Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 6 No. 5 (2022) 866 - 871 ISSN Media Electronic: 2580-0760

House Prices Segmentation Using Gaussian Mixture Model-Based Clustering

Muhammad Hafidh Raditya[1], Indwiarti[2], Aniq Atiqi Rohmawati[3]
[1,2,3]Informatics, School of Computing, Telkom University
[1]mhafidhraditya@student.telkomuniversity.ac.id, [2]indwiarti@telkomuniversity.ac.id, [3]aniqatiqi@telkomuniversity.ac.id

Abstract

House is a place for humans to live and a main necessity for humans. For years, the need for houses is increasing and varied so that it affects the selling price of the house. Therefore, more research is needed to learn about the selling price of houses. This research is only focusing on house price segmentation in DKI Jakarta using the Gaussian Mixture Model-Based Clustering Method with the Expectation-Maximization algorithm. The goal of this research is to make a house price segmentation model so that we can obtain useful information for the potential buyer. Clustering with GMM utilize the log-likelihood function to optimize the GMM parameters. The result of this research is houses in DKI Jakarta can be segmented into 3 different clusters. The first cluster is for the low-profile houses. The second cluster is for the mid-profile houses. The third cluster is for the high-profile houses. The silhouette score that was produced by the clustering method is 0.60866 meaning that this score is quite good because it's close to a value of 1.

Keywords: *segmentation, clustering, Gaussian Mixture Model*

1. Introduction

House is one of the main needs for human and it's very important. House is the place for human to seek for protection from the outside world. As time goes by, human needs for house is getting higher, and the house prices are getting varied based on the house parameter like area of the house, house location, numbers of room, etc.

Because of the varied house prices, further research to analyze the correlation between house prices and the other house parameter needs to be done. The results of the analysis are expected to be able to group into different categories so people that want to buy houses could have enough information in buying a house that match their budget and criteria. There is already previous research that did a house prices segmentation using hedonic regression [1].

Clustering is one of the most important unsupervised learning processes in machine learning [2]. Clustering is a process to segment a group of similar data into a same cluster, but different if they are compared to another data that is part of another cluster [3]. Clustering is often used for data segmentation process, so clustering is a good choice for the approach of this research.

Hierarchical clustering is the first clustering method that was used by the biologists and social scientists. At that time cluster analysis is already a branch of multivariate statistics analysis [4]. But in this research, the method that will be used is the gaussian mixture model-based clustering. Gaussian mixture model (GMM) is a simple model that can be used for classification or clustering compared to another method [5]. GMM is a probability density function that is represented by a group of gaussian function component [6]. In GMM-based clustering, each cluster are represented by gaussian distribution or normal distribution with three parameters like *mean* (μ), covariance (σ), and weight (π) [7]GMM can be modelled by using formula 1.

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \sigma_k) \quad (1)$$

$p(x)$ is the probability density function, K is the number of clusters, $\mathcal{N}$ is the number of observations, μ_k is the mean of cluster k, σ_k is the covariance of cluster k, and π_k is the weight of cluster k.

On these days there are so much research that already used GMM as the method of the clustering. For example, there is research that used GMM-based clustering to identify daily electricity usage profiles of a building [8]. The result of the research is GMM-based clustering can provide useful information of the pattern

Accepted: 29-08-2022 | Received in revised: 15-10-2022 | Published: 31-10-2022

of daily electricity usage profiles. Furthermore, the research also stated that GMM-based clustering performs better than the hierarchical clustering based on the silhouette score. The comparison of the GMM-based clustering and the hierarchical clustering for this problem can be seen on Table 1.

Table 1. GMM-based clustering vs hierarchical clustering comparison for identifying daily electricity usage profile

Clustering method	Optimal cluster(s)	Silhouette score
GMM	2	0.719
Hierarchical clustering	2	0.696

There is another research that compared GMM-based clustering and K-Means clustering using the Google Cluster Trace dataset [9]. The result is the number of optimal clusters of GMM-based clustering is way more than the number of optimal clusters on K-Means clustering, resulting in more detailed information in GMM-based clustering result. Although the GMM-based clustering's computation time is much higher than the K-means clustering's computation time. But this thing is understandable considering the differences of the number of the optimal clusters between the two methods. The result of the research can be seen on Table 2.

Table 2. GMM-based clustering vs K-Means clustering comparison using the Google Cluster Trace dataset

Clustering method	Optimal cluster(s)	Computation time (ms)
GMM	12	312.03
K-Means	2	46.86

The are two goals of this research: (1). Identify how to optimize the parameters of GMM, (2). Identify the result of the house prices segmentation using GMM-based clustering then analyze the results

2. Research Methods

2.1 Data Collection

The dataset that is used in this research is a group of house prices data that are collected from www.olx.co.id website. The data is collected using the free web scraping tool that are available on Google Chrome as an extension. This tool utilizes the HTML elements, that each element is unique based on each attribute of the house prices. The dataset was obtained in January 2022. The attributes that are available on this dataset can be seen in Table 3.

Table 3. Dataset attributes

Attributes	Description
Land area	Land area in m^2
Building area	Building area in m^2
Bedrooms	Number of bedrooms
Bathrooms	Number of bathrooms
Floors	Number of floors
Location	Location of the house. Consist of province, city, and district
Price	The price of the house

The dataset can be downloaded by clicking on the URL https://bit.ly/3fh4uCD.

2.2 Data Exploration and preprocessing

This step will do the data exploration, which is a process required to understand the data before the preprocessing can be done. Data preprocessing is a process to prepare and transform the raw data to a form that is can be used in the modelling process [10]. There are several steps in the data preprocessing part. **Error! Reference source not found.** is the shape of the raw dataset.

Land area	Building area	Bedrooms	Bathrooms	Floors	Location	Price
60 m2	60 m2	3	2	2.0	Cempaka Putih, Jakarta Pusat, Jakarta D.K.I.	Rp 550.000.000
87 m2	85 m2	3	2	2.0	Sawah Besar, Jakarta Pusat, Jakarta D.K.I.	Rp 795.000.007
79 m2	60 m2	2	2	2.0	Cempaka Putih, Jakarta Pusat, Jakarta D.K.I.	Rp 589.300.023
144 m2	41 m2	2	1	1.0	Cempaka Putih, Jakarta Pusat, Jakarta D.K.I.	Rp 495.000.000
90 m2	60 m2	4	2	1.0	Kelapa Gading, Jakarta Utara, Jakarta D.K.I.	Rp 2.100.000.000
…	…	…	…	…	…	…

Figure 1. Raw dataset

The raw dataset has 2098 line of data. The first step that need to be done is renaming all of the attribute's name. after that the data type of each attribute needs to be uniformed. The numerical attributes will be changed to float so the data processing will be easier, and the categorical data will be changed to string. For example, the price attribute is a string type, so it needs to be changed to float. Furthermore, the location attribute will be split into three different attributes that are district, city, and province. Figure 2 is the result of the preprocessed dataset.

Land area	Building area	Bedrooms	Bathrooms	Floors	Price	District	City	Province
60.0	60.0	3.0	2.0	2.0	550000000	Cempaka Putih	Jakarta Pusat	Jakarta D.K.I.
87.0	85.0	3.0	2.0	2.0	795000007	Sawah Besar	Jakarta Pusat	Jakarta D.K.I.
79.0	60.0	2.0	2.0	2.0	589300023	Cempaka Putih	Jakarta Pusat	Jakarta D.K.I.
144.0	41.0	2.0	1.0	1.0	495000000	Cempaka Putih	Jakarta Pusat	Jakarta D.K.I.
90.0	60.0	4.0	2.0	1.0	2100000000	Kelapa gading	Jakarta Utara	Jakarta D.K.I.
…	…	…	…	…	…	…	…	…

Figure 2. Preprocessed dataset

After the raw dataset are transformed, the next step is to handle the missing values. The only attribute that has missing values is floors attribute. By assuming that there is no house that has zero floor, then the missing values will be changed to the value of 1.

The final step in the preprocessing part is the outlier handling. The attributes that will be checked for the outliers are price, land area, and building area. The

DOI: https://doi.org/10.29207/resti.v6i5.4459

outlier detection is utilizing the interquartile range concept [11]. All the data that are considered outliers will be dropped.

After the preprocessing part is done, the amount of data that was initially 2098 lines, now the data are only consist of 1548 lines.

2.3. Clustering Process

The preprocessed data will be clustered based on its price attribute. First thing to do is to identify the number of optimal clusters for the data using the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC). Both are useful methods for model selection and on GMM-based clustering, both methods are very useful to identify the number of optimal clusters [12], [13]. BIC is utilizing the log-likelihood of the data [12], meanwhile AIC is utilizing the sum square error of the data [13]. The lower the score of BIC and AIC, the better the quality of the clusters [12], [13]. Figure 3 is the result of BIC and AIC of the current data.

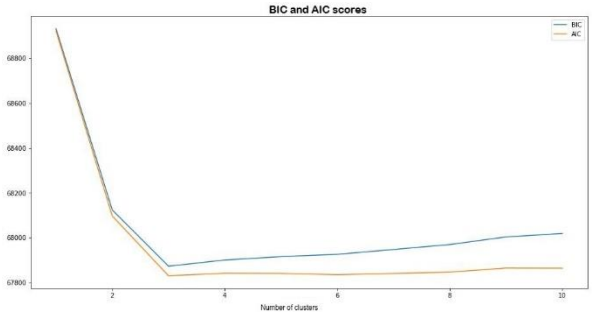

Figure 3. Graph of the identification process of finding the optimal number of clusters

From Figure 3 can be seen that both BIC and AIC produce the same answer of three optimal clusters. But there is a slight difference that is quite interesting. On the AIC graph that is colored in orange, the score of the AIC for the number of clusters above three looks constant kind of constant. It's like there is no difference of the AIC score for the number of clusters above three. But for the BIC graph that is colored in blue, there can be seen the increasing trend of the BIC scores for the number of clusters above 3. From this observation, it can be concluded that BIC is more precise of finding the optimal number of clusters rather than AIC.

After identifying the optimal number of clusters are three, the next step is for the GMM modelling. Expectation-maximization is an algorithm to ensure the convergence of a function [14]. The EM algorithm utilize the log-likelihood of the GMM function to optimize the three parameters of GMM. [14]. The process of EM will be repeated until the log-likelihood of the function is convergence, or in other words the value of the log-likelihood is constant or not changing anymore for each iteration [15]. Figure 4 and Figure 5 is the result of the clustering with three clusters and Figure 6 is the visualization of the density plot for each cluster.

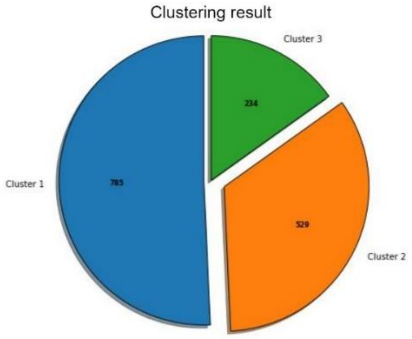

Figure 4. Pie chart of the clustering result

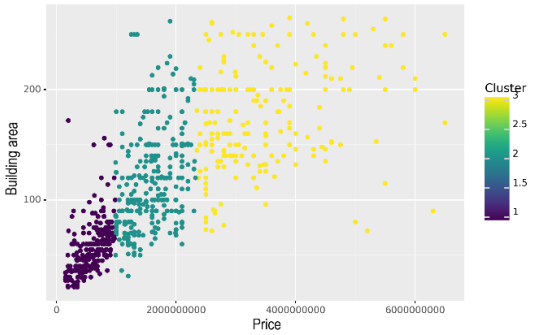

Figure 5. Scatter plot of the clustering result

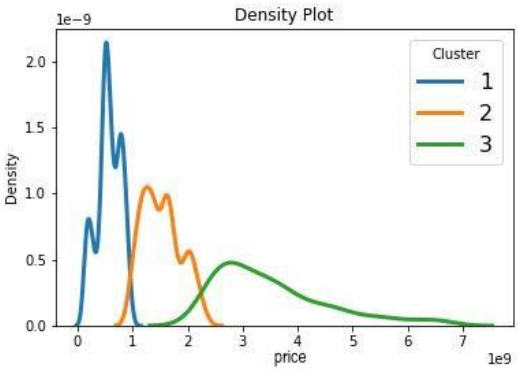

Figure 6. Density plot of each cluster

From Figure 6, cluster 1 is the densest cluster, followed by cluster 2, then cluster 3. Based on Figure 4, cluster 1 has 785 data, cluster 2 has 529 data, and cluster 3 has 234 data. If we look at the shape of the data distribution in Figure 5, cluster 1 and cluster 2 have the shape that tend to be ellipses, while cluster 3 has the shape that tends to be a circle. From this observation, it can be concluded that GMM-based clustering can produce the more flexible form of clusters. The clustering process is already converged on each cluster because the amount of maximum iteration that was set in this process is 100, but the process is already converged and stopped at the 18th iteration. The details of the three gaussian parameters for each cluster can be seen on Table 4.

DOI: https://doi.org/10.29207/resti.v6i5.4459

Table 4. Gaussian parameter values of each cluster

Parameter	Cluster 1	Cluster 2	Cluster 3
Weight	0.47367921	0.33842949	0.1878913
Covariance	5.35732443 x 10^{16}	2.16710375 x 10^{17}	1.70950349 x 10^{18}
K-Means	5.74828282 x 10^{8}	1.46268685 x 10^{9}	3.07239510 x 10^{9}

Clustering with these three clusters produce the silhouette score of 0.60866. silhouette is a method to test the quality of a cluster [16]. The closer the silhouette score to 1, the better the clustering result [17]. So, the score of 0.60866 is a very good score

Because this is a one-dimensional clustering, there need to be a comparation with at least two-dimensional clustering. First is the two-dimensional clustering between price and building area attributes. **Error! Reference source not found.** and **Error! Reference source not found.** are the result of the clustering.

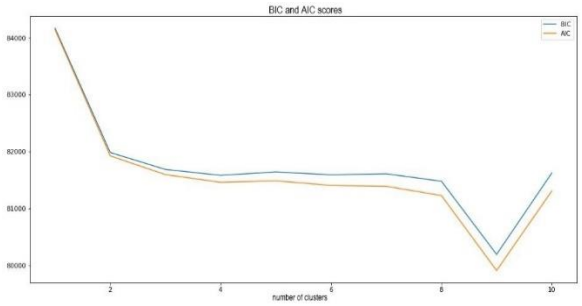

Figure 7. Number of optimal clusters for price - building area two-dimensional clustering

Figure 8. Scatter plot of price - building area two-dimensional clustering

Based on **Error! Reference source not found.**, this two-dimensional clustering using price and building area attributes have the optimal number of clusters of nine. But if we look at **Error! Reference source not found.**, there is one cluster (marked by the red circles) that only have 2 data. This situation is not effective because those 2 data could be included in another cluster that have the closest similarity to them. Furthermore, this clustering process only produce the silhouette score of 0.02074, a value that is closer to 0 rather than closer to 1. This means that the quality of the cluster is bad.

Next is the two-dimensional clustering using price and land area attributes. **Error! Reference source not found.** and **Error! Reference source not found.** are the result of the clustering

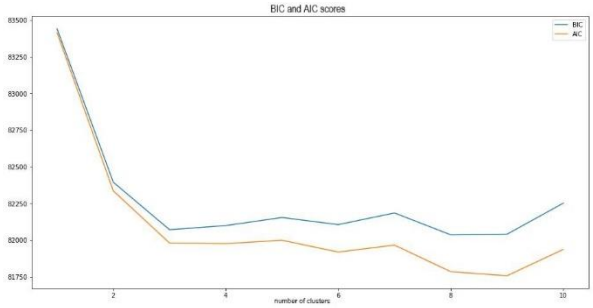

Figure 9. Number of optimal clusters for price - land area two-dimensional clustering

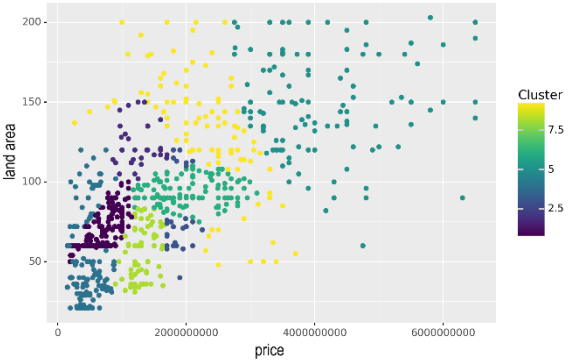

Figure 10. Scatter plot of price - land area two-dimensional clustering

Based on **Error! Reference source not found.**, the two-dimensional clustering using price and land area attributes have the optimal number of clusters of nine. Based on **Error! Reference source not found.**, this time the data distribution is good. There are no clusters that have very few data compared to the others. But the problem is still the same. The silhouette score for this clustering is only 0.056460, a value that is closer to 0 rather than closer to 1. This means that the quality of the cluster is bad.

So, the one-dimensional clustering has the better performance compared to the two-dimensional clustering. In the next section there will be a further evaluation of the one-dimensional clustering.

3. Results and Discussions

Before we evaluate the results, we need to analyze the correlation all attributes toward price attribute to decide which attributes to be evaluated. On this step, the Pearson correlation coefficient is used to analyze that [18]. The result of the correlation coefficient can be seen on **Error! Reference source not found.**.

Table 5. Pearson correlation coefficient towards price attribute

DOI: https://doi.org/10.29207/resti.v6i5.4459

Attributes	Correlation coefficient
Land area	0.689191
Building area	0.797858
Bedrooms	0.609563
Bathrooms	0.614404
Floors	0.329670

Land area and building area attributes have the highest correlation coefficient towards price attribute. Therefore, these two attributes will be selected to be evaluated in the next step. Beside these two, the district and city attributes will be selected too to see what the unique location on each cluster is. The goal of the analysis process is to get information from the clustering process. The numerical attributes will be visualized in histogram, while the categorical attributes will be visualized in bar plot. The analysis process will be carried out per cluster. Figure 11, Figure 12, and Figure 13 are the results of the clustering process for each cluster.

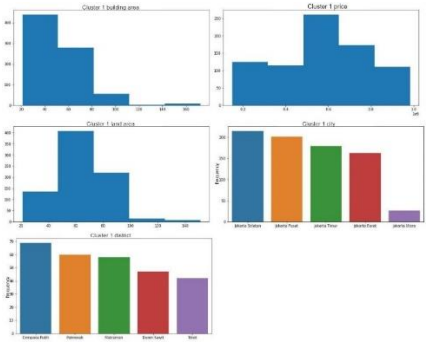

Figure 11. Clustering result for cluster 1

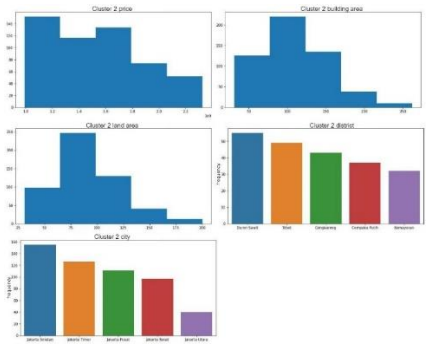

Figure 12. Clustering result for cluster 2

On cluster 1, the house prices range from 150,000,000 - 982,300,000 IDR. But if we look at the histogram, majority of the price on cluster 1 range from 450,000,000 – 650,000,000 IDR. The land area of cluster 1 houses ranges from 21 m^2 – 151 m^2. But if we look at the histogram, majority of the land area on cluster 1 range from 45 m^2 – 75 m^2. The building area of cluster 1 houses ranges from 21 m^2 – 172 m^2. But if we look at the histogram, majority of the building area on cluster 1 range from 21 m^2 – 50 m^2. The houses on cluster 1 mostly located on South Jakarta and Central Jakarta. If we look at the district, houses on cluster 1 mostly located in Cempaka Putih, Central Jakarta.

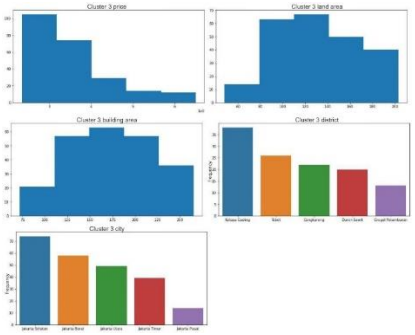

Figure 13. Clustering result for cluster 3

On cluster 2, the house prices range from 990,000,000 – 2,325,000,000 IDR. But if we look at the histogram, majority of the price on cluster 2 range from 990,000,000 – 1,800,000,000 IDR. The land area of cluster 2 houses ranges from 31 m^2 – 200 m^2. But if we look at the histogram, majority of the land area on cluster 2 range from 60 m^2 – 100 m^2. The building area of cluster 2 houses ranges from 31 m^2 – 262 m^2. But if we look at the histogram, majority of the building area on cluster 2 range from 75 m^2 – 125 m^2. The houses on cluster 2 mostly located on South Jakarta and East Jakarta. If we look at the district, houses on cluster 2 mostly located in Duren Sawit, East Jakarta.

On cluster 3, the house prices range from 2,350,000,000 – 6,500,000,000 IDR. But if we look at the histogram, majority of the price on cluster 3 range from 2,350,000,000 – 4,000,000,000 IDR. The land area of cluster 3 houses ranges from 48 m^2 – 203 m^2. But if we look at the histogram, majority of the land area on cluster 3 range from 80 m^2 – 140 m^2. The building area of cluster 3 houses ranges from 72 m^2 – 265 m^2. But if we look at the histogram, majority of the building area on cluster 3 range from 110 m^2 – 225 m^2. The houses on cluster 3 mostly located on South Jakarta and West Jakarta. If we look at the district, houses on cluster 3 mostly located in Kelapa Gading, North Jakarta.

The summary of the analysis explanation of Figure 11, Figure 12, and Figure 13, can be seen on Table 6.

Table 6. Clustering results analysis

Attributes	Cluster 1	Cluster 2	Cluster 3
Price (IDR)	450,000,000 – 650,000,000	990,000,000 – 1,800,000,000	2,350,000,000 – 4,000,000,000
Land area	45 m^2 – 75 m^2	60 m^2 – 100 m^2	80 m^2 – 140 m^2
Building area	21 m^2 – 50 m^2	75 m^2 – 125 m^2	110 m^2 – 225 m^2
District	Cempaka Putih, Central Jakarta	Duren Sawit, East Jakarta	Kelapa Gading, North Jakarta
City	South Jakarta and Central Jakarta	South Jakarta and East Jakarta	South Jakarta and West Jakarta

DOI: https://doi.org/10.29207/resti.v6i5.4459

4. Conclusion

In GMM-based clustering process, the GMM parameters were initialized randomly then optimized using the EM algorithm that utilizes the log-likelihood function. The process will be repeated until the log-likelihood function converged.

Silhouette is a score to define the quality of a clustering process. Silhouette score ranged from 0 to 1. The closer it gets to 1, the better the quality of the clustering. In this research, the number of mixture models or the number of clusters are three clusters with the silhouette score of 0.60866, which is a very good score considering that its close to 1. Cluster 1 is the cluster with low profile houses that the price mostly ranged on 450,000,000 – 650,000,000 IDR. The houses in cluster 1 mostly located in Cempaka Putih, Central Jakarta. Cluster 2 is the cluster with mid profile houses that the price mostly ranged on 990,000,000 – 1,800,000,000 IDR. The houses in cluster 2 mostly located in Duren Sawit, East Jakarta. Lastly, cluster 3 is the cluster with high profile houses that the price mostly ranged on 2,350,000,000 – 4,000,000,000 IDR. The houses in cluster 3 mostly located in Kelapa Gading, North Jakarta.

There are also some suggestions for the potential buyers. If the potential buyers have a low budget approximately 650,000,000 IDR at most, then it is recommended to find houses in Central Jakarta area. If the potential buyers have a high budget more than 2,000,000,000 IDR, then it is recommended to find houses in North Jakarta. But if the potential buyers haven't decided how many budgets that they got and haven't conduct an in-depth survey about houses in DKI Jakarta, then it is recommended for the buyers to find houses in South Jakarta because houses in South Jakarta have varied prices.

Lastly there are also some suggestions for the next research that may be conducted. First, because this research only used one-dimensional clustering, then the next research could be using two or more-dimensional clustering. And because this research only used the dataset from the DKI Jakarta area, then it is recommended to use a bigger dataset. For example, house prices dataset on Java Island.

Reference

[1] M. Yazdani, "House Price Determinants and Market Segmentation in Boulder, Colorado: A Hedonic Price Approach," Aug. 2021, https://doi.org/10.48550/arxiv.2108.02442.

[2] T. S. Madhulatha, "An Overview on Clustering Methods," May 2012, https://doi.org/10.48550/arxiv.1205.1117.

[3] A. Saxena *et al.*, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, Dec. 2017, https://doi.org/10.1016/J.NEUCOM.2017.06.053.

[4] M. S. Yang, C. Y. Lai, and C. Y. Lin, "A robust EM clustering algorithm for Gaussian mixture models," *Pattern Recognition*, vol. 45, no. 11, pp. 3950–3961, Nov. 2012, https://doi.org/10.1016/J.PATCOG.2012.04.031.

[5] H. Ling and K. Zhu, "Predicting Precipitation Events Using Gaussian Mixture Model," *Journal of Data Analysis and Information Processing*, vol. 05, no. 04, pp. 131–139, Oct. 2017, https://doi.org/10.4236/JDAIP.2017.54010.

[6] D. Reynolds, "Gaussian Mixture Models," *Encyclopedia of Biometrics*, pp. 827–832, 2015, https://doi.org/10.1007/978-1-4899-7488-4_196.

[7] X. He, D. Cai, Y. Shao, H. Bao, and J. Han, "Laplacian regularized Gaussian mixture model for data clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 9, pp. 1406–1418, 2011, https://doi.org/10.1109/TKDE.2010.259.

[8] K. Li, Z. Ma, D. Robinson, and J. Ma, "Identification of typical building daily electricity usage profiles using Gaussian mixture model-based clustering and hierarchical clustering," *Applied Energy*, vol. 231, pp. 331–342, Dec. 2018, https://doi.org/10.1016/J.APENERGY.2018.09.050.

[9] E. Patel and D. S. Kushwaha, "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model," *Procedia Computer Science*, vol. 171, pp. 158–167, Jan. 2020, https://doi.org/10.1016/J.PROCS.2020.04.017.

[10] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," *Journal of Engineering and Applied Sciences*, vol. 12, no. 16, pp. 4102–4107, Sep. 2017, https://doi.org/10.3923/JEASCI.2017.4102.4107.

[11] H. P. Vinutha, B. Poornima, and B. M. Sagar, "Detection of outliers using interquartile range technique from intrusion dataset," *Advances in Intelligent Systems and Computing*, vol. 701, pp. 511–518, 2018, https://doi.org/10.1007/978-981-10-7563-6_53/COVER.

[12] S. Watanabe SWATANAB, "A Widely Applicable Bayesian Information Criterion," *Journal of Machine Learning Research*, vol. 14, pp. 867–897, 2013, https://doi.org/10.5555/2567709.2502609.

[13] J. E. Cavanaugh and A. A. Neath, "The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 11, no. 3, p. e1460, May 2019, https://doi.org/10.1002/WICS.1460.

[14] N. Sammaknejad, Y. Zhao, and B. Huang, "A review of the Expectation Maximization algorithm in data-driven process identification," *Journal of Process Control*, vol. 73, pp. 123–136, Jan. 2019, https://doi.org/10.1016/J.JPROCONT.2018.12.010.

[15] S. F. Qonita, "Segmentasi Citra MRI Tumor Otak Menggunakan Gaussian Mixture Model dan Hybrid Gaussian Mixture Model - Spatially Variant Finite Mixture Model Dengan Algoritma Expectation-Maximization," *Institut Teknologi Sepuluh Nopember*, 2018.

[16] A. Aditya, I. Jovian, and B. N. Sari, "Implementasi K-Means Clustering Ujian Nasional Sekolah Menengah Pertama di Indonesia Tahun 2018/2019," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 4, no. 1, pp. 51–58, Jan. 2020, https://doi.org/10.30865/MIB.V4I1.1784.

[17] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, no. C, pp. 53–65, Nov. 1987, https://doi.org/10.1016/0377-0427(87)90125-7.

[18] J. Deng, Y. Deng, and K. H. Cheong, "Combining conflicting evidence based on Pearson correlation coefficient and weighted graph," *International Journal of Intelligent Systems*, vol. 36, no. 12, pp. 7443–7460, Dec. 2021, https://doi.org/10.1002/INT.22593.

DOI: https://doi.org/10.29207/resti.v6i5.4459