Jurnal Varian
Vol. 6, No. 1, Oktober 2022, Hal. 71–80
e-ISSN: 2581-2017

Application of Soft-Clustering Analysis Using Expectation Maximization Algorithms on Gaussian Mixture Model

Andi Shahifah Muthahharah[1], Muhammad Arif Tiro[2], Aswi Aswi[3]
[1,2,3]Statistics Study Program, Universitas Negeri Makassar, Indonesia

Article Info

Article history:

Received : 07-17-2022
Revised : 10-23-2022
Accepted : 11-13-2022

Keywords:

Sentiment Analysis;
GMM;
Mixed Model;
Said Index;
Water Quality.

ABSTRACT

Research on soft-clustering has not been explored much compared to hard-clustering. Soft-clustering algorithms are important in solving complex clustering problems. One of the soft-clustering methods is the Gaussian Mixture Model (GMM). This study aims to determine the number of clusters formed by using the GMM method. The data used in this study is synthetic data on water quality indicators obtained from the Kaggle website. The stages of the GMM method are: imputing the Not Available (NA) value, checking the data distribution, conducting a normality test, standardizing the data and estimating the parameters with the Expectation Maximization (EM) algorithm. The best number of clusters is based on the biggest value of the Bayesian Information Creation (BIC). The results showed that the best number of clusters was 3 clusters. Cluster 1 consisted of 1110 observations with low-quality category, cluster 2 consisted of 499 observations with medium quality category, and cluster 3 consisted of 1667 observations with high-quality category. The results of this study recommend that the GMM method can be grouped correctly when the variables used are generally normally distributed. This method can be applied to real data, both in which the variables are normally distributed or mixture of Gaussian and non-Gaussian.

Accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: https://doi.org/10.30812/varian.v6i1.2142

Corresponding Author:

Aswi Aswi
Statistics Study Program, Universitas Negeri Makassar
Email: aswi@unm.ac.id

A. INTRODUCTION

Era Society 5.0 emphasizes that humans are the center of balance in economic progress and the resolution of social problems by a system that integrates the virtual world and the real world. One of the technological advances that have been developed since the industrial revolution era to the Society 5.0 era is machine learning. Machine learning is a technology that allows computers to run to solve problems on their own without having to be programmed explicitly (Samuel, 2000). Machine learning is used to make computers perform sophisticated tasks without human intervention on the basis of learning and continuously improving experience to understand the complexity of problems and the need for adaptability (Alzubi et al., 2018). In the world of machine learning, the terms supervised learning and unsupervised learning are commonly used.

The unsupervised learning technique discussed in this study is clustering. Clustering is a technique of grouping data into several clusters so that the data in one cluster has a maximum similarity level and the data between clusters has a minimum similarity (Tan et al., 2006). Clustering techniques are divided into two, namely hard-clustering and soft-clustering. In soft-clustering, data points can be grouped in more than one cluster. An example of soft-clustering is the Gaussian Mixture Model (GMM) algorithm. GMM is a probabilistic model that assumes all data points are generated from a mixture of some Gaussian distributions with unknown parameters.

A comparison of the model-based method with the K-Mean method in cluster analysis has been investigated using both simulation data and Iris secondary data (Pardede, 2007). The results showed that the model-based method was more effective in separating

overlapping clusters than the K-Mean method (Pardede, 2007). One of the model-based methods is GMM. Another research on GMM has also been conducted (Yeung et al., 2001). Their results concluded that data transformation could result in normal data. Although the original data did not fully meet the Gaussian mixed assumption even after transformation, model-based clustering still resulted in higher quality clusters and suggested a better number of clusters (Yeung et al., 2001). In addition, (Kassambara, 2017) in his book also gives an example that in model-based clustering such as GMM, data standardization is first carried out before carrying out the GMM analysis. GMM, one of the soft-clustering methods, has not been widely explored. This study aims to determine the best number of clusters formed from synthetic data on Water Quality Indicators using the GMM method.

B. LITERATURE REVIEW

1. Cluster Analysis

Cluster analysis is defined as an algorithm used to group data points, where data points that have similar characteristics are placed into one group. Clustering analysis is divided into two, namely hard-clustering and soft-clustering. In hard-clustering, each data point has a clear boundary in its grouping, whether it belongs to a cluster or not. An example of hard-clustering is the K-mean algorithm. Whereas in soft-clustering, data points can be grouped in more than one cluster. An example of soft-clustering is the GMM algorithm.

2. Gaussian Mixture Model (GMM)

GMM is one of the soft-clustering methods in machine learning that uses a continuous probability distribution. GMM is a parametric probability density function represented as the sum of the weights of the Gaussian component density (Reynolds, 2009). GMM involves a mixture of several Gaussian distributions. GMM is used to classify data points into different clusters based on their probability distribution. GMM consists of two parts, namely the mean vector (μ) & the variance/covariance matrix $(\sum)$. GMM is useful in situations where the cluster has an "elliptical" shape.

The GMM equation, which is the weighted sum of M components of the Gaussian density, is written as follows (Mohammed et al., 2016).

$$P(x|\theta) = \sum_{k=1}^{M} w_k p(x|\theta_k) \tag{1}$$

where

θ_k	: the mean and covariance of k-th components
w_k	: the weight of k-th component
$p(x\theta_k)$	: Gaussian density.

3. Parameter Estimation with the Expectation Maximization (EM) Algorithm

The Expectation Maximization (EM) algorithm was first introduced in a classic magazine in 1977 by (Dempster et al., 1977) in the Journal of The Royal Statistical Society. The use of the EM algorithm in the Gaussian mixed model is to maximize the likelihood function related to the parameters consisting of the mean and covariance of the components and coefficients of the mixture (Bishop and Nasrabadi, 2006).

The EM steps are described as follows (Bishop and Nasrabadi, 2006).

1. Initializes the mean k, covariance k, and coefficient of mix k, and evaluates the initial log-likelihood value.
2. E-step. Evaluates the value using the current parameter.

$$\gamma(Z_{nk}) = \frac{\pi_k N\left(X_n|\mu_k, \sum_k\right)}{\sum_{j=1}^{K} \pi_j N\left(X_n|\mu_j, \sum_j\right)} \tag{2}$$

where:

DOI: https://doi.org/10.30812/varian.v6i1.2142

$\gamma(Z_{nk})$: estimated value of E-step
π : sample weight
N : the number of observations
x_n : n-th data
μ : sample mean
$\sum$: covariance matrix

3. M-step. Re-estimate the parameter using the current value.

$$\mu_k^{new} = \frac{1}{N_K}\sum_{n=1}^{N}\gamma(z_{nk})x_n \tag{3}$$

$$\sum_k^{new} = \frac{1}{N_K}\sum_{n=1}^{N}\gamma(z_{nk})(x_n-\mu_k^{new})(x_n-\mu_k^{new})^T \tag{4}$$

$$\pi_k^{new} = \frac{N_k}{N} \tag{5}$$

$$N_k = \sum_{n=1}^{N}\gamma(z_{nk}) \tag{6}$$

where:
$\gamma(Z_{nk})$: estimated value
π : sample weight
N : the number of observations
x_n : n-th data
μ : sample mean
$\sum$: covariance matrix

4. Evaluating the log-likelihood value

$$\ln p(X|\mu,\sum,\pi) = \sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k N|\left(x_n|\mu_k,\sum_k\right)\right\} \tag{7}$$

where:
π : sample weight k
N : the number of observations
x_n : n-th data
μ : sample mean k
$\sum$: covariance matrix k

The next step is to check convergence through the calculation of the log-likelihood value at the end of each EM step. The EM has converged if the log-likelihood value at the end of each EM step does not change significantly. If the convergence criteria are not met, then return to step 2.

4. Determination of the Best Number of Clusters with Bayesian Information Creation (BIC)

Bayesian Information Creation (BIC) is used to select the best model. The BIC formula is given in the following equation.

$$BIC = 2\ln f(X|\underline{\hat{\theta}}) + s\ln N \tag{8}$$

where:
$\ln f(X|\underline{\hat{\theta}})$: the maximum value of the log-likelihood function of the estimated model
N : number of observations
s : number of parameters

By using the mclust package, it is concluded that the greater the BIC value, the stronger the evidence for the goodness of the model and the number of clusters (Fraley and Raftery, 2002). The mclust package does not involve negative components.

DOI: https://doi.org/10.30812/varian.v6i1.2142

C. RESEARCH METHOD

1. Data Source

The data used in this study is secondary data, namely synthetic data on Water Quality Indicators obtained from the Kaggle website and can be accessed via the link https://www.kaggle.com/adityakadiwal/water-potability. The data used consisted of 3276 samples from 10 variables, namely pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity, and Potability. The scale used in the data is a numerical scale on the variables pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity, and a binary scale on the variables Potability. The Missingness Map of synthetic data on Water Quality Indicators can be seen in Figure 1.

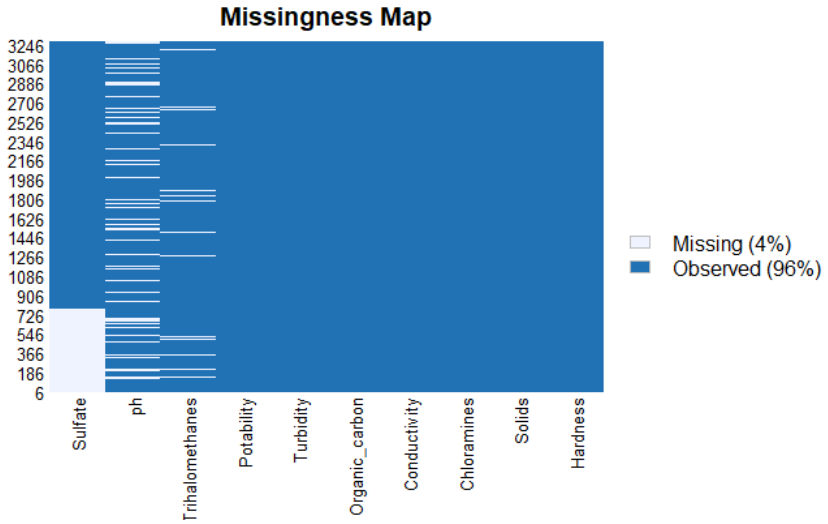

Figure 1. Missingness Map of Synthetic Water Quality Indicator

2. Research Procedure

The research procedure carried out is given in Figure 2.

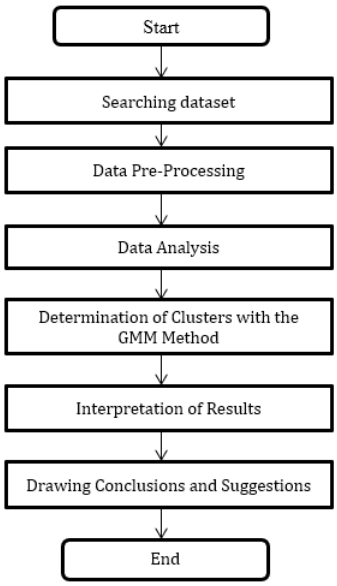

Figure 2. Research Procedure

DOI: https://doi.org/10.30812/varian.v6i1.2142

3. Data Analysis Technique

The steps taken in this research are as follows.

1. Perform pre-processing data which includes imputed NA values and descriptive analysis of Synthetic Water Quality Indicator variables.
2. Checking the distribution of data for each variable through the Cullen and Frey graph.
3. Test the normality assumption through the Shapiro-Wilk test.
4. Standardize data.
5. Estimating parameters using the EM Algorithm using the mclust library in R software. The steps of the EM algorithm briefly can be seen as follows (Bishop and Nasrabadi, 2006).
   (a) Initializes the mean k, covariance k, and coefficient of mix k, and evaluates the initial log likelihood value.
   (b) E-step. Evaluates values using existing parameters.
   (c) M-step. Re-estimate parameters using existing values.
   (d) Evaluate the log-likelihood value. The next step is to check the convergence of the parameters or log-likelihood values. If the convergence criteria are not met, then return to step b (E-step).
6. Determine the best number of clusters with Bayesian Information Creation (BIC).
7. Display grouping visualization with GMM.
8. Draw conclusions.

D. RESULTS AND DISCUSSION

1. NA Value Imputation

The imputation of the NA value is carried out when the researcher finds the NA value in the data. This causes the accuracy of the data to be less than optimal or less good. Therefore, an imputation was carried out to handle the NA value using the Amelia package on R. Amelia's package uses the multiple imputation method. The multiple imputation method is one method for estimating the value of the missing observations. The possibility of obtaining the correct estimate is greater than just imputing once, because the missing observed values is imputed several times (Utami and Danardono, 2019).

Amelia package combines the bootstrap method and the Expectation Maximization algorithm to perform data imputation. As postulated by (Efron, 1992), the bootstrap distribution is obtained by replacing the unknown distribution with the empirical distribution of the data in the statistical function, and then resampling the data to obtain the Monte Carlo distribution which generates random variables, with the theoretical approximation of the bootstrap method for dynamic dimensions can be seen in (Tiro, 1991). The data is known to contain the NA value in the pH variable at about 491, the Sulfate variable at about 781, and the Trihalomethanes variable at about 162.

2. Descriptive Analysis

The descriptive analysis of the complete data characteristics used for each variable is given in Table 1.

Table 1. Descriptive Analysis of Synthetic Water Quality Indicators

Variable	Minimum Value	Median	Maximum Value	Mean
pH	0	7.05	14	7.08
Hardness (mg/L)	47.43	196.97	323.12	196.37
Solids (mg/L)	320.9	20927.8	61227.2	22014.1
Chloramines (mg/L)	0.35	7.13	13.12	7.12
Sulfate (mg/L)	129.0	333.9	481	334.4
Conductivity (μS/cm)	181.5	421.9	753.3	426.2
Organic Carbon (mg/L)	2.20	14.22	28.30	14.28
Trihalomethanes (mg/L)	0.73	66.66	124	66.53
Turbidity (NTU)	1.45	3.95	6.73	3.96

3. Data Distribution with Cullen and Frey Graphs

By checking the distribution of data with the Cullen and Frey graph, it was found that the pH, Hardness, Chloramines, and Sulfate variables were between the normal distribution and the logistic distribution. The Solids variable was between the

DOI: https://doi.org/10.30812/varian.v6i1.2142

lognormal distribution, the gamma distribution, and the beta distribution. The Conductivity variable was in between the normal distribution and the beta distribution. Meanwhile, three variables are in a normal distribution, namely Organic Carbon, Trihalomethanes, and Turbidity. Therefore, it can be concluded that the three variables meet the assumption of normality. One of the results of the normality test of the Organic Carbon variable using the Cullen and Frey graph can be seen in Figure 3.

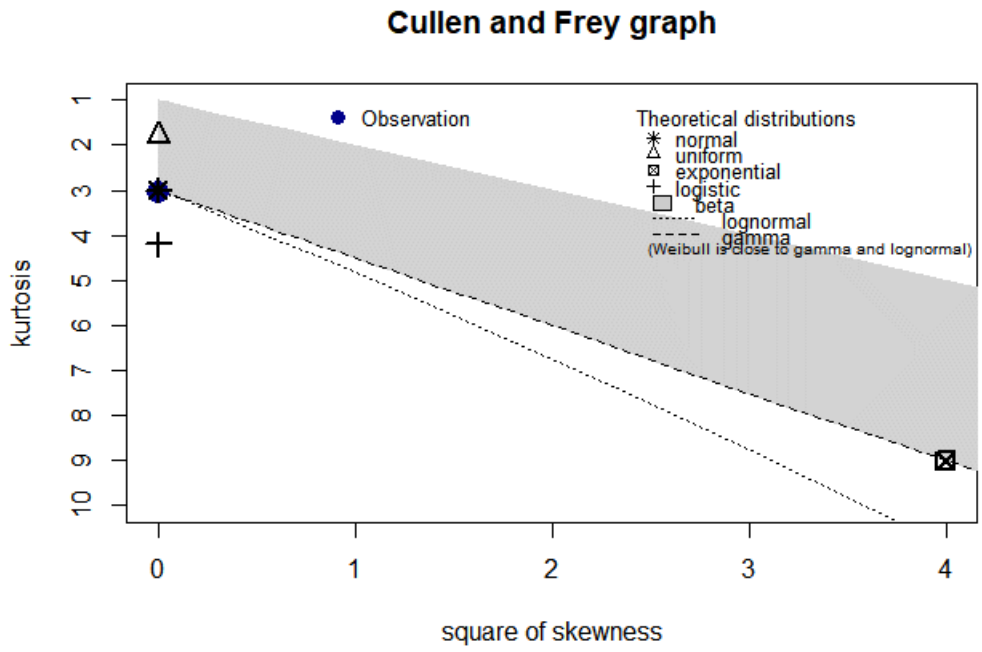

Figure 3. Cullen and Frey Chart of Organic Carbon Variables

The normality test on water quality indicator data variables using the Shapiro-Wilk test is presented in Table 2.

Table 2. Shapiro-Wilk test on Synthetic Water Quality Indicator Variables

Variable	p-value
pH	< 0.001
Hardness (mg/L)	< 0.001
Solids (mg/L)	< 0.001
Chloramines (mg/L)	< 0.001
Sulfate (mg/L)	< 0.001
Conductivity (μS/cm)	< 0.001
Organic Carbon (mg/L)	0.62
Trihalomethanes (mg/L)	0.08
Turbidity (NTU)	0.93

Based on Table 2, it can be seen that the six variables of pH, Hardness, Solids, Chloramines, Sulfate, and Conductivity do not meet the assumption of normality. This can be seen from the p-value < 0.05. Furthermore, the p-values of the Organic Carbon, Trihalomethanes, and Turbidity variables were 0.62, 0.08, and 0.93 respectively, which were greater than 0.05. Therefore, it can be concluded that the three variables meet the assumption of normality. Although there is a mixture of Gaussian and non-Gaussian, GMM is a flexible method for modeling different multidimensional distributions of data (VanderPlas, 2016).

Furthermore, the data was transformed so that the data had a mean of 0 and a variance of 1. Three types of transformations commonly used, namely logarithmic transformation, square root transformation, and data standardization were carried out (Yeung et al., 2001). They concluded that data transformation could increase normality in the data.

4. Parameter Estimation with Expectation Maximization (EM) Algorithm

The parameter estimation results using the EM Algorithm are shown in Table 3. Furthermore, the graph for selecting the best model related to the number of clusters can be seen in Figure 4.

Table 3. Parameter Estimation Results with EM Algorithm

Log-likelihood	n	df	BIC	ICL
4947.36	3276	56	9441.43	7081.71

Based on Figure 4, it is known that the number of the best clusters for water quality indicator data is 3 clusters. The selection of the best model can be seen at point VVI (light purple) with the description below.

1. The number of clusters $n = 1$ has a BIC value that is around 9100
2. The number of clusters $n = 2$ has a BIC value that is around 9300
3. The number of clusters $n = 3$ has a BIC value that is around 9400

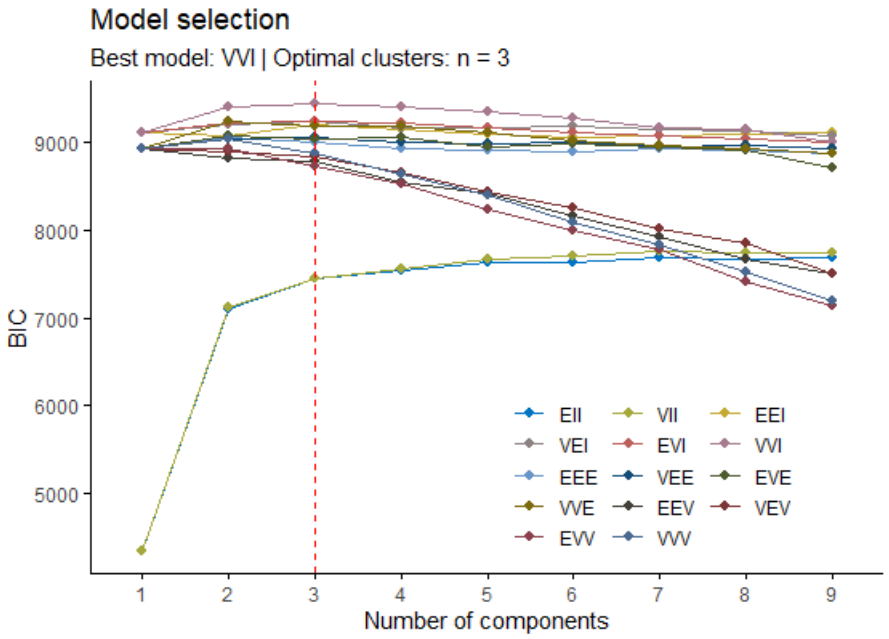

Figure 4. Selection of the Number of Water Quality Indicator Clusters

Based on the mclust package, in general, the best model is obtained from the highest BIC value (Androniceanu et al., 2020) (Fraley and Raftery, 2007) (Pardede, 2013). In addition, (Fraley et al., 2012) also stated that the selection of the number of clusters can be seen based on the adjusted vertical axis to display the maximum value. Based on Table 3 and Figure 4, it can be concluded that the best model is the model with the optimal number of clusters $n = 3$ with BIC value = 9441.43.

The mclust package in R has several types of models for modeling GMM which are presented in Table 4 below (Scrucca et al., 2016).

Table 4. Types of Models in Package mclust in R

Model Type	Model Name	Information
Univariate Mixture	E	Equal variance (one dimension)
	V	Unequal variance (one dimension)
Multivariate Mixture	EII	Round, same volume
	VII	Round, the volume is not the same
	EEI	Diagonal, volumes and shapes are the same
	VEI	Diagonal, varying volume, same shape
	EVI	Diagonal, same volume, varied shape
	VVI	Diagonal, volume and shapes vary
	EEE	Ellipsoidal, same volume, shape, and orientation
	VEE	Ellipsoidal, same and shape orientation
	EVE	Ellipsoidal, same and volume orientation
	VVE	Ellipsoidal, same orientation
	EEV	Ellipsoidal, same volume, same shape
	VEV	Ellipsoidal, same shape
	EVV	Ellipsoidal, equal volume
	VVV	Ellipsoidal, volume, shape, and orientation varies

5. Visualization of Clustering with Gaussian Mixture Model (GMM)

Figure 5 shows the distribution plot of water quality clusters with 3 clusters.

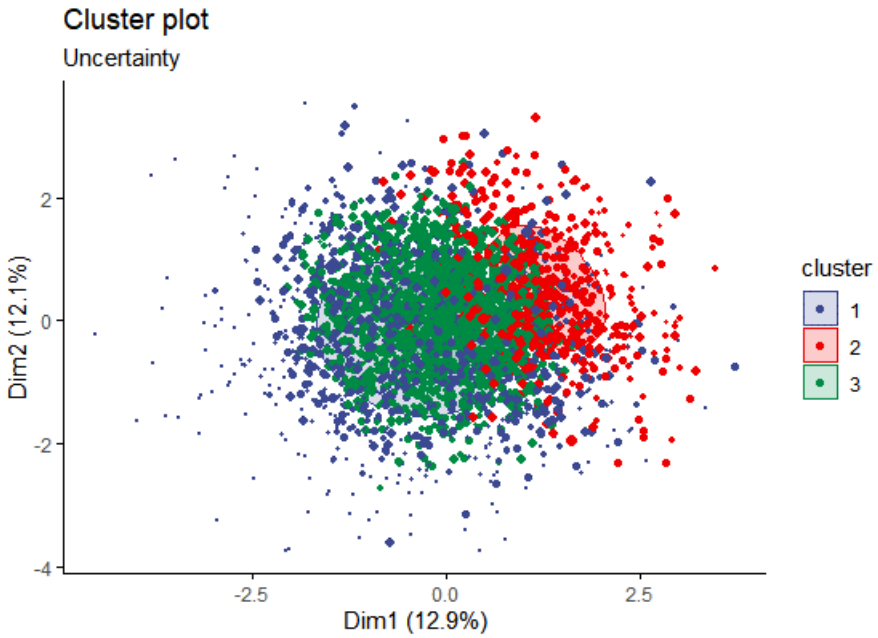

Figure 5. A plot of the Water Quality Cluster Distribution

Clusters are formed based on the observed values. Based on Figure 5, it can be seen that the blue observation value is cluster 1, the red observation value is cluster 2, and the green observation value is cluster 3.

In addition, in Figure 6, the distribution of water quality groupings based on their observed values using GMM is also presented.

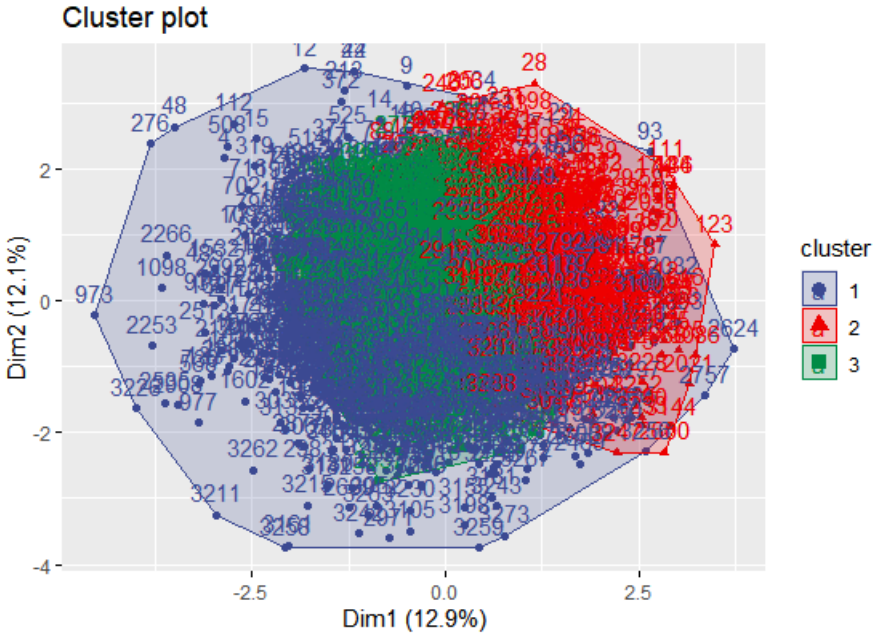

Figure 6. A Plot of Cluster Distribution of Water Quality Grouping Based on Observation Values

E. CONCLUSION AND SUGGESTION

In general, it can be concluded that the best number of clusters is 3 clusters. The parameter estimation results using the Expectation Maximization (EM) algorithm state that the maximum log-likelihood value is 4947.36 with the best number of clusters being 3 clusters. This can be seen based on the largest BIC value, which is 9441.43. With the GMM method, it is known that cluster 1 consists of 1110 observations with poor water quality clusters, cluster 2 consists of 499 observations with moderate quality clusters, and cluster 3 consists of 1667 observations with the best water authenticity cluster. A more precise grouping can be produced by the GMM method if the variables used are generally normally distributed. This method is flexible and can be applied to real data, both in which the variables are normally distributed and which have a mixture of Gaussian and non-Gaussian. Further research is recommended to use more complex data, such as data in the form of images.

REFERENCES

Alzubi, J., Nayyar, A., and Kumar, A. (2018). Machine learning from theory to algorithms: an overview. In *Journal of physics: conference series*, volume 1142, page 012012. IOP Publishing.

Androniceanu, A., Kinnunen, J., and Georgescu, I. (2020). E-government clusters in the eu based on the gaussian mixture models. *Administratie si Management Public*, (35):6–20.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer.

Fraley, C. and Raftery, A. (2007). Model-based methods of classification: using the mclust software in chemometrics. *Journal of Statistical Software*, 18:1–13.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631.

Fraley, C., Raftery, A. E., Murphy, T. B., and Scrucca, L. (2012). mclust version 4 for r: normal mixture modeling for model-based clustering, classification, and density estimation. Technical report, Technical report.

Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning*, volume 1. Sthda.

Mohammed, M., Khan, M. B., and Bashier, E. B. M. (2016). *Machine learning: algorithms and applications*. Crc Press.

Pardede, T. (2007). Perbandingan metode model-based dengan metode k-mean dalam analisis cluster. *Jurnal Matematika Sains dan Teknologi*, 8(2):98–108.

Pardede, T. (2013). Kajian metode berbasis model pada analisis kelompok dengan perangkat lunak mclust. *Jurnal Matematika Sains dan Teknologi*, 14(2):84–100.

Reynolds, D. (2009). Gaussian mixture models.

Samuel, A. L. (2000). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 44(1.2):206–226.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *The R journal*, 8(1):289.

Tan, P., Steinbach, M., and Kumar, V. (2006). Instructors solution manual. *Introduction to Data Mining*.

Tiro, M. A. (1991). *Edgeworth expansion and bootstrap approximation for M-estimators of linear regression parameters with increasing dimensions*. Iowa State University.

Utami, R. S. and Danardono, D. (2019). Metode multiple imputation untuk mengatasi kovariat tak lengkap pada data kejadian berulang. *Journal of Fundamental Mathematics and Applications (JFMA)*, 2(2):47–57.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987.