with few parameters and meeting statistical assumptions, it also has drawbacks. One is long computation time, especially with many groups or datasets. Model-based clustering also faces challenges in estimating the correct number of groups [14].

Banfield and Raftery developed a framework for model-based clustering using the eigenvalue decomposition of the covariance matrix Σ.

$$\Sigma_g = \lambda_g D_g A_g D_G^T$$

Where:

λ_g is a scalar value that indicates the volume of the Ellipses.

D_g is an orthogonal matrix of eigenvectors that Represent the orientation of the principal. Components.

A_g is a diagonal matrix with elements proportional to the eigenvalues and indicates the contours of the Density function. [12].

Model Finite Mixture

Finite mixture models provide significant flexibility in modeling data with multiple modes, skewness, and non-standard distribution characteristics. However, this flexibility is balanced by an increase in the number of parameters as the number of components increases [15].

Assume a random vector variable x with dimension p comes from a *finite mixture* distribution with probability density function:

$$f(x \mid \vartheta) = \sum_{g=1}^{G} \pi_g f_g(x \mid \theta_g) \qquad (2.3)$$

where $\vartheta = (\pi_1, \pi_2, ..., \pi_G, \theta_1, \theta_2, ..., \theta_g)$ the vector parameter $f_g(x \mid \theta_g)$ is called the probability density function of x with group parameter θ_G, G is the number of groups, and π_g is the weight or mixing proportion of group g subject to the following constraints [13]:

$$0 \le \pi_j \le 1, g = 1, 2, ..., G \text{ dan } \sum_{g=1}^{G} \pi_g = 1$$

Integrated Completed Likelihood

The ICL (Integrated Completed Likelihood) criterion has proven to be a popular approach in model-based clustering, as it automatically selects the number of clusters in a mixture model. This approach effectively maximizes the likelihood of the complete data, including allocating observations to clusters in the model selection criteria [16].

The selection of the best model-based clustering can use the ICL criterion. The principle is to maximize the likelihood function of the complete data. Therefore, the formula for ICL can be expressed as follows:

$$ICL_g = \ln f(y_i) - \frac{p}{2}\ln(n) \qquad (2.4)$$

Where $f(y_i) = f(x_i, \widehat{z}_i)$ is the likelihood function of the complete data. p is the total number of parameters, and n is the number of observations [11].

Community Walfare

Social welfare is the condition where all basic needs are met, especially fundamental ones such as food, clothing, housing, education, and health care. Here are some indicators of welfare [17].

HDI - The United Nations Development Programme (UNDP) has been using the Human Development Index (HDI) since 1990 to assess a country's human development achievements and releases it in an annual report known as the Human Development Report (HDR) [18].

Poor Population - The Copenhagen Social Development Action Programme in 1995, a high-level meeting worldwide, is evidence of this. Poverty, unemployment, and social exclusion are some social issues that require immediate attention and are important to be the main agenda in every country [19].

Open Unemployment Rate - Dealing with unemployment is one of the most difficult problems. Despite experiencing a decrease, there are still many unemployed people in Indonesia. Human development is the key to shaping a country's ability to develop its capabilities to create job opportunities to reduce the unemployment rate [20].

Gross Regional Domestic Product (GRDP) - One important metric in this assessment is Gross Regional Domestic Product (GRDP), which illustrates the importance of understanding the economic conditions of a region in a specific period. GRDP provides a comprehensive overview of the economic contribution of an area during a certain period [21].

Health - The reasons behind the decline in human quality of life, individually and collectively, remain a matter of debate. Part of this problem is the difficulty of conducting research on humans that can identify causal relationships. It is important to acknowledge that this issue is highly complex and influenced by many factors [22].

Results and Discussion

Data Description

Data description is an effort to present data in a way that is easy to understand and can be interpreted well. In this study, the data used consists of five independent variables, namely: Human Development Index (HDI) (X_1), Poor Population (X_2), Open Unemployment Rate (X_3), Gross Regional Domestic Product (X_4), and Health (X_5).

The data used in this study is secondary and obtained from the Central Bureau of Statistics of North Sumatra from 2018 to 2022. The research area is as follows:

Table 1. List of Regencies/Cities in North Sumatra

No	Wilayah	No	Wilayah
1	Nias	18	Serdang Bedagai
2	Mandailing Natal	19	Batu Bara
3	Tapanuli Selatan	20	Padang Lawas Utara
4	Tapanuli Tengah	21	Padang Lawas
5	Tapanili Utara	22	Labuhan Batu Selatan
6	Toba Samosir	23	Labuhan Batu Utara
7	Labuhan Baut	24	Nias Utara
8	Asahan	25	Nias Barat
9	Simalungun	26	Sibolga
10	Dairi	27	Tanjung Balai
11	Karo	28	Pematang Siantar
12	Deli Serdang	29	Tebing Tinggi
13	Langkat	30	Medan
14	Nias Selatan	31	Binjai
15	Humbang Hasundutan	32	Padangsidimpuan
16	Pakpak Bharat	33	Gunungsitoli
17	Samosir		

Multivariate Outlier Detection

The Mahalanobis Distance for each observation can be calculated and will indicate the distance of an observation from the mean of all variables in a multidimensional space [23].

The outlier detection method is carried out by calculating Mahalanobis and robust distances. These results are then compared with the cut-off value from the distribution $\chi^2_{p;0,05}$, Because this study uses 5 variables, the degrees of freedom are k=5, and the significance level is 0.05. The cut-off values generated each year will vary. Points outside this boundary are considered outliers and marked with a special symbol.

To determine whether regencies/cities are detected as outliers or not, outlier tests will be conducted on the five variables: HDI, poor population, LFPR, GRDP, and health. The mean value (μ) is obtained as follows :

$$\mu_{1.2018} = (2.63+2.85+2.99+\ldots+2.96) / 33 = 3.03$$

$$\mu_{2.2018} = (16.37+9.58+9.16+ \ldots +18.44) /33 = 11.325$$

$$\mu_{3.2018} = (1.62+4.43+5.28+\ldots+5.92) /33 = 4.593$$

$$\mu_{4.2018} = (0.47+1.69+1.72+\ldots+ 0.67) /33 = 3.03$$

$$\mu_{5.2018} = (20.04+14.3+8.46+\ldots+14.49) /33 = 12.154$$

After obtaining the mean (μ), the next step is determining the mean difference value ($x - \mu$) for Nias City.

$$x-\mu = [2.63\text{-}3.03 \quad 16.37\text{-}11.325 \quad 1.62\text{-}4.593 \quad 0.47\text{-}3.03 \quad 20.04\text{-}12.154]$$
$$x-\mu = [\text{-}0.400 \quad 5.044 \quad \text{-}2.973 \quad \text{-}2.560 \quad 7.885]$$

Next, $(x-\mu)^T$ the process is as follows:

$$(x-\mu)^T = \begin{bmatrix} \text{-0.400} \\ 5.044 \\ \text{-2.973} \\ \text{-2.560} \\ 7.885 \end{bmatrix}$$

The covariance matrix values for the year 2018 were calculated using Excel, and the covariance value for Nias Regency is as follows:

$$S = \begin{bmatrix} 0.042 & -0.742 & 0.303 & 0.548 & -0.431 \\ -0.742 & 24.553 & -3.847 & -7.816 & 10.485 \\ 0.303 & -3.847 & 7.205 & 5.182 & -2.307 \\ 0.548 & -7.816 & 5.182 & 29.278 & -4.207 \\ -0.431 & 10.485 & -2.307 & -4.207 & 11.271 \end{bmatrix}$$

Next, to calculate the inverse (S^{-1}) of the covariance matrix, the following steps are taken:

$$S^{-1} = \begin{bmatrix} 97.832 & 2.733 & -2.208 & 0.569 & 1.079 \\ 2.733 & 0.208 & -0.047 & 0.115 & -0.050 \\ -2.208 & -0.047 & 0.219 & -0.033 & -0.013 \\ 0.569 & 0.115 & -0.033 & 0.185 & -0.023 \\ 1.079 & -0.050 & -0.013 & -0.023 & 0.169 \end{bmatrix}$$

Then, from these matrices, the Mahalanobis Distance can be calculated using the formula:

$$MD = \sqrt{(x-\mu)^T S^{-1} (x-\mu)}$$

So :

$$MD = \sqrt{(x-\mu)^T S^{-1} (x-\mu)}$$

$$MD = \begin{bmatrix} \text{-0.400} \\ 5.044 \\ \text{-2.973} \\ \text{-2.560} \\ 7.885 \end{bmatrix} x \begin{bmatrix} 97.832 & 2.733 & -2.208 & 0.569 & 1.079 \\ 2.733 & 0.208 & -0.047 & 0.115 & -0.050 \\ -2.208 & -0.047 & 0.219 & -0.033 & -0.013 \\ 0.569 & 0.115 & -0.033 & 0.185 & -0.023 \\ 1.079 & -0.050 & -0.013 & -0.023 & 0.169 \end{bmatrix} x \begin{bmatrix} 2.63\text{-}3.03 & 16.37\text{-}11.325 & 1.62\text{-}4.593 & 0.47\text{-}3.03 & 20.04\text{-}12.154 \end{bmatrix}$$

$$MD = \sqrt{7.387}$$
$$MD = 2.718$$

The same method calculates the Mahalanobis Distance for the next regency up to North Padang Lawas Regency. Similarly, the same process is carried out for the other 33 regencies, because this study uses 5 variables, the degrees of freedom k=5 and the significance level 0.05 are:

$$C_k = X^2_{5-1;0,05}$$

Meaning based on the Chi-Square table with degrees of freedom and significance of 0.95 is:

$$C_k = X^2_{5-1;0,05}$$
$$C_k = X^2_{4;0,05}$$
$$C_k = 3.356$$

Thus, the results to indicate whether each regency is an outlier or not are presented in the following table.

Table 2. Mahalanobis Distance and Robust Distance of Regencies/Cities in 2018.

No	Kabupaten	Mahalobis Distance	Robust Distance	Keterangan
1	Nias	2.718	2.652	Bukan Outlier
2	Mandailing Natal	1.964	2.026	Bukan Outlier
3	Tapanuli Selatan	1.764	1.631	Bukan Outlier
4	Tapanuli Tengah	1.135	1.14	Bukan Outlier
5	Tapanuli Utara	2.364	2.179	Bukan Outlier
6	Toba Samosir	2.391	2.164	Bukan Outlier
7	Labuhan Batu	1.771	2.195	Bukan Outlier
8	Asahan	1.253	1.846	Bukan Outlier
9	Simalungun	0.832	1.923	Bukan Outlier
10	Dairi	1.591	1.417	Bukan Outlier
11	Karo	2.443	2.292	Bukan Outlier
12	Deli Serdang	2.33	7.515	Outlier
13	Langkat	1.066	2.264	Bukan Outlier
14	Nias Selatan	2.386	2.355	Bukan Outlier
15	Humbang Hasundutan	2.355	2.153	Bukan Outlier
16	Pakpak Barat	1.867	1.933	Bukan Outlier
17	Samosir	1.821	1.813	Bukan Outlier
18	Serdang Bedagai	2.583	2.681	Bukan Outlier
19	Batu Bara	1.032	1.748	Bukan Outlier
20	Padang Lawas Utara	0.783	0.726	Bukan Outlier
21	Padang Lawas	1.637	1.62	Bukan Outlier
22	Labuhan Batu Selatan	1.363	1.858	Bukan Outlier
23	Labuhan Batu Utara	0.621	0.973	Bukan Outlier
24	Nias Utara	3.769	4.895	Outlier
25	Nias Barat	3.694	3.717	Outlier

26	Sibolga	2.012	1.944	Bukan Outlier
27	Tanjung Balai	0.963	1.048	Bukan Outlier
28	Pematang Siantar	3.256	2.908	Bukan Outlier
29	Tebing Tinggi	1.761	1.702	Bukan Outlier
30	Medan	5.07	19.52	Outlier
31	Binjai	1.689	1.843	Bukan Outlier
32	Padangsidimpuan	1.569	1.769	Bukan Outlier
33	Gunungsitoli	1.913	2.108	Bukan Outlier

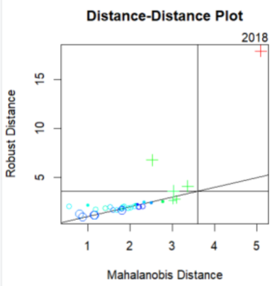

Figure 1. Plot of Mahalanobis Distance against Robust Distance for 2019.

The image shows a plot of Mahalanobis Distance and robust distance for Regencies/Cities in North Sumatra for 2018. From the visualization, four points are significantly far from the cut-off intersection line, indicating that four regencies/cities are outliers in the multivariate data. The regencies/cities detected as outliers through this analysis are Sibolga, Pematang Siantar, and Padangsidempuan.

Then, the same procedure was carried out for the next 5 years, and the results are as follows:

Table 3. List of Regencies/Cities Detected as Outliers.

Tahun	Outlier	Kabupaten/Kota
2018	4	Deli Serdang, Nias Utara, Nias Barat, Medan
2019	4	Deli Serdang, Nias Utara, Nias Barat, Medan
2020	4	Deli Serdang, Nias Utara, Nias Barat, Medan
2021	4	Deli Serdang, Nias Utara, Nias Barat, Medan
2022	4	Deli Serdang, Nias Utara, Nias Barat, Medan

Model-Based Clustering with Integrated Completed Likelihood Criterion

Model-based clustering (MBC) is capable of identifying at least 28 models with a maximum number of groups of 9 using the teigen package in R software. The selection of the optimal group is based on the largest value of ICL[24].

Clustering of Regencies/Cities in 2018

The Teigen package in the R programming language can identify 28 possible models with a maximum number of groups of up to 9 groups for MBC mixture t multivariate with ICL criterion.

The 2018 analysis revealed the highest ICL value of -186.1064. This figure was achieved using two cluster groups in the CICC model.

The marginal contour plot results for the Community Welfare Indicators in 2018 are as follows:

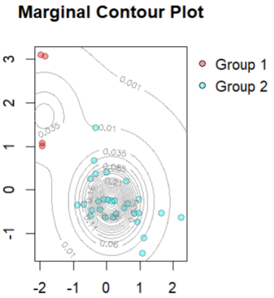

Figure 2. Marginal Contour Plot Data for the Year 2018

By observing Figure 1, which displays the visualization of the formed group members, further information regarding the clustering results of regencies/cities in North Sumatra in 2018 can be found in the following table:

Table 4. Clustering Results of Regencies/Cities in North Sumatra in 2018.

Cluster I		Cluster II	
1.	Nias	1.	Mandailing Natal
2.	Nias Selatan	2.	Tapanuli Selatan
3.	Nias Utara	3.	Tapanuli Tengah
4.	Nias Barat	4.	Tapanili Utara
		5.	Toba Samosir
		6.	Labuhan Baut
		7.	Asahan
		8.	Simalungun
		9.	Dairi
		10.	Karo
		11.	Deli Serdang
		12.	Langkat
		13.	Humbang Hasundutan
		14.	Pakpak Bharat
		15.	Samosir
		16.	Serdang Bedagai
		17.	Batu Bara
		18.	Padang Lawas Utara
		19.	Padang Lawas
		20.	Labuhan Batu Selatan
		21.	Labuhan Batu Utara
		22.	Sibolga
		23.	Tanjung Balai
		24.	Pematang Siantar
		25.	Tebing Tinggi
		26.	Medan
		27.	Binjai
		28.	Padangsidimpuan
		29.	Gunungsitoli

Clustering of Regencies/Cities in 2019

The data analysis results for 2019 show that the highest ICL value is -172.8015. This value was achieved when the data was clustered into 3 groups using the CICU model.

The marginal contour plot of the Community Welfare Indicator data in 2019 yields the following results:

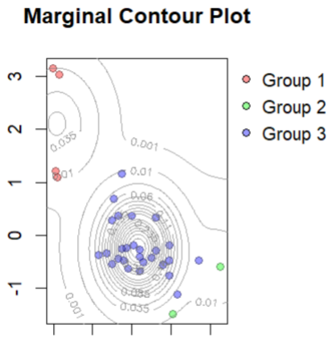

Figure 3. Marginal Contour Plot Data for the Year 2019

Here is a detailed table of the clustering results of regencies/cities 2019.

Table 5. Clustering Results of Regencies/Cities in North Sumatra in 2019.

Cluster I		Cluster II		Cluster III	
1	Nias	1	Mandailing Natal	1	Deli Serdang
2	Nias Selatan	2	Tapanuli Selatan	2	Medan
3	Nias Utara	3	Tapanuli Tengah		
4	Nias Barat	4	Tapanuli Utara		
		5	Toba Samosir		
		6	Labuhan Batu		
		7	Asahan		
		8	Simalungun		
		9	Dairi		
		10	Karo		
		11	Langkat		
		12	Humbang Hasundutan		
		13	Pakpak Barat		
		14	Samosir		
		15	Serdang Bedagai		
		16	Batu Bara		
		17	Padang Lawas Utara		
		18	Padang Lawas		
		19	Labuhan Batu Selatan		
		20	Labuhan Batu Utara		
		21	Sibolga		
		22	Tanjung Balai		
		23	Pematang Siantar		
		24	Tebing Tinggi		
		25	Binjai		
		26	Padangsidimpuan		
		27	Gunungsitoli		

Clustering of Regencies/Cities in 2020

The data analysis results in 2020 show that the highest ICL value is -175.2595. This value was achieved when the data was clustered into 2 groups using the CIUC model. From this model, it can be seen that a clustering pattern with the following characteristics emerged:

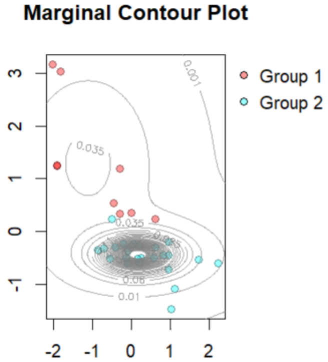

Figure 4. Marginal Contour Plot Data for the Year 2020

Considering Figure 4, which shows the visual representation of the formed group members, differences between group members are indicated through color variations. Further details regarding the clustering results of regencies/cities in North Sumatra in 2020 can be found in the following table.

Table 6. Clustering Results of Regencies/Cities in North Sumatra in 2020.

Cluster I		Cluster II	
1.	Nias	1	Mandailing Natal
2.	Nias Selatan	2	Tapanuli Selatan
3.	Nias Utara	3	Tapanili Utara
4.	Nias Barat	4	Toba Samosir
5.	Tapanuli Tengah	5	Labuhan Baut
6.	Samosir	6	Asahan
7.	Sibolga	7	Simalungun
8.	Tanjung Balai	8	Dairi
9.	Gunungsitoli	9	Karo
		10	Deli Serdang
		11	Langkat
		12	Humbang Hasundutan
		13	Pakpak Bharat
		14	Serdang Bedagai
		15	Batu Bara
		16	Padang Lawas Utara
		17	Padang Lawas
		18	Labuhan Batu Selatan
		19	Labuhan Batu Utara
		20	Pematang Siantar
		21	Tebing Tinggi
		22	Medan
		23	Binjai
		24	Padangsidimpuan

Clustering of Regencies/Cities in 2021

Clustering of regencies/cities in North Sumatra in 2021 was carried out until reaching 2 groups, resulting in 2 groups. This is because the largest ICL value was obtained at G=2, with the largest value being -174.0472. The appropriate framework model for the matrix in this case is the UIUC model.

The clustering process of regencies/cities in North Sumatra in 2021 can be visualized through a marginal contour plot, as shown in the following image.

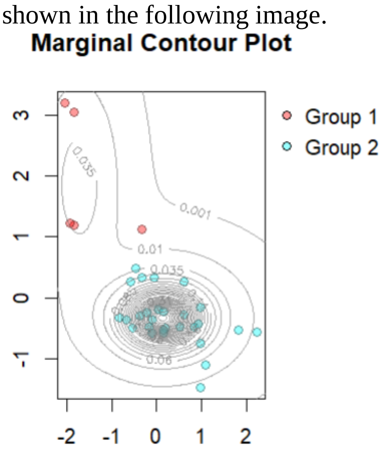

Figure 5. Marginal Contour Plot Data for the Year 2021

The visualization of the formed group members in 2021 is shown in Figure 5. Each group is represented in the contour plot. Further information regarding the clustering results of regencies/cities in North Sumatra in 2021 can be found in the following table.

Table 7. Clustering Results of Regencies/Cities in North Sumatra in 2021.

Cluster I		Cluster II	
1.	Nias	1.	Mandailing Natal
2.	Nias Selatan	2.	Tapanuli Selatan
3.	Nias Utara	3.	Tapanuli Tengah
4.	Nias Barat	4.	Tapanili Utara
5.	Gunungsitoli	5.	Toba Samosir
		6.	Labuhan Baut
		7.	Asahan
		8.	Simalungun
		9.	Dairi
		10.	Karo
		11.	Deli Serdang
		12.	Langkat
		13.	Humbang Hasundutan
		14.	Pakpak Bharat
		15.	Samosir
		16.	Serdang Bedagai
		17.	Batu Bara
		18.	Padang Lawas Utara
		19.	Padang Lawas
		20.	Labuhan Batu Selatan
		21.	Labuhan Batu Utara
		22.	Sibolga
		23.	Tanjung Balai
		24.	Pematang Siantar
		25.	Tebing Tinggi
		26.	Medan
		27.	Binjai
		28.	Padangsidimpuan

Clustering of Regencies/Cities in 2022

The data analysis results for 2022 show that the highest ICL value is -191.6845. This value was achieved when the data was clustered into 2 groups using the CICC model. From this model, it can be seen that a clustering pattern with the following characteristics emerged.

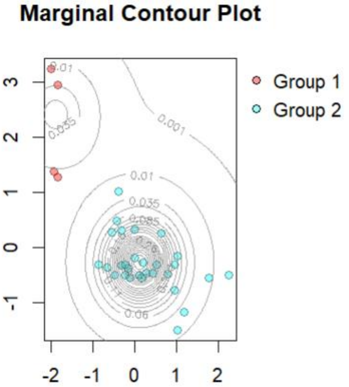

Figure 6. Marginal Contour Plot Data for the Year 2022

Figure 6 shows the visualization of the formed group members in 2022. Theoretically, each group's contour plots should have similar shapes and volumes. However, differences in the results can be caused by the perspective of contour extraction and different variable combinations. More detailed information regarding the clustering results of regencies/cities in North Sumatra in 2022 can be found in the following table.

Table 8. Clustering Results of Regencies/Cities in North Sumatra in 2022.

Cluster I		Cluster II	
1.	Nias	1.	Mandailing Natal
2.	Nias Selatan	2.	Tapanuli Selatan
3.	Nias Utara	3.	Tapanuli Tengah
4.	Nias Barat	4.	Tapanili Utara
		5.	Toba Samosir
		6.	Labuhan Baut
		7.	Asahan
		8.	Simalungun
		9.	Dairi
		10.	Karo
		11.	Deli Serdang
		12.	Langkat
		13.	Humbang Hasundutan
		14.	Pakpak Bharat
		15.	Samosir
		16.	Serdang Bedagai
		17.	Batu Bara
		18.	Padang Lawas Utara
		19.	Padang Lawas
		20.	Labuhan Batu Selatan
		21.	Labuhan Batu Utara
		22.	Sibolga
		23.	Tanjung Balai
		24.	Pematang Siantar
		25.	Tebing Tinggi
		26.	Medan
		27.	Binjai
		28.	Padangsidimpuan
		29.	Gunungsitoli

Cluster Similarity Test

Cluster similarity testing aims to identify significant differences between the formed groups. This process involves testing for mean differences using the Manova method, a multivariate analysis to assess whether the population mean vectors are similar.

Table 9. Results of Group Equality Test with Manova Test.

Tahun	Nilai Pillai's Trace	F	p-value
2018	0.700	12.584	0.000
2019	1.514	16.814	0.000
2020	0.635	9.399	0.000
2021	0.758	16.929	0.000
2022	0.738	15.235	0.000

If the p-value < α at a significance level of 0.05, the decision from the test is to reject the null hypothesis H_0 . The results of the cluster similarity test with Pillai's Trace statistic show that the p-value generated is smaller than the significance level α (0.05) for each year [25]. This indicates significant differences between the mean vectors of the groups in each year. Therefore, it can be concluded that Group 1 significantly differs from the other groups each year. With the rejection H_0 of the null hypothesis, cluster analysis for each city in North Sumatra can be performed.

Conclusion

Based on the analysis and discussion, it can be concluded that model-based clustering can help with grouping. Based on the distance-distance plot for outlier detection, it was found that at the 90th percentile of the data used in this study, there were 3 outliers each year. The formed clusters show that in 2018, 2020, 2021, and 2022, 2 clusters were formed each year, which is the ideal number to use. However 2019, there were only 3 clusters, with the regencies/cities of Nias, South Nias, North Nias, and West Nias consistently in Cluster I. The clustering results from 2018 to 2022 show that Cluster I represents regencies/cities with low HDI and GDP compared to those in Cluster II and III.

References

[1] Sukmasari, D. (2020). Konsep Kesejahteraan Masyarakat Dalam Perspektif Al-Qur'an. *At-Tibyan*, *3*(1), 1–16.

[2] Basofi, A., & Santoso, D. B. (2017). Analisis Pengukuran Kesejahteraan Di Indonesia Jurnal Ilmiah. *Jurnal Ilmiah Mahasiswa FEB*, *10*(2), 1–16.

[3] Qona'ah, N., Devi, A. R., & Dana, I. M. G. M. (2020). Laboratory Clustering using K-Means, K-Medoids, and Model-Based Clustering. *Indonesian Journal of Applied Statistics*, *3*(1), 64.

[4] Martias, L. D. (2021). Statistika Deskriptif Sebagai Kumpulan Informasi. *Fihris: Jurnal Ilmu Perpustakaan Dan Informasi*, *16*(1), 40. https://doi.org/10.14421/fhrs.2021.161.40-59

[5] Nasution, L. E. (2017). STATISTIK DESKRIPTIF. Jurnal Hikmah, 14(1), 49-55.

[6] Ulinnuh, N., & Veriani, R. (2020). Analisis Cluster dalam Pengelompokan Provinsi di Indonesia Berdasarkan Variabel Penyakit Menular Menggunakan Metode Complete Linkage , Average Linkage dan Ward. *InfoTekJar : Jurnal Nasional Informatika Dan Teknologi Jaringan*, *5*(1), 101–108.

[7] Hair, Joseph F., Black, WC, Babin, BJ, et al. 2009. Multivariate Data Analysis (7th ed). Upper Saddle River:Prentice-Hall International, Inc.

[8] Nafisah, Q., & Chandra, N. E. (2017). Analisis Cluster Average Linkage Berdasarkan Faktor-Faktor Kemiskinan di Provinsi Jawa Timur. *Zeta - Math Journal*, *3*(2), 31–36.

[9] Simamora, B. 2005. Multivariate Marketing Analysis. Ed. 1. Jakarta: PT Gramedia Reader.

[10] Nisa, K., Sari, R. F., Cipta, H., & Husein, I. (2020). Cluster Analysis Using the Hierarki Method For Grouping Sub-Districts in The District Steps Based on Health Indicators. ZERO: Jurnal Sains, Matematika Dan Terapan, 4(1), 28.

[11] Agustini, M. (2017). *Model-Based clustering* dengan *Distribusi t Multivariat* Menggunakan Kriteria *Integrated Completed Likelihood* dan *Minimum Message Length.*

[12] Azhar, M. (2020). Pemetaan Tingkat Konsumsi Energi Bahan Bakar Minyak (BBM) Kabupaten/Kota di Jawa Timur dengan *Model Based Clustering*.

[13] Akhyar, S. (2017). Pengelompokan Kabupaten/Kota di Jawa Timur Berdasarkan Indikator Pembangunan Ekonomi Menggunakan *Model Based Clustering*.

[14] Bhagat, A., Kshirsagar, N., Khodke, P., Dongre, K., & Ali, S. (2016). Penalty Parameter Selection for Hierarchical Data Stream Clustering. *Procedia Computer Science*, *79*(May), 24–31.

[15] Melnykov, V., & Maitra, R. (2010). Finite mixture models and model-based clustering. *Statistics Surveys*, *4*(October), 80–116.

[16] Bertoletti, M., Friel, N., & Rastelli, R. (2015). Choosing the number of clusters in a finite mixture model using an exact integrated completed likelihood criterion. *Metron*, *73*(2), 177–199.

[17] Bustaman, N., Yulyanti Shinta, & Dewi, S. K. (2021). *Analisis Faktor-Faktor…{Bustamam, dkk}|*. *32*(1), 85–92. https://journal.uir.ac.id/index.php/kiat

[18] Indonesia, U. I. (2008). *INDEKS PEMBANGUNAN MANUSIA INDONESIA Mohammad Bhakti Setiawan & Abdul Hakim*. 18–26

[19] Waluyo, D. (2006). Studi Tentang Bentuk Kemiskinan Penduduk di Desa Cindogo Kecamatan Tapen Kab. Bondowoso. *Jurnal Humanity*, *1*(2), 11482.

[20] Mahroji, D., & I. Nurkhasanah. 2019. "Pengaruh Indeks Pembangunan Manusia terhadap Tingkat Pengangguran di Provinsi Banten" Jurnal Ekonomi-Qu, 9(1).

[21] Hasibuan, R. R. A., Kartika, A., Suwito, F. A., & Agustin, L. (2022). Pengaruh Produk Domestik Regional Bruto (PDRB) terhadap Tingkat Kemiskinan Kota Medan. *Reslaj : Religion Education Social Laa Roiba Journal*, *4*(3), 683–693.

[22] Jacob, D. E., & Sandjaya. (2018). Faktor faktor yang mempengaruhi kualitas hidup masyarakat Karubaga district sub district Tolikara propinsi Papua. *Jurnal Nasional Ilmu Kesehatan (JNIK)*, *1*(69), 1–16.

[23] Arie, K. B., Hoyyi Abdul, & Abdul, M. M. (2013). Analisis Faktor-Faktor Yang Mempengaruhi Keputusan Pembelian dan Kepuasan Konsumen Pada Notebook Merek Acer. Jurnal Gaussian, 2(1), 29-38

[24] Uly Aldini, & Wara Pramesti. (2020).Pengelompokan Provinsi Di Indonesia Berdasarkan Indikator Mutu Pendidikan Sekolah Menengah Pertama tahun 2016-2018 Menggunakaan Model Based Clustering. *J Statistika: Jurnal Ilmiah Teori Dan Aplikasi Statistika*, *13*(2), 25–38.

[25] Purnomo, Sutadji, E., Utomo, W., Purnawirawan, O., Farich, R., A.S., S., M., R. F., Carina, A., & R., N. G. (2022). *Analisis Data Multivariat*.