Journal of Information System and Technology Research

journal homepage: https://journal.aira.or.id/index.php/jistr/

Determination of Tuberculosis Risk Clusters Based on Health Factors in East Java Using Fuzzy Gustafson Kessel

Naila Mughnifa Qalbi[1], Aviolla Terza Damaliana[2], Mohammad Idhom [3]

[1,2, 3] Data Science , Universitas Pembangunan Nasional "Veteran" Jawa Timur

ARTICLE INFO

Article history:

Received February 18, 2026
Accepted 29 May, 2026
Available online May 31, 2026

Keywords:

Regional Clustering
Fuzzy Gustafson Kessel
Tuberculosis Health
Modifird partition coeffient(MPC)
Clustering

How to Cite :
N. M. Qalbi, A. T. Damaliana, and M. Idhom, "Determination of Tuberculosis Risk Clusters Based on Health Factors in East Java Using Fuzzy Gustafson Kessel," *Journal of Information System and Technology Research*, vol. 5, no. 2, pp. 206–218, May 2026.

ABSTRACT

Tuberculosis (TB) is an infectious disease caused by *Mycobacterium tuberculosis* and remains a major public health problem in Indonesia, particularly in East Java Province. This study aims to group tuberculosis risk levels across 38 districts/cities in East Java Province based on health determinants using the Fuzzy Gustafson-Kessel (FGK) clustering method. The data were obtained from the Central Statistics Agency of East Java Province and the East Java Provincial Health Office in 2024, using four main variables: the number of Diabetes Mellitus (DM) patients, malnourished toddlers, Human Immunodeficiency Virus (HIV) patients, and productive-age active smokers. The FGK method was applied because it can form elliptical clusters through the Mahalanobis distance approach, making it suitable for data with non-homogeneous distribution characteristics. The optimal number of clusters was determined using the Modified Partition Coefficient (MPC). The results show that the four-cluster solution achieved the highest MPC value of 0,987 indicating good cluster partition quality. These four clusters represent tuberculosis risk groups categorized as low priority, medium priority, and high priority across districts/cities in East Java Province. The clustering results can serve as a basis for determining intervention priorities and supporting more targeted public health policy planning.

Corresponding Author:

Naila Mughnifa qalbi
Department of Data Science, Universitas Pembangunan Nasional "Veteran" Jawa Timur.
Email: nailamq0104@gmail.com

1. INTRODUCTION

Tuberculosis (TB) is an infectious disease caused by *Mycobacterium tuberculosis* that primarily attacks the lungs but may also affect other organs. TB remains one of the major infectious diseases contributing to global morbidity and mortality. According to the World Health Organization (WHO), in 2023 there were approximately 10.8 million TB cases globally, with around 1.09 million deaths among people not infected with HIV[1]. Indonesia is one of the countries with the highest TB burden and ranks second after India, with approximately 1,090,000 TB cases and around 125,000 deaths annually, equivalent to approximately 14 deaths every hour [2]. This condition indicates that TB is still a serious public health problem requiring continuous control efforts.

In Indonesia, the burden of TB is unevenly distributed across provinces. West Java Province has the highest number of TB cases, with 234,710 cases, followed by Central Java and East Java[3]. East Java Province is one of the regions with a high TB burden and population density of around 870 people/km². Based on data from the Central Statistics Agency of East Java Province, there were 84,628 TB cases in 2023, with a case detection rate of 91.20% and a treatment success rate of 88.5%. However, in 2024, the number of TB cases increased to 88,733 cases. TB treatment achievement is obtained by

dividing the number of TB cases detected, treated, and reported by the estimated number of TB incidents in the same year[4].The increase in TB cases, accompanied by a decrease in case detection and treatment success rates, indicates that TB control in East Java still faces substantial challenges[5]. Therefore, a more targeted and data-driven strategy is needed, particularly to identify districts and cities with different levels of TB risk.

The vulnerability of a region to TB is closely related to the distribution of risk factors that weaken the immune system and increase exposure to infection. WHO states that TB is more easily transmitted to individuals with weakened immune systems, including people with diabetes mellitus, people living with HIV, malnourished toddlers, and smokers. Diabetes mellitus is one of the important risk factors because individuals with diabetes have a two to three times greater risk of developing TB infection. Hyperglycemia in people with diabetes can impair immune system function, making the body more susceptible to *Mycobacterium tuberculosis* infection [6] In addition, TB also affects children as a vulnerable population. Hyperglycemia in people with diabetes mellitus contributes to a decline in immune system function.

In addition, TB also affects children as a vulnerable population. The Head of the Surabaya City Health Office reported that TB cases among children aged 1–14 years reached approximately 1,300 cases. This indicates that TB transmission does not only occur among adults but also affects children through household and community transmission[7]. Tuberculosis is not limited to adults but has also had a significant impact on children as a vulnerable population. The high number of TB cases in children indicates the potential for active transmission, both within the family and community. This situation also reflects suboptimal health and immune systems in some children, making them more susceptible to infectious diseases like tuberculosis[8]. Children exposed to TB are at risk of impaired growth and development, which may affect their quality of life in the future[9]. Even after successful TB treatment, children and adolescents may still experience respiratory problems and disability that can reduce their quality of life, ability to participate in activities, and growth potential[10].

HIV infection is also strongly associated with TB incidence because it weakens the immune system and increases the risk of developing active TB. People living with HIV are approximately 30 times more likely to develop active TB compared to HIV-negative individuals. The interaction between HIV and TB can accelerate disease progression, as each infection worsens the clinical course of the other[11]. Research conducted by [12] Research shows that poor nutrition in toddlers contributes to an increased risk of tuberculosis. Toddlers with inadequate nutritional status have weaker immune systems, making them susceptible to Mycobacterium tuberculosis infection. Furthermore, research conducted by [8] Research shows that poor nutrition in toddlers contributes to an increased risk of tuberculosis. Toddlers with inadequate nutritional status have weaker immune systems, making them susceptible to Mycobacterium tuberculosis infection. Furthermore, research conducted by [7] proved that active smokers have a 118.774 times higher risk of suffering from tuberculosis.

Another study also showed a positive and significant relationship between smoking habits and TB incidence, where smokers had a 3.81 times greater risk of developing TB compared to non-smokers[13] Smoking risk is particularly relevant in the productive age group, namely 15–64 years, because exposure to nicotine, carbon monoxide, and other harmful substances can impair oxygen supply, damage the respiratory tract, and weaken the immune system [14] Furthermore, the male population has also been reported as a significant variable influencing TB incidence, which may be related to behavioral factors, smoking prevalence, and higher exposure to outdoor activities [15].

The high number of tuberculosis cases in East Java, if not promptly addressed, could potentially lead to a TB endemic. Therefore, grouping districts/cities based on disease factors that weaken the immune system is a strategic step in disease control efforts. The results of this analysis can help identify districts/cities with different risk levels, allowing for more targeted health intervention priorities. Thus, the district/city grouping strategy is expected to minimize the risk of tuberculosis transmission and prevent the spread of TB into an endemic condition. The grouping in this study focused on mapping data on diabetes mellitus sufferers, malnourished toddlers, HIV sufferers, and productive-age smokers to identify priority district/city areas for immediate treatment. Based on the description above, the research problem in this study is the increasing number of tuberculosis cases in East Java Province accompanied by a declining case detection rate and treatment success rate, while the distribution of TB risk factors varies across districts/cities, requiring a more targeted clustering approach[15]. The main purpose of clustering is to identify unseen patterns in data by grouping entities that have similar characteristics and separating entities that have different traits into other groups[16].

The research, conducted by [17] conducted a comparison of the Fuzzy C Means and Fuzzy Gustafson Kessel methods in determining tuberculosis risk clusters in East Java in 2021. The results of the study showed that the Fuzzy Gustafson Kessel method was superior in producing optimal clusters. However, the study only used data on Diabetes Mellitus (DM), HIV, and toddlers with poor nutritional status, without including the variable of active smokers. However, based on survey results and previous studies, active smokers are known to have a high vulnerability to tuberculosis transmission. Therefore, this study aims to complement previous studies by using the Fuzzy Gustafson Kessel method in grouping districts/cities in East Java Province based on more comprehensive tuberculosis risk factors. The Fuzzy Gustafson Kessel method is one of the soft clustering methods developed from the Fuzzy C Means method. In the Fuzzy C Means algorithm, distance calculations use Euclidean distances which produce spherical clusters[18].

The formation of spherical clusters assumes that the data variance is the same in all directions and does not consider the correlation between variables. As a result, it is less than optimal if the data structure has an asymmetric distribution or different variances for each variable.This was reinforced by research conducted by [19] comparing the Fuzzy C Means and

Fuzzy Gustafson Kessel methods in grouping provinces in Indonesia based on crime factors. The study concluded that the Fuzzy Gustafson Kessel method is the best method as indicated by the smaller standard deviation ratio value in the cluster, resulting in a more optimal grouping. Research conducted by[20] shows that the Fuzzy Gustafson-Kessel (FGK) method has superior capabilities in handling large-scale datasets.

This is demonstrated by its ability to produce high-quality clusters by utilizing the Mahalanobis distance as a measure of closeness between data. Initialization of the parameters c (number of clusters) and m (fuzzy weight) is the initial step required before starting the analysis. The parameter c indicates the number of clusters to be formed, while the parameter m determines the level of fuzziness in the cluster division, the greater the value of m, the more spread the object's membership degree to the cluster. Evaluation of cluster results using MPC by calculating the membership degree value of each object to all formed clusters[21]. Research conducted by [22] successfully identified Pedestrian Traffic Behavior Patterns with validity discounting using MPC.The purpose of this study is to determine the tuberculosis risk cluster in each district/city in East Java using the Fuzzy Gustafson Kessel (FGK) method. The results of this analysis are to help identify districts/cities with different risk levels, thus enabling the determination of more targeted health intervention priorities. Thus, the district/city grouping strategy is expected to be able to minimize the risk of tuberculosis transmission and prevent the spread of TB to become an endemic condition.

2. RESEARCH METHOD

This research was conducted through several stages of data processing, starting with data preparation and continuing through the application of the methods used. All these stages are visualized in the research workflow in Figure 1, which systematically depicts the research process.

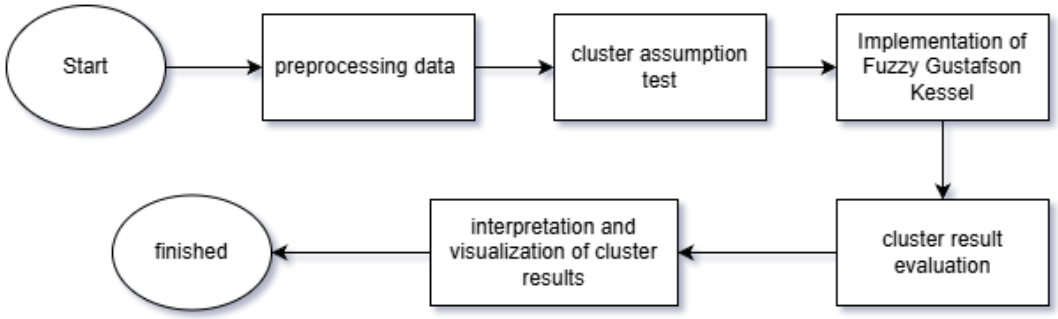

Figure 1. Research Workflow
(Source: Developed by the Author)

The research workflow is illustrated in Figure 1. The process begins with data preprocessing, which aims to prepare the dataset before clustering analysis. This stage includes data collection, variable selection, descriptive statistical analysis, checking data completeness, and data standardization. Standardization is conducted to ensure that all variables are on a comparable scale, thereby preventing variables with larger numerical ranges from dominating the clustering process. The next stage is the cluster assumption test, which is carried out to ensure that the data are appropriate for clustering analysis. This stage includes multicollinearity testing using the Variance Inflation Factor (VIF) and sampling adequacy testing using the Kaiser-Meyer-Olkin (KMO) measure. The VIF test is used to detect high correlations among variables, while the KMO test is used to assess whether the correlation structure among variables is adequate for further analysis. Data that meet these assumptions can then be processed using the clustering method.

After the assumption testing, the Fuzzy Gustafson-Kessel method is implemented. This stage begins with the initialization of parameters, including the number of clusters, fuzziness parameter, convergence threshold, maximum iteration, and initial membership matrix. The process continues by calculating the cluster centers based on the membership degree of each object. Furthermore, the covariance matrix for each cluster is computed to capture the variance and correlation structure of the data. The Mahalanobis distance is then calculated to measure the distance between each object and each cluster center. The membership matrix and cluster centers are updated repeatedly until the algorithm reaches convergence.

The clustering results are then evaluated using the Modified Partition Coefficient (MPC). This evaluation is used to measure the clarity of the fuzzy partition and to determine the optimal clustering result. A higher MPC value indicates clearer cluster membership, while a lower value indicates more ambiguous membership. The final stage is the interpretation and visualization of the clustering results. Each cluster is interpreted based on the characteristics of tuberculosis risk factors, and the results are visualized through tables, plots, and maps to support the identification of priority areas for tuberculosis control in East Java Province.

2.1 Dataset

The data used in this study are secondary data obtained from the Central Statistics Agency (BPS) and the East Java Provincial Health Office for the year 2024. The dataset consists of tuberculosis risk factor variables at the district/city level in East Java Province, including the number of diabetes mellitus cases, HIV cases, malnourished toddlers, active smokers, and other relevant variables used to support tuberculosis risk clustering analysis.

2.2 Data Preprocessing

The data used is secondary data from the Central Statistics Agency (BPS) and the East Java Provincial Health Office (DINKES). Before the clustering process was carried out, data preprocessing was conducted to ensure that the dataset was ready for analysis. The preprocessing stage began with descriptive statistical analysis to identify the general characteristics of the data, including the minimum value, maximum value, mean, and standard deviation of each variable. This step was used to provide an initial overview of the distribution and variation of tuberculosis risk factor variables across districts/cities in East Java Province. Table 1 presents the description and data type of each variable used in this study.

Table 1. Research Variable Data

Variable	Name Variable	Data Types
X_1	Hiv	People
X_2	Diabetes melitus	People
X_3	Malnourished Toddlers	People
X_4	Active Smokers Ages 15-24	People
X_5	Active Smokers Ages 25-34	People
X_6	Active Smokers Ages 35-44	People
X_7	Active Smokers Ages 45-55	People
X_8	Active Smokers Ages 55-64	People

(Source: Developed by the Author)

In addition, the characteristics of the data were visualized in the form of a map to illustrate the spatial distribution of tuberculosis risk factors in East Java Province. This visualization provides a clearer understanding of regional differences before the clustering process is performed. Figure 2 presents the spatial representation of the data characteristics used in this study. After the descriptive analysis and spatial visualization, data standardization was conducted to transform all variables into a comparable scale. Standardization is necessary because the variables have different units of measurement and value ranges. Thus, this process ensures that each variable contributes proportionally to the clustering analysis and prevents variables with larger numerical values from dominating the model.

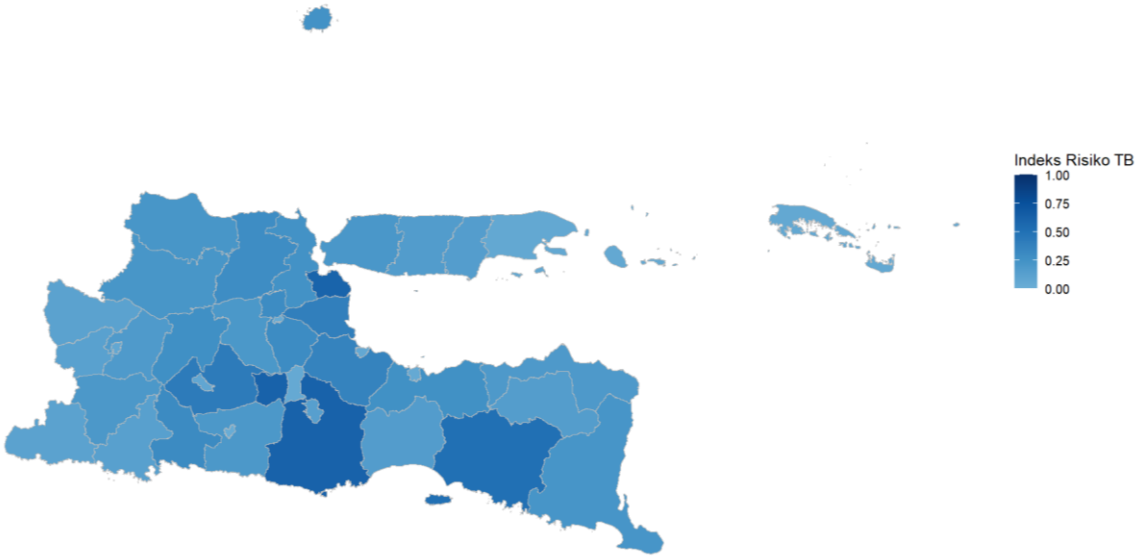

Figure 2. Distribution Map of Data Characteristics
(Source: Developed by the Author)

Figure 2 shows the spatial distribution of tuberculosis risk factor characteristics in East Java Province. The map provides an initial overview of regional variation based on the TB risk index, where darker blue indicates higher risk characteristics and lighter blue indicates lower risk characteristics. This visualization supports the preliminary understanding of spatial patterns before the clustering analysis is performed.

2.3 Research Methods

This study uses the Fuzzy Gustafson Kessel method which applies the concept of fuzzy logic where a value is not absolutely true or false, but can be between the two. The fuzzy clustering method does not force each data point to enter a particular cluster, but rather provides a degree of membership in each cluster. The value of the degree of membership is in the range of 0 to 1[23]. Before cluster modeling is carried out, there are assumptions that must be met, namely a representative sample and the absence of multicollinearity. The representative sample test is carried out using Kaiser Meyer Olkin (KMO) if the KMO value is in the range of 0.5 to 1 indicating that the sample has represented the population. The formula is shown in equation (1).

$$\text{KMO} = \frac{\Sigma_{i=1}^{p}\Sigma_{j}^{p} r_{ij}^{2}}{\Sigma_{i=1}^{p}\Sigma_{j=1}^{p} r_{ij}^{2} + \Sigma_{i=1}^{p}\Sigma_{j=1}^{p} r_{ij}^{2}} \quad (1)$$

Then, a non-multicollinearity test is carried out to determine whether or not there is a correlation between independent variables using the Variance Inflation Factor (VIF). The model is said not to experience multicollinearity if the VIF value for each variable is less than 10 (VIF < 10) [24]. The non-multicollinearity test equation can be seen in equation (2).

$$\text{VIF} = \frac{1}{1-R_j^2} \quad (2)$$

Next, data standardization is performed using the Z-Score to equalize the scale between variables. This standardization is done by subtracting each data value from its average value, then dividing it by the standard deviation of each variable, so that the distance between data points can be measured proportionally. The results of this z-score are in the range of -2 to +2 z-score equation can be seen in equation (3).

$$\text{Z} = \frac{x_i - \bar{x}}{s} \quad (3)$$

After all clustering assumptions have been met, the next step is to perform cluster modeling. The cluster modeling in this study uses the Fuzzy Gustafson-Kessel method, the application of which can be seen in the equation below.

1. Initialize the initial parameters such as determining the number of initial clusters (c≥ 2), fuzzifier value (fuzziness level) (m>1), maximum iteration (t_{max}), smallest error (ε) and initial iterations (t =1)
2. Generating random numbers with matrix symbols $\boldsymbol{U}_0$. Functions of the matrix $\boldsymbol{U}_0$ This is a randomly selected membership with a value range of 0 to 1. Which is in accordance with the formula $\sum \mu_{ik}^{m} = 1$.

$$\boldsymbol{U}_0 = \begin{bmatrix} \mu_{11} & \cdots & \mu_{1c} \\ \vdots & \ddots & \vdots \\ \mu_{1n} & \cdots & \mu_{nc} \end{bmatrix} \quad (4)$$

3. Calculate the value of the kth cluster center (v_k) on k = 1,2,…,c and j = 1,2,…,m with the equation (5).

$$v_{kj} = \frac{\Sigma_{i=1}^{n}(\mu_{ik})^{m} x_{ij}}{\Sigma_{i=1}^{n}(\mu_{ik})^{m}} \quad (5)$$

4. Calculate the value of the covariance matrix in the cluster with the equation (6)

$$F_i = \frac{\Sigma_{i=1}^{n}(\mu_{ik})^{m}((X_{ij}-V_{kj})^{T})(X_{ij}-V_{kj})^{T}}{\Sigma_{i=1}^{n}(\mu_{ik})^{m}} \quad (6)$$

5. Calculating the mahalanobis distance ($\mathbf{D}_{ic_i}^2$)

$$D_{ikAk}^2 = (X_{ij} - V_{kj})^T \left[(\rho_I \det (F_i)^{\frac{1}{n}} F_1^{-1} \right] (X_{ij} - v_{kj}) \quad (7)$$

6. Calculate the value of the objective function in the iteration - t (P_t)

$$J_t = \sum_{k=1}^{C} \sum_{j_i=1}^{p} (\mu_{ik})^m \times D_{ikAk}^2 \quad (8)$$

7. Updating the value of the membership function, namely U_{t+1}

$$U_{(t+1)} = \left[\sum_{h=1}^{c} \left(\frac{D_{i_k}^2}{D_{i_h}^2} \right)^{\frac{1}{m-1}} \right]^{-1} \quad (9)$$

8. Repeat steps 3 to 7 until the objective function difference condition is met. $|P_t - P_{t-1}| < \varepsilon$ and the iteration reaches the maximum limit $(t > t_{max})$ fulfilled.

Evaluation of cluster results using MPC is a development of the Partition Coefficient (PC) validity index, which has a weakness, namely the value of the index experiences monotonic changes as the number of clusters increases. To overcome this weakness, the Modified Partition Coefficient (MPC) is an improvement because it can provide more accurate partition quality results without being affected by the number of clusters used. MPC measures the distance between the degree of membership and the cluster center. The MPC equation can be seen in the equation (10).

$$\text{MPC} = 1 - \frac{c}{c-1}(1 - PC) \quad (10)$$

9. With PC calculations using the equation (11) :

$$PC(c) = \frac{1}{p} \sum_{i=1}^{p} \sum_{k=1}^{c} (\mu_{ik})^2 \quad (11)$$

10. Good quality clustering results are indicated by an MPC value close to 1 because the cluster separation is clearer.

3. RESULTS AND DISCUSSION

The results of the discussion regarding descriptive statistics were conducted on each variable with a total of 38 observations representing districts/cities in East Java Province. Descriptive statistics serve to provide an overview of the data characteristics. A summary of the descriptive statistics for all variables is presented in Table 2 below.

3.1 Data Preparation

The data preparation stage began with the collection of secondary data obtained from the Central Statistics Agency (BPS) and the East Java Provincial Health Office. The data used in this study consist of tuberculosis risk factor variables at the district/city level in East Java Province in 2024. One of the variables used in this study is the number of active smokers by age group. This variable was obtained through an estimation process using data from the National Socio-Economic Survey (Susenas). The percentage of the population who smoked in each age group was multiplied by the projected population data published by BPS for the corresponding year. Through this calculation, the estimated number of active smokers in each age group for every district/city in East Java Province was obtained.

3.2 Preprocessing Data

Data preprocessing was carried out to prepare the dataset before the clustering process. In this study, the preprocessing stage began with data standardization, because the variables used have different units of measurement and value ranges. For example, the number of HIV cases, diabetes mellitus cases, malnourished toddlers, and active smokers by age group have different numerical scales. Table 2 below is the result of the data standardization process.

Table 2. Data standardization

No	HIV	Diabetes Melitus	Malnourished Toddlers	Active Smokers Ages 15-24	Active Smokers Ages 25-34	Active Smokers Ages 35-44	Active Smokers Ages 45-55	Active Smokers Ages 55-64
1	-1.04080	-0.634165	-0.750060	-0.406822	-0.27275	-0.09243	-0.01641	-0.64638
2	-0.12730	-0.352569	-0.190157	-0.794286	-0.79552	0.069052	0.35721	1.06507
3	-0.95074	-0.634262	-0.625806	-0.246430	-0.22858	-0.00045	-0.35629	-0.0708
4	1.44237	-0.286163	0.008865	0.240757	0.33570	0.492954	0.50453	0.51057
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
38	-1.02794	-0.757829	-0.994227	-0.810771	-0.48331	-0.58164	-0.89396	-0.48073

(Source: Developed by the Author)

The standardization process was conducted to transform all numerical variables into a comparable scale. The results show that each variable was converted into a standardized score with a mean close to 0 and a standard deviation of 1. Positive values indicate that the corresponding district/city has a variable value above the overall average, while negative values indicate a value below the average. For instance, the standardized values show relatively high scores for HIV, diabetes mellitus, malnourished toddlers, and active smokers in several districts/cities, indicating that these regions have higher tuberculosis risk factor characteristics compared to others. This process ensures that variables with larger numerical ranges, particularly active smoker variables, do not dominate the clustering process. Therefore, the standardized data are considered suitable for further clustering analysis using the Fuzzy Gustafson-Kessel method.

3.3 VIF Assumption Test

The results of the Z-score standardization for all 38 districts/cities in East Java Province are presented in Table 2. After the data standardization process, the next step is to test for multicollinearity using the Variance Inflation Factor (VIF). Multicollinearity testing is conducted to ensure that there is no high linear correlation between independent variables, which could affect the stability of the clustering results. A variable is declared free from multicollinearity if it has a VIF value of less than 10 [25]. The VIF calculation formula is shown in Equation (2), and the complete VIF test results for each variable are presented in Table 3.

Table 3. Results of Non-Multicollinearity Test

Variable	VIF
Hiv	3.43
Diabetes mellitus	4.46
Malnourished Toddlers	2.09
Active Smokers Ages 15-24	2.25
Active Smokers Ages 25-34	2.83
Active Smokers Ages 35-44	3.94
Active Smokers Ages 45-55	2.18
Active Smokers Ages 55-64	1.36

(Source: Developed by the Author)

Based on Table 3, the results of the multicollinearity test show that all variables have Variance Inflation Factor (VIF) values below 10, with the highest VIF value recorded for the Diabetes Mellitus variable at 4.46 and the lowest for the Active Smokers Ages 55–64 variable at 1.36. Since all VIF values are well below the threshold of 10, it can be concluded that there is no multicollinearity among the variables used in this study. This indicates that each variable contributes independent information to the analysis, so the clustering process can proceed to the next stage.

3.4 KMO Assumption Test

Subsequently, a sampling adequacy test was conducted for all variables using the Kaiser-Meyer-Olkin (KMO) measure. The KMO test is used to assess whether the correlation pattern between variables is suitable for further multivariate analysis. A KMO value greater than 0.5 indicates that the sampling adequacy assumption has been met and the data is appropriate for further analysis [26]. The KMO calculation formula is presented in Equation (1), and the KMO test results are shown in Table 4.

Table 4. Kaiser Mayer Olkin (KMO) Test

Overall KMO value = 0.74	
Variable	Overall KMO
Hiv	0.71
Diabetes mellitus	0.73
Malnourished Toddlers	0.71
Active Smokers Ages 15-24	0.64
Active Smokers Ages 25-34	0.84
Active Smokers Ages 35-44	0.77
Active Smokers Ages 45-55	0.79
Active Smokers Ages 55-64	0.48

(Source: Developed by the Author)

Based on Table 4 regarding the statistical results of the sample adequacy test for each variable, the total overall KMO value was 0.74, which (0.7 ≥0.5) means it failed to reject H_0 so that the sample used is sufficient to continue the cluster analysis process. After all clustering assumptions are met, the next step is to conduct the cluster analysis process using the Fuzzy Gustafson Kessel (FGK) method.

3.5 Results of the Gustafson Kessel fuzzy modeling

After all clustering assumptions are met, the next step is to conduct the cluster analysis process using the Fuzzy Gustafson Kessel (FGK) method. In this study, the initial parameter initialization uses three clusters, from clusters 2 to 4, with the parameter m = 2, t maximum = 1000, $\varepsilon = 10^{-5}$. The next step was to calculate the cluster centers based on the membership degree of each object. After the cluster centers were obtained, the covariance matrix for each cluster was computed to capture the variance and correlation structure of the data. The Mahalanobis distance was then calculated to measure the distance between each object and each cluster center. This distance was used to update the membership values, where objects closer to a cluster center had higher membership degrees in that cluster.

The process was repeated iteratively by recalculating the cluster centers, covariance matrices, Mahalanobis distances, and membership values until convergence was achieved or the maximum number of iterations was reached. The results of the Fuzzy Gustafson-Kessel modeling with various combinations of $c$and $m$are presented in Table 5. Each model was validated using the Modified Partition Coefficient (MPC). MPC evaluates the quality of fuzzy clustering based on the membership degree of each object in the formed clusters. A higher MPC value indicates that the objects have more distinct membership in a particular cluster, while a lower MPC value indicates that the membership is more ambiguous. Thus, the optimal clustering model is selected based on the highest MPC value.

Table 5. Cluster Result Output

Number of clusters (c)	Weighting rank (m)	Iterations	objective function	MPC
2	2	20	64.28679	0.785
3	2	20	52.81224	0.987
4	2	20	9.225466	0.632

(Source: Developed by the Author)

Based on Table 5 presents the clustering results obtained from the Fuzzy Gustafson-Kessel (FGK) method for varying numbers of clusters (c = 2, 3, and 4), all using a fuzzy weighting parameter of m = 2. The evaluation of the optimal number of clusters is based on the Modified Partition Coefficient (MPC) value, where a higher MPC value closer to 1 indicates a better and more distinct cluster partition. For c = 2, the algorithm converged after 20 iterations with an objective function value of 64.28679 and MPC value of 0.785. For c = 3, the algorithm converged faster at 20 iterations, producing an objective function value of 52.81224 and mpc value of 0.987. For c = 4, the algorithm converged at 20 iterations with an objective function value of 9.225466 and the MPC value of 0.632.

Based on these results, the configuration of four clusters three is determined as the optimal solution in this study, as it produces the highest MPC value of 0. 0.987, which is the closest to 1 among all tested configurations. An MPC value approaching 1 indicates that each district/city has a clearly dominant membership degree to one particular cluster, meaning the cluster boundaries are well-defined and the partition quality is high. Conversely, an MPC value close to 0 would suggest high overlap between clusters, indicating unclear grouping. Therefore, the MPC value of 0.987 obtained in this study confirms that the four-cluster solution provides a clustering structure with high clarity, low inter-cluster overlap, and strong separation between cluster members. Figure 3 below presents the visualization of the clustering results for $c = 3$ and $m = 2$ using Principal Component Analysis (PCA). This visualization is used to illustrate the distribution pattern of districts/cities in East Java Province based on the formed clusters.

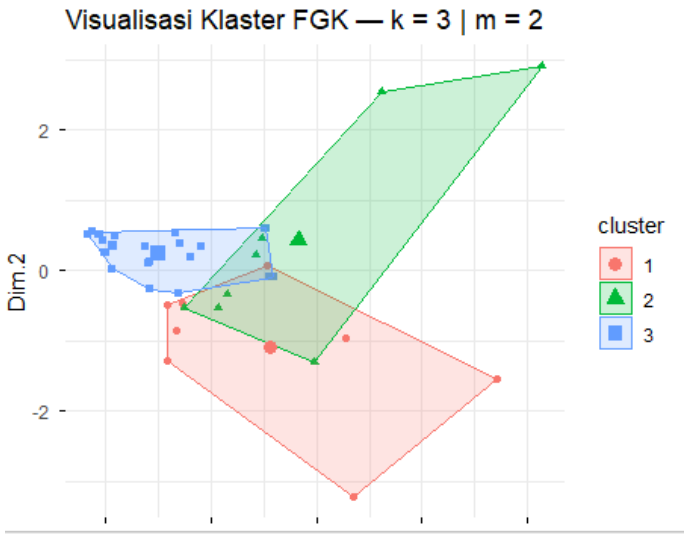

Figure 3. Visualization of cluster *c* =3, *m* =2
(Source: Developed by the Author)

Based on Table 5, the Fuzzy Gustafson-Kessel (FGK) clustering results show that the three-cluster configuration (c=3) with m=2 produced the highest MPC value of 0.987. This indicates that the clustering structure has clear membership, low overlap among clusters, and good partition quality. Therefore, c=3 was selected as the optimal number of clusters in this study. Figure 3 presents the PCA visualization of the clustering results for c=3 and m=2. This visualization illustrates the distribution pattern of districts/cities in East Java Province based on the formed clusters. Each point represents a district/city, while different colors and shapes indicate different cluster memberships. The visualization shows that the FGK method is able to separate the data into three groups with relatively distinct patterns. Although several points appear close to one another or slightly overlap, this condition is reasonable in fuzzy clustering because each district/city has a degree of membership in every cluster. Overall, the PCA visualization supports the MPC validation results, indicating that the three-cluster solution provides a good representation of tuberculosis risk grouping in East Java Province.

3.6 Cluster Interpretation

After the optimal cluster was determined using the Modified Partition Coefficient (MPC) validation, the next step was to perform cluster profiling. Cluster profiling aims to identify and describe the characteristics of each cluster based on the average value of the variables used in the study. Through this process, each cluster can be interpreted according to the dominant tuberculosis risk factors, such as HIV, diabetes mellitus, malnourished toddlers, and active smokers by age group.Cluster profiling is important because the clustering results only show the grouping of districts/cities, while the profile of each cluster explains the meaning and characteristics of the formed groups. Therefore, the profiling process helps determine which clusters have higher, moderate, or lower tuberculosis risk characteristics. The results of the cluster profiling are presented in Table 6 below.

Table 6. Average Profile of Variable Values for Each Cluster

Variable	Cluster 1	Cluster 2	Cluster 3
Hiv	97.75	154.62	75.59
Diabetes mellitus	29.34	43.92	13.45
Malnourished Toddlers	4,550	5,196	1,786
Active Smokers Ages 15-24	421,971	229,142	110,320
Active Smokers Ages 25-34	546,169	577,742	171,070
Active Smokers Ages 35-44	326,149	560,414	231,575
Active Smokers Ages 45-55	489,959	398,922	219,877
Active Smokers Ages 55-64	539,423	26,163	102,263

(Source: Developed by the Author)

Based on Table 6, Cluster 2 is categorized as high priority because it has the highest average scores for several key tuberculosis risk variables: HIV (154.62%), diabetes mellitus (43.92%), malnutrition in children under five

(5,196.75%), active smokers aged 25–34 (577,742,543), and active smokers aged 35–44 (560,414,700). High scores for these risk factors indicate that areas in Cluster 2 have greater tuberculosis vulnerability characteristics than other clusters.

Cluster 1 is categorized as medium priority. This cluster has relatively high scores for several variables, particularly active smokers aged 15–24 (421,971,032), smokers aged 45–55 (489,959,713), and active smokers aged 55–64 (539,423,485). Furthermore, the HIV, diabetes mellitus, and malnutrition rates in Cluster 1 are also at a moderate level. Therefore, this cluster exhibits significant risk characteristics, but not as high as Cluster 2 for key variables such as HIV, diabetes mellitus, and malnutrition in children.

Meanwhile, Cluster 3 is categorized as low priority because it has the lowest average scores for most tuberculosis risk variables. This cluster has an HIV score of 75.59, diabetes mellitus of 13.45, malnutrition in children of 1,786.95, and active smoking rates in most age groups, which are lower than those in Clusters 1 and 2. Therefore, areas in Cluster 3 have relatively lower tuberculosis risk characteristics than the other clusters. Furthermore, Table 7 presents the grouping results of districts/cities into three clusters based on the highest membership degree.

Table 7. Results of grouping regions based on district/city

Regency/City	Cluster
Pacitan, Ponorogo, Trenggalek, Tulungagung, Lumajang, Bondowoso, Situbondo, Nganjuk, Madiun, Magetan, Ngawi, Lamongan, Pamekasan, Sumenep, Kota Kediri, Kota Blitar, Kota Malang, Kota Probolinggo, Kota Pasuruan, Kota Mojokerto, Kota Madiun, Kota Batu	Cluster 3 (Low Priority)
Blitar, Kediri, Malang, Jember, Banyuwangi, Mojokerto, Bojonegoro, Bangkalan	Cluster 1 (Medium Priority)
Probolinggo, Pasuruan, Sidoarjo, Jombang, Tuban, Gresik, Sampang, Kota Surabaya	Cluster 2 (High Priority)

(Source: Developed by the Author)

Based on Table 6, each cluster shows different characteristics of tuberculosis risk factors. Cluster 2 is categorized as the high-priority cluster because it has the highest average values for several key risk factors, including HIV, diabetes mellitus, malnourished toddlers, active smokers aged 25–34 years, and active smokers aged 35–44 years. These characteristics indicate that districts/cities in Cluster 2 have a higher vulnerability to tuberculosis and require more intensive intervention.

Cluster 1 is categorized as the medium-priority cluster. This cluster has relatively high values in several smoking-related variables, particularly active smokers aged 15–24 years, 45–55 years, and 55–64 years. Although the values of HIV, diabetes mellitus, and malnourished toddlers are lower than those in Cluster 2, Cluster 1 still shows moderate tuberculosis risk characteristics and therefore requires preventive monitoring.

Meanwhile, Cluster 3 is categorized as the low-priority cluster because it has the lowest average values for most tuberculosis risk factor variables. This indicates that districts/cities in Cluster 3 have relatively lower tuberculosis risk characteristics compared to the other clusters. However, routine surveillance and health promotion are still needed to prevent an increase in risk.

Based on the cluster profiling results, the priority classification can be summarized as follows: Cluster 2 represents the high-priority group, Cluster 1 represents the medium-priority group, and Cluster 3 represents the low-priority group. Furthermore, Table 7 presents the distribution of districts/cities in each cluster based on the highest membership degree.

3.7 Map visualization

After the clustering results, through visualization, the clustering results can be interpreted more easily, particularly in identifying the differences in regional characteristics and the distribution of districts/cities within each cluster. Figure 4 presents the spatial distribution of tuberculosis risk clusters in East Java Province based on the Fuzzy Gustafson-Kessel clustering results. The map shows the grouping of districts/cities into three priority levels: high, medium, and low.

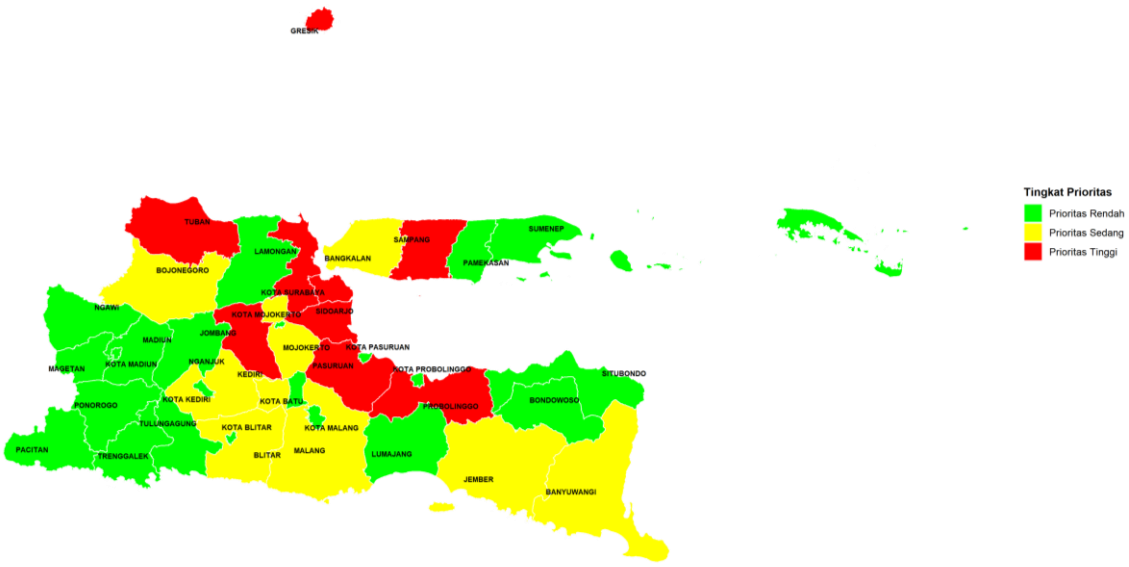

Figure 4. Cluster result visualization map
(Source: Developed by the Author)

The red color represents the high-priority cluster, the yellow color represents the medium-priority cluster, and the green color represents the low-priority cluster. The high-priority cluster, shown in red, consists of Probolinggo, Pasuruan, Sidoarjo, Jombang, Tuban, Gresik, Sampang, and Surabaya City. These areas are categorized as high priority because they have higher average values for several tuberculosis risk factors, such as HIV, diabetes mellitus, malnourished toddlers, and active smokers in productive age groups.

The medium-priority cluster, shown in yellow, consists of Blitar, Kediri, Malang, Jember, Banyuwangi, Mojokerto, Bojonegoro, and Bangkalan. These areas have moderate tuberculosis risk characteristics, with relatively high values in several smoking-related variables but lower values than the high-priority cluster in some main risk factors.Meanwhile, the low-priority cluster, shown in green, consists of Pacitan, Ponorogo, Trenggalek, Tulungagung, Lumajang, Bondowoso, Situbondo, Nganjuk, Madiun, Magetan, Ngawi, Lamongan, Pamekasan, Sumenep, Kediri City, Blitar City, Malang City, Probolinggo City, Pasuruan City, Mojokerto City, Madiun City, and Batu City. These areas generally have lower average values for most tuberculosis risk factor variables compared to the other clusters.

Overall, the map provides a clearer spatial representation of tuberculosis risk distribution across East Java Province. The visualization helps identify areas that require different levels of intervention priority. Regions in the red cluster require greater attention in tuberculosis control programs, while regions in the yellow cluster require preventive monitoring. The green cluster represents areas with relatively lower risk characteristics, although routine surveillance and prevention efforts are still needed.

4. CONCLUSION

This study found that tuberculosis risk in East Java Province can be grouped into three clusters based on health risk factors using the Fuzzy Gustafson-Kessel (FGK) method. The data were considered suitable for clustering analysis, as indicated by an overall KMO value of 0.74 and VIF values below 10, showing that the variables met the sampling adequacy requirement and did not indicate serious multicollinearity. The clustering evaluation showed that the three-cluster configuration ($c = 3$, $m = 2$) produced the highest MPC value of 0.987, indicating a clear fuzzy partition, low overlap among clusters, and good cluster separation. This result confirms that the FGK method is able to form representative clusters for regional tuberculosis risk analysis in East Java Province.

The main empirical finding shows that cluster 2 is the highest-priority cluster**.** This cluster consists of Probolinggo, Pasuruan, Sidoarjo, Jombang, Tuban, Gresik, Sampang, and Surabaya City. These areas have higher average values for several major tuberculosis risk factors, including HIV, diabetes mellitus, malnourished toddlers, and active smokers in productive age groups. Therefore, districts/cities in Cluster 2 should receive greater intervention priority through strengthened early TB case detection, routine screening for vulnerable groups, improved treatment coverage, and smoking prevention programs.

Methodologically, this study contributes by applying the FGK method to tuberculosis risk clustering at the district/city level. The FGK method is appropriate for regional health data because it uses Mahalanobis distance and can adapt to differences in variance and covariance structure among variables. The use of MPC also provides an objective basis for determining the

optimal cluster configuration.The clustering results can support policymakers in designing more targeted tuberculosis control strategies in East Java Province.

High-priority areas require intensive intervention, medium-priority areas require preventive monitoring, and low-priority areas still require routine surveillance to prevent an increase in tuberculosis risk. However, this study is limited by the use of cross-sectional secondary data, so it cannot explain causal relationships or temporal changes in tuberculosis risk. Future research is recommended to include longitudinal data, socio-economic variables, environmental variables, and comparisons with other clustering methods to obtain more comprehensive and robust results.

5. ACKNOWLEDGEMENT

The author expresses sincere gratitude to God Almighty for the grace that made this research possible to be completed. Appreciation is extended to the supervising lecturer for valuable guidance, feedback, and continuous support during the preparation of this article, as well as to Universitas Pembangunan Nasional "Veteran" Jawa Timur for providing academic facilities. The author also thanks the Central Statistics Agency (BPS) and the East Java Provincial Health Office for providing the data used in this study, and all parties who contributed directly or indirectly to the completion of this research

6. REFERENCES

[1] UNICEF, "Desk Review: Pediatric Tuberculosis with a Focus on Indonesia," 2022.
[2] Tim kerja TBC, "Gerakan Indonesia Akhiri TBC dengan Komitmen dan Aksi Nyata (GIATKAN)," *kemenkes*. [Online]. Available: https://ayosehat.kemkes.go.id/gerakan-indonesia-akhiri-tbc-dengan-komitmen-dan-aksi-nyata-giatkan
[3] Brilliant Ayang Iswenda, "Jawa Barat Jadi Provinsi dengan Kasus TBC Terbanyak Sepanjang 2024," GoodStats Indonesia. [Online]. Available: https://goodstats.id/article/jawa-barat-menjadi-provinsi-dengan-kasus-tbc-terbanyak-sepanjang-2024-cCYo6
[4] J. Epidemiologi and K. Komunitas, "Jekk kk 1," vol. 10, no. 2, pp. 1–9, 2025.
[5] G. W. Regression, "Indonesian Journal of Applied Statistics," vol. 6, no. 2, pp. 116–124, 2023.
[6] W. Gotera *et al.*, "Diabetes Melitus sebagai Faktor Risiko Tuberkulosis Diabetes Mellitus as A Risk Factor for Tuberculosis," vol. 27, no. 3, pp. 273–281.
[7] C. M. Dewani, D. Anandari, and S. Rahardjo, "Media Kesehatan Masyarakat Indonesia Analisis Faktor Risiko dan Pemetaan Kasus Tuberkulosis Paru di Wilayah Kerja Puskesmas Baturraden I Kabupaten Banyumas," vol. 2024, no. 23, pp. 223–228, 2024.
[8] S. Aliviana, P. Sari, D. Astuti, R. Widyastuti, and M. Malang, "Identifikasi Faktor Risiko Terhadap Terjadinya Penyakit Tuberculosis," vol. 4, no. 2, pp. 124–132, 2023.
[9] J. Wang, "Global burden and trend of tuberculosis in children and adolescents (under 15 years old) from 1990 to 2021 , with projections to 2040," no. June, pp. 1–13, 2025, doi: 10.3389/fpubh.2025.1578658.10.3389/fpubh.2025.1578658.
[10] K. Romanowski, S. S. Chiang, S. A. Land, M. M. Van Der Zalm, and J. R. Campbell, "Articles Tuberculosis-associated respiratory impairment and disability in children and adolescents : a systematic review," *eClinicalMedicine*, vol. 81, no. February, p. 103107, 2025, doi: 10.1016/j.eclinm.2025.103107.10.1016/j.eclinm.2025.103107.
[11] Anisa Putri Utami, Ika Restu Kaeksi, Nisa Wahyuningsih, and Liss Dyah Dewi Arini, "Infeksi Menular Seksual," *J. Mhs. Ilmu Kesehat.*, vol. 3, no. 1, pp. 208–215, 2025, doi: 10.59841/jumkes.v3i1.2323.10.59841/jumkes.v3i1.2323.
[12] N. Hasnanisa, S. Prasetyo, Y. Handayani, P. Studi, K. Masyarakat, and F. K. Masyarakat, "Faktor-faktor Tuberkulosis Paru : Analisis Spasial Factors of Pulmonary Tuberculosis : Spatial Analysis Partisipan dan Desain Studi," vol. 15, no. September, pp. 107–118, 2023.
[13] U. B. Dwipa, "Korelasi Antara Merokok dan Infeksi Tuberkolosis (TBC)," vol. 19, no. 03, pp. 163–167, 2024.
[14] F. N. Poetri, "Hubungan antara Kenaikan Kasus TBC dan Kebiasaan Merokok di Kalangan Remaja dan Dewasa." [Online]. Available: https://terbesarjogja.com/hubungan-antara-kenaikan-kasus-tbc-dan-kebiasaan-merokok-di-kalangan-remaja-dan-dewasa/#:~:text=Menurut Badan Penelitian dan Pengembangan Kesehatan%2C individu yang,untuk terkna tuberkulosis dibandingkan mereka yang tidak merokok
[15] P. Tuberkulosis and T. B. C. Di, "Faktor-faktor yang memengaruhi penyebaran penyakit tuberkulosis (tbc) di provinsi jawa barat," vol. 9, no. 3, pp. 165–170, 2020.
[16] E. Fortuna, D. A. Prasetya, and K. M. Hindrayani, "X-Means Clustering for Segmenting Property Customer Payment Behaviors," vol. 5, no. 1, pp. 1–14, 2026.

[17] A. Nurzida, I. T. Utami, and M. Y. Rochayani, "Perbandingan Metode Fuzzy C-Means Dan Gustafson-Kessel Dalam Penentuan Cluster Tingkat Risiko Penularan Tuberculosis Terhadap Penyakit Di Jawa Timur," *J. Gaussian*, vol. 13, no. 2, pp. 373–382, 2024, doi: 10.14710/j.gauss.13.2.373-382.10.14710/j.gauss.13.2.373-382.

[18] M. Studi, K. P. Kec, and N. Kab, "Penerapan Fuzzy C-Means dalam Sistem Pendukung Keputusan untuk Penentuan Penerima Bantuan Langsung Masyarakat (BLM) PNPM- 2007 Pemerintah Indonesia mencanangkan Program Nasional Pemberdayaan Masyarakat," pp. 264–273.

[19] B. Destia and M. D. Kartikasari, "COMPARISON OF FUZZY C-MEANS AND FUZZY GUSTAFSON-KESSEL CLUSTERING METHODS IN PROVINCIAL GROUPING IN INDONESIA BASED ON CRIMINALITY-RELATED FACTORS," vol. 17, no. 2, pp. 1093–1102, 2023.

[20] M. Fajri, Rais, and L. Handayani, "Regions Grouping in Central Sulawesi Province By Transmitted Disease Using Fuzzy Gustafson Kessel," *Barekeng*, vol. 17, no. 1, pp. 275–284, 2023, doi: 10.30598/barekengvol17iss1pp0275-0284.10.30598/barekengvol17iss1pp0275-0284.

[21] K. Nida, M. N. Hayati, and R. Goejantoro, "Implementasi Metode Fuzzy Possibilistic C-Means pada Pengelompokan Provinsi di Indonesia Berdasarkan Data Jumlah Kejadian dan Dampak Bencana Banjir," *J. Math. Educ. Sci.*, vol. 7, no. 1, pp. 33–42, 2024, doi: 10.32665/james.v7i1.1919.10.32665/james.v7i1.1919.

[22] P. Saeipour, P. Sarbakhsh, S. Salemi, and F. B. Aghdam, "A Fuzzy Clustering Approach to Identify Pedestrians ' Traffic Behavior Patterns," vol. 23, no. 3, 2023, doi: 10.34172/jrhs.2023.127.10.34172/jrhs.2023.127.

[23] rahmania azwarini, "Fuzzy Gustafson Kessel Clustering," 2025. [Online]. Available: https://exsight.id/blog/2025/03/19/fuzzy-gustafson-kessel-clustering/

[24] T. Kyriazos and M. Poga, "Dealing with Multicollinearity in Factor Analysis : The Problem , Detections , and Solutions," pp. 404–424, 2023, doi: 10.4236/ojs.2023.133020.10.4236/ojs.2023.133020.

[25] S. K. Hati and V. Aryati, "PENELITIAN MANAJEMEN SUMBER DAYA MANUSIA," vol. 1, pp. 94–102, 2022.

[26] P. Kabupaten, K. Di, P. Jawa, B. Jumlah, and P. Peternakan, "Article Info:," vol. 11, pp. 366–376, 2023, doi: 10.14710/j.gauss.11.3.366-376.10.14710/j.gauss.11.3.366-376.