Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.13888
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Evaluation of Cluster Models for Creating Profiles of Home Buyers

Made Dhanita Listra Prashanti Dewi[1)*], Ito Wasito[2)]
[1)]Pradita University, Indonesia
[1)]made.dhanita@student.pradita.ac.id, [2)]ito.wasito@pradita.ac.id

Submitted : Jul 16, 2024 | **Accepted** : Aug 7, 2024 | **Published** : Oct 3, 2024

Abstract: The property industry in Indonesia is currently a dynamic and continuously evolving field, in line with rapid economic growth and urbanization. Shifts in lifestyle patterns, infrastructure development, and changes in government policies have had a significant impact on how properties are marketed in Indonesia. With a growing population and increasing purchasing power, the Indonesian property market is becoming more complex. Therefore, strategies are needed to segment consumer groups for effective marketing in the housing sector. This research will delve deeper into consumer segmentation in home selection, a technique that divides consumer diversity into distinct groups based on characteristics and behavior. By using an extensive dataset involving demographic data such as location, age, gender, occupation, and many other variables, clustering algorithms can uncover complex patterns to determine consumer segments in their home selection. The algorithms to be used for this study are K-Means clustering, the Gaussian Mixture model, and Hierarchical clustering. By using these three data clustering models, we can determine which algorithm produces the most ideal results for customer profiling. The results demonstrate that the K-Means algorithm outperforms the others in accurately identifying distinct consumer segments, hence producing customer profiles. Therefore, customer profiling can also be used by the marketing division as a tool to aid in promotions in order to better understand their target audience, hence creating a successful marketing campaign.

Keywords: Customer Profiling; Gaussian Mixture Model; Hierarchical Clustering; K-Means Clustering; Property;

INTRODUCTION

A growing property company has several projects running simultaneously in different locations. Consequently, the target consumers vary along with the house prices. Therefore, there is a need for clear "Customer Profiling" for each project to make marketing strategies more effective. A “customer profile” is a description of a business’s ideal customer which includes their demographic, psychographic and behavioral pattern (Galic, 2024). Here, the use of data science opens up opportunities for creating accurate customer profiles using the wealth of existing demographic information of past home buyers. Previous research conducted by Abdulhafedh in 2021 has shown success in using clustering algorithms such as K-Means, Hierarchical Clustering, and PCA to create customer profiles and define the marketing strategy of a credit card company using transaction history (Abdulhafedh, 2021).

The main motivation for this study is to create an "Effective Marketing Strategy Plan". When a property company has multiple projects running, the marketing strategy must differ for each. However, not all past marketing campaigns done by the company have been successful. This is due to various factors, such as insufficient budget for the target, unclear "call to action," and especially when the campaign strategy does not align with the market segment. The primary cause of this problem is having overly general information about their target audience, without any detailed insights. Therefore, a detailed customer profiling for each project is necessary to make marketing strategies more effective. With accurate consumer profiling, the property company can gain a deeper understanding of consumer preferences in home selection, encompassing factors like location, amenities, security, and more. Data science can be leveraged to create customer profiles by analyzing demographic data such as gender, age, region, and others. Consequently, a data science model can identify patterns to help predict a consumer's home selection based on their demographic data.

There are two objectives for this research: one from the company's perspective and the other from the field of data science. From the data science perspective, the primary objective is to compare the algorithm models (K-Means Clustering, Hierarchical Clustering, and Gaussian Mixture Model) to determine which one produces the

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.13888
e-ISSN : 2541-2019
p-ISSN : 2541-044X

most accurate clustering results for developing customer profiling for each project. The main objective from the company's perspective is to enhance brand awareness and increase sales. With increased sales, the company will grow and boost its revenue.

The research involves evaluating several data science models to find the most optimal model for creating customer profiling. From the research results, various aspects of the data science models can be developed and improve specifically for the topic of customer profiling in the property industry. Data scientists can ensure the robustness of their models and make more accurate predictions or recommendations, ultimately contributing to the successful application of applied data science techniques.

LITERATURE REVIEW

A literature study is conducted for this research to ensure that the research methodology design is clear and structured. The literature review will focus on studies related to the three clustering methodologies to be used: K-means clustering, the Hierarchical model, and the Gaussian Mixture model.

K-Means Clustering

One study conducted by Tabianan et al. in 2022 demonstrated that K-Means Clustering can be used for consumer segmentation based on customer purchasing behavior data on e-commerce platforms. However, when using large datasets with the K-Means algorithm, data smoothing is necessary for optimal results (Tabianan et al., 2022). In 2021, Ghazal et al. published a journal on using Manhattan Distance to evaluate the performance of K-Means clustering for more optimal results. Moreover, the algorithm's speed increased compared to using other distance formulas like Euclidean distance (M. Ghazal et al., 2021). In 2024, Huang et al. researched the effectiveness of K-Means clustering in detecting fraud financial data, showing that the results are more optimal when cluster analysis is applied to the dataset (Huang et al., 2024). Furthermore, another research done in the same year by Sarkar et al. showed that k-mean clustering shows a high purity rate of 0.95 with AI data of a customer segmentation analysis (Sarkar et al., 2024). This further highlight that K-Mean clustering maybe suitable for the topic of customer profiling.

K-means clustering is a data grouping method that divides a dataset into "k" groups called clusters, where each data point is placed in the cluster with the nearest centroid based on Euclidean distance. The process begins with random initialization of centroids, followed by iterations of assigning data points to the nearest cluster and updating the centroid positions based on the average of the data points in each cluster. The primary objective is to minimize the sum of squared distances between each data point and its centroid. Below is the formula for K-Means Clustering is:

$$A = \sum_{a=1}^{l} \sum_{b=1}^{n} \|x_b^a - C_a\|^2 \quad (1)$$

In this formula, A is the objective function, l is the total number of clusters in the dataset, m is the total number of cases, and C is the centroid for the cluster.

Hierarchical Clustering

Hierarchical clustering is another clustering method that will be employed in this study. A journal by Ghosal et al. in 2020 used hierarchical clustering to analyze the death rate of several countries in pandemic. The results produced from utilizing hierarchical clustering helps them analyze the behavior of the population and effectiveness in their lockdown restriction (Ghosal et al., 2020). Hence, showing that hierarchical clustering can produce compelling outcome of information from raw data. In 2019 a journal was released to proven that the average linkage in hierarchical clustering to define the distance between clusters can be further optimized. Charikar et al. uses an algorithm that is utilizes by semidefinite programming solution which employs vector representation to define hierarchical clustering at every level of detail. Additionally, it applies spreading metric constraints to enhance the robustness of the solution (Charikar et al., 2019). Another study done in 2020, research by Shetty, P., and Singh, S. showed that hierarchical clustering with Euclidean distance could produce clearer and more accurate clustering results (Shetty & Singh, 2021). Nevertheless, this algorithm version requires considerable time when handling larger datasets. In 2021, Ji et al. developed a hierarchical clustering method that accommodates data with time constraints. Ji et al. used data from distribution network operations, and their method demonstrated more accurate and efficient results (Ji et al., 2021).

Hierarchical clustering is a method in data science for grouping similar data points into clusters, ultimately forming a hierarchical structure. There are two types of hierarchical clustering: agglomerative clustering and divisive hierarchical clustering. For this study, agglomerative clustering will be used. Agglomerative clustering employs a bottom-up approach. This method starts with each data point as an individual cluster. In each iteration, the algorithm combines the two most similar clusters into a new combined cluster. To select the two clusters to be merged, the Euclidean distance formula will be used. This process is repeated until all data points are contained

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.13888
e-ISSN : 2541-2019
p-ISSN : 2541-044X

within one cluster or a predefined stopping criterion is met. The result is a dendrogram with a tree-like structure representing the hierarchy of cluster combinations. Visualization of the hierarchical clustering process can be seen in Figure 1.

$$ED = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (2)$$

The formula is used to calculate the Euclidean Distance. Here, *ED* stands for Euclidean Distance, and *(x, y)* represents the coordinates of the data points. The coordinates x_2 and $y_2$are the locations of data point 2, while x_1 and y_1 are the locations of data point 1.

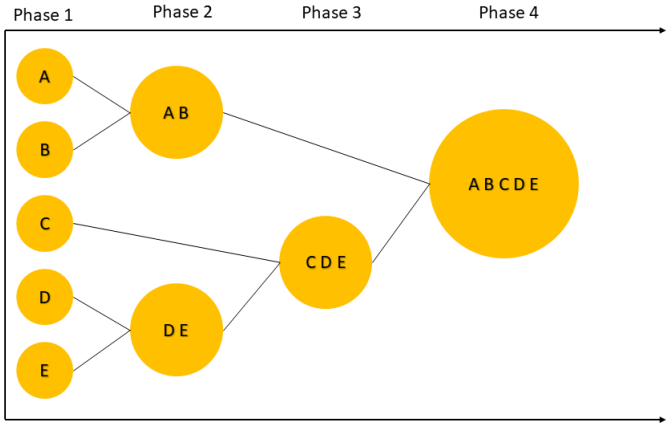

Fig. 1 Visualization of Hierarchical Clustering

Gaussian Mixture Model

A study conducted by Patel, E. and Kushawaha, D.S. in 2020 showed that the Gaussian Mixture Model is ideal for data that requires deeper and more detailed analysis. However, using the Gaussian Mixture Model is more time-consuming (Patel & Kushwaha, 2020). In 2021, Jagannathan et al. demonstrated that the Gaussian Mixture Model can also be used for detecting moving images in autonomous vehicles. This algorithm helps detect vehicles from images that have been cleaned of noise (Jagannathan et al., 2021). Additionally in the same year, Muyuan Chen and Steven J. Ludtke described an optimal way to use the Gaussian model. They started with the lowest resolution value and Gaussian function. Each additional Gaussian was placed close to the existing Gaussian. This approach can improve the convergence results of the Gaussian Mixture Model method (Chen & Ludtke, 2021). Another study done by Zhao et al. in 2023, addresses the complex problem of "Generalized Category Discovery", where both known and unknown categories exist in unlabeled data. Zhao et al. propose using a semi-supervised Gaussian Mixture Model and prototypical contrastive learning to find a stable and accurate clustering of unlabeled images. The framework shows a strong performance across various challenging datasets (Zhao et al., 2024). Hence shows that semi-supervised Gaussian Mixture Model can be an optimal way of grouping raw data into a structured format.

The Gaussian Mixture Model is a probabilistic model used for clustering in machine learning. This model assumes that the data is generated from multiple Gaussian distributions, also known as normal distributions. During the training phase, the Gaussian Mixture Model employs the Expectation-Maximization algorithm. Initially, the algorithm randomly selects parameters such as means, covariances, and weights. The algorithm then iterates between two steps: the "expectation" step and the "maximization" step. In the expectation step, the algorithm calculates the probability for each data point belonging to a cluster based on the estimated parameter values. In the maximization step, the algorithm updates the parameters to maximize the probability of the data points within the determined clusters. The expectation and maximization steps are repeated until convergence is achieved. Visualization of this process can be seen in Figure 2.

$$p(x) = \sum_{m=1}^{m} \pi_m N(X|\mu_m, \sum m) \quad (3)$$

The formula mentioned pertains to the Gaussian Mixture Model. Here, *m* represents the number of components in the mixture model, and π_m denotes the mixing coefficient that estimates the density of each Gaussian component. A component of the mixture model is derived from the Gaussian density value, represented by $N(X|\mu_m, \sum m)$ in the formula. Each component m includes all variables in the Gaussian distribution, such as the mean μ_m, covariance $\sum m$, and mixing coefficient π_m.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.13888
e-ISSN : 2541-2019
p-ISSN : 2541-044X

METHOD

Based on the literature study, the algorithms chosen for this research are K-Means (Tabianan et al., 2020), Hierarchical Clustering (Shetty and Singh, 2020), and Gaussian Mixture Model (Patel and Kushawa, 2021). The K-Means clustering by Tabianan et al. is selected for this study because their selected K-Mean methodology, as indicated by the literature review comparing various K-Means methodologies, showed more accurate results. For hierarchical clustering, the formula by Shetty and Singh will be used because their methodology utilizes Euclidean distance, which supports the type of data to be used in this research. Lastly, for the Gaussian Mixture Model, the formula by Patel and Kushawa will be applied because their model is deemed ideal for more detailed analysis.

Data Description

The data source that is used for this study is obtained from a property company X that has several projects located in the Jabodetabek area. The data used is specifically from the year 2023 and includes information from 375 customers with 9 variables or attributes. The dataset used for this study is considered complete because it includes all 9 variables required for mortgage applications. These variables and their respective data types are as follows: Project Name (Nominal), Selling Price (Ratio), Gender (Nominal), Age (Ratio), Status (Nominal), Residence (Nominal), Last Education (Nominal), Business Sector (Nominal), and Total Income (Ratio).

Methodology Procedure

The methodological procedure for this study begins with a literature review. The purpose of this stage is to analyze various research and related sources that have been conducted previously to provide a strong methodological foundation for further research planning. After that, data preparation will be carried out for the study. This stage ensures that all data is complete and accurate. Next is the implementation of the chosen clustering models using Python. The first methodology is K-Means Clustering, the second is Hierarchical Clustering, and the last is the Gaussian Mixture Model. The results of these methodologies will then be analyzed using the Silhouette and Variance Ratio metrics. This stage determines the ideal clustering model for customer profiling.

Metrics of Evaluation

To evaluate the performance of the three clustering algorithms, this study will use the Silhouette and Variance Ratio metrics. Silhouette is a method used to evaluate the clustering results of a data science model. This method measures how similar an object is to its own cluster (cohesion) and also compares it with other clusters (separation). The Silhouette score ranges from -1 to 1. If the model has a higher score, it indicates that the model has clear cluster separation. On the other hand, if the score is lower or close to -1, it means that the model does not separate the clusters well enough (Bhardwaj, 2020).

$$SO = \frac{C(i)-d(i)}{\max(C(i),d(i))} \quad (4)$$

The formula mentioned is for calculating the silhouette score. Here $c(i)$ refers to the average distance of data point i to other data points within the same cluster, and $d(i)$ refers to the distance to the nearest different cluster from data point i. Below Figure 2 show a visualization, for the silhouette metric.

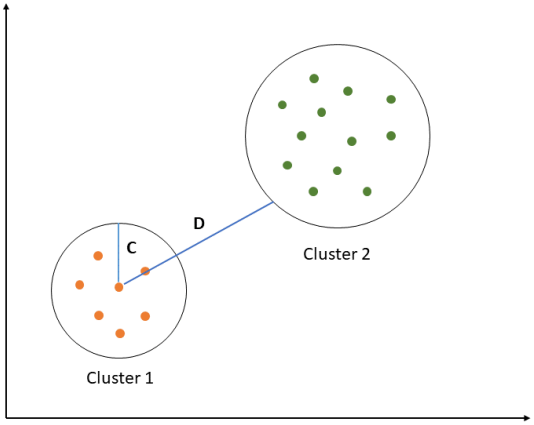

Fig. 2 Visualization of Silhouette Metric

Another metric that will be used is the variance ratio. The variance ratio is a method commonly used to evaluate the efficiency of a forecasting or trading model. It is calculated by dividing the variance of the model's predictions

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.13888
e-ISSN : 2541-2019
p-ISSN : 2541-044X

by the actual variance. A higher ratio indicates better predictive efficiency, meaning the model can capture more variability in the data (Fogarty et al., 2021).

$$F = \frac{s_1^2}{s_2^2} \quad (5)$$

The formula mentioned is for the variance ratio. In this formula, S_1^2 and S_2^2 represent the sample variances of one population (obtained from the model's predictions) and the actual results of another population, respectively.

RESULT

This section will begin with the presentation of the results from the three algorithms: K-Means Clustering, Hierarchical Clustering, and Gaussian Mixture Model. It will then proceed with an analysis of the results and the effectiveness of each algorithm.

Using K-Means Clustering will yield several results, such as cluster centers or centroids. The centroids represent the central point of each cluster found by K-Means. These centroid points represent the average result of each data point in that cluster. In K-Means Clustering, there are 3 clusters, each with its own centroids based on the 9 variables in the data. The centroid data points found for all variables can be found in Table 1. Figure 3 shows a visualization of the clusters with a scatter plot for 2 variables: Income and House Selling Price. The star points indicate the centroids of each cluster.

The scatter plot results show that the first cluster, colored yellow, has data indicating that house prices are lower when income is lower. It can be concluded that the customer profile for the yellow cluster prioritizes price over other aspects such as design or land area. The marketing division can use this information to create promotions that focuses more on “hard selling” content, such as price plays and bigger promotions. The second cluster, colored purple, has a higher range. This is because some individuals with high incomes are in this cluster, but the house prices are in the mid-range. This cluster suggests that consumers in this group consider not only price but also other factors like location, house design, and amenities. This is because the centroid of this cluster is higher than that of the yellow cluster. The third cluster, colored green, shows that higher incomes result in higher house prices. However, this cluster also has the most notable separation compared to the others. Therefore, the results from this cluster are not accurate.

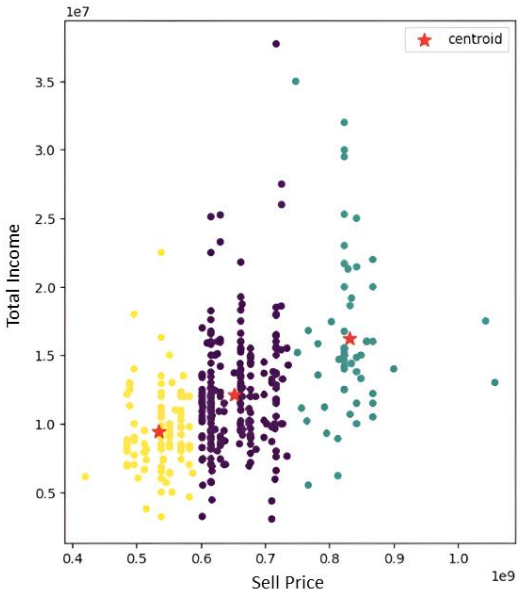

Fig. 3 Scatter Plot of Total Income and Sell Price Clusters from K-Mean Clustering

Table 1 Centroid’s Result from K-Mean Clustering

Attributes	Centroid 0	Centroid 1	Centroid 2
Project name	3.0536e+00	3.3000e+00	2.5107e+00
Selling price	6.5207e+08	8.3147e+08	5.4327e+08
Gender	4.4843e-01	3.8333e-01	3.4782e-01
Age	3.0089e+01	3.0666e+01	3.0206e+01

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.13888
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Status	4.8430e-01	4.5000e-01	4.8913e-01
Residence	2.8130e+00	2.2166e+00	1.5000e+00
Last Education	9.0134e-01	7.3333e-01	1.2065e+00
Business Sector	2.9874e+01	3.8633e+01	2.8706e+01
Total Income	1.2672e+07	1.6283e+07	9.4220e+06

To evaluate the effectiveness of K-Means Clustering, the results from the silhouette score and variance ratio will be used. Here are the results obtained from the silhouette and variance ratio:

Variance Ratio for Each Cluster:
Cluster 1: 0.102
Cluster 2: 0.041
Cluster 3: 0.024
Silhouette Score: 0.6362271293879871

The results show that K-Means Clustering using the available data provides a low variance ratio and a fairly high silhouette score. Each cluster shows a variance ratio close to 0, indicating that each cluster is more compact and well-separated. Meanwhile, the silhouette score of 0.63, which is close to 1, indicates good cohesion and separation within each cluster.

The result of hierarchical clustering is a dendrogram. This dendrogram shows how each data point merges into a cluster. The diagram below shows the dendrogram results for the last 100 data points to present the results clearly. From the bottom-up approach, it consists of 3 distinct clusters. The middle cluster, colored green in the diagram, has the most different height compared to the other 2 clusters. This is because this cluster has fewer merging points. This indicates that the objects within the cluster have higher differences. Meanwhile, the final cluster, colored red, has good merging points as the height differences between each data point are not too far apart compared to the first cluster, colored orange.

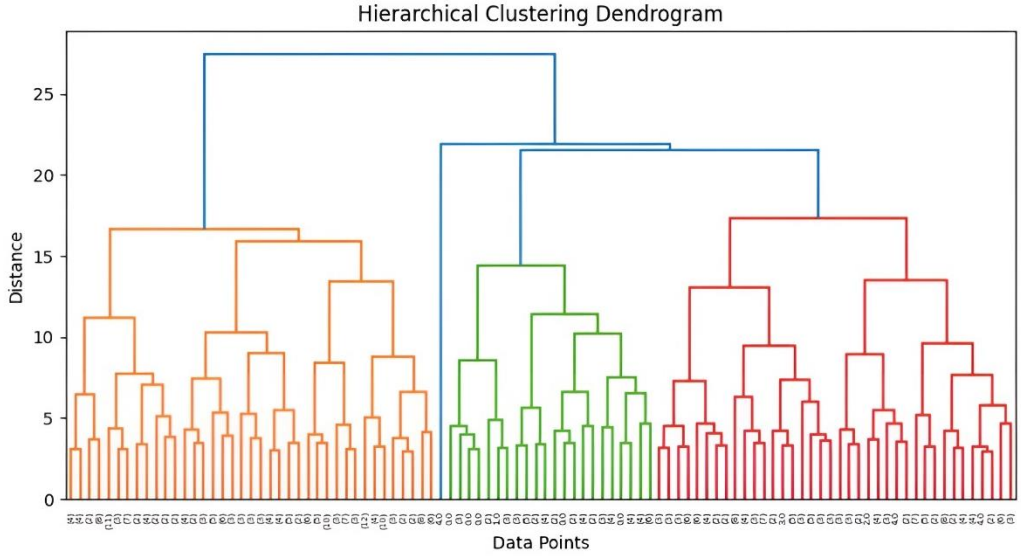

Fig. 4 Dendrogram Result from Hierarchical Clustering

To evaluate the effectiveness of Hierarchical Clustering, the results from the silhouette score and variance ratio will be used. Here are the results obtained from the silhouette score and variance ratio:

Explained Variance Ratio: [0.17978279, 0.17105316]

The results from the variance ratio and silhouette score indicate suboptimal performance. This is because the variance ratio is still too large. A smaller variance ratio suggests that the data has high-dimensional characteristics, and the patterns within the data cannot be detected by the dimensions used by the algorithm. Therefore, the hierarchical clustering algorithm is not suitable for this data.

The results from the Gaussian Mixture Model produced 3 clusters. To analyze the results of the Gaussian

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.13888
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Mixture Model, the mean value of each cluster will be used. Table 2 will show the mean value results for each cluster.

Table 2 Mean Value Results of Each Clusters from Gaussian Mixture Model

Attributes	Cluster 0	Cluster 1	Centroid 2
Project name	3.313	2.378	2.087
Selling price	6.261	6.533	7.342
Gender	0.480	0.486	0.162
Age	29.07	36.432	31
Status	0.395	0.891	0.562
Residence	2.015	1.729	2.162
Last Education	0.876	2.162	0.625
Business Sector	32.523	10.864	35.387
Total Income	1.127e+07	1.683e+07	1.418e+07

Based on the mean values, Cluster 1 has higher mean values compared to the other clusters. This can be seen from features such as "Last Education," "Total Income," and "Age." This indicates that Cluster 1 consists of individuals with higher income, older age, and higher education levels. This information can be used by the marketing division for strategic planning. To sell higher-priced houses, they should target older individuals more frequently. Additionally, Cluster 1 has the lowest nominal value in the "Business Sector" attribute. This indicates that the business sectors in this cluster are not very diverse. Consequently, the marketing division can put their focus on specific business sectors when it comes to corporate promotion. Clusters 0 and 1 have mean values that are not significantly different, except for the "Gender" and "Project" features. Cluster 2 has the lowest mean value in the "Gender" feature, indicating a gender imbalance in this grouping. However, this information also suggests high separation, resulting in inaccurate clustering. On the other hand, the first cluster has the highest mean value for the "Project" feature. This implies that consumers in the first cluster have more varied choices in selecting clusters or projects from the property company. Therefore, fair marketing is necessary to ensure that all projects receive high exposure, making it more likely for people to buy houses from property X.

However, the effectiveness of the clustering accuracy from the Gaussian Mixture Model needs to be tested with the variance ratio and silhouette score. Below are the results obtained from the silhouette score and variance ratio:

Component 0: Explained Variance Ratio = 0.224
Component 1: Explained Variance Ratio = 0.320
Component 2: Explained Variance Ratio = 0.456
Silhouette Score: 0.07099137464329296

The variance ratio for the Gaussian Mixture Model shows high results. Cluster 2 has the highest variance ratio, approaching the value of 1. This indicates that the data points in this cluster are more spread out. It suggests heterogeneity within the cluster, implying that the data points may come from different subgroups or have diverse characteristics. Moreover, the silhouette score for the Gaussian Mixture Model is very low, indicating suboptimal clustering and poor separation. Therefore, the results from the Gaussian Mixture Model clustering are inaccurate.

DISCUSSIONS

Based on the comparison of variance ratio and silhouette scores for each algorithm, the K-Means clustering algorithm is shown to be the most suitable for this dataset. K-Means clustering achieved a silhouette score of 0.63, which is closer to one, indicating better-defined clusters. The variance ratio ranges from 0.02 to 0.1, suggesting compact clusters with minimal spread. On the other hand, the Gaussian Mixture Model performed the worst for this dataset, with a silhouette score of 0.07 and a variance ratio ranging from 0.22 to 0.45. This poor performance is likely due to the Gaussian Mixture Model's suitability for more complex multidimensional data, which is not the case here. Hierarchical clustering also showed suboptimal results with a silhouette score of 0.17 and a variance ratio of 0.17. Hierarchical clustering's difficulty in handling large datasets and automatically determining the number of clusters contributed to its lower performance. K-Means clustering emerges as the most appropriate algorithm for this dataset because it is simpler compared to other algorithms and offers higher flexibility for scalability. The dataset, while not overly complex, has a large number of instances, making K-Means particularly suitable for demographic datasets in home purchasing analysis. For future studies, additional evaluation metrics such as Adjusted Rand Index and statistical validation using premutation tests should be considered to further substantiate these findings.

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

CONCLUSION

The conclusion of this study highlights that each algorithm has its strengths and limitations, which should be considered depending on the nature of the data and the analysis objectives. However, the study indicates that the K-Mean clustering algorithm is the most suitable for demographic data in home purchasing. Furthermore, the clustering results from K-Mean can be used for creating customer profiles to aid marketing planning. By better understanding target consumers through customer profiling, the marketing division can allocate more marketing budget effectively to targeted segments. This is because these consumers are more likely to purchase homes from property X, leading to increased sales and income flow. K-Means is valued for its simplicity, computational efficiency, and ability to handle datasets with clear and separated clusters. In practice, the choice of model depends on data characteristics, analysis needs, and the availability of computational resources.

REFERENCES

Abdulhafedh, A. (2021). Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation. *Journal of City and Development*, *3*(1), 12–30. https://doi.org/10.12691/jcd-3-1-3

Bhardwaj, A. (2020, May 26). *Silhouette Coefficient*. Medium. https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c

Charikar, M., Chatziafratis, V., & Niazadeh, R. (2019). Hierarchical Clustering better than Average-Linkage. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 2291–2304). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611975482.139

Chen, M., & Ludtke, S. J. (2021). Deep learning-based mixed-dimensional Gaussian mixture model for characterizing variability in cryo-EM. *Nature Methods*, *18*(8), 930–936. https://doi.org/10.1038/s41592-021-01220-5

Fogarty, J. J., Rensing, K., & Stuckey, A. (2021). *Chapter 11 Variance Ratio Test | Introduction to R and Statistics*. https://saestatsteaching.tech/section-varianceratio

Galic, D. (2024, June 11). *What are customer profiles? A complete guide, examples, and free templates*. Zendesk. https://www.zendesk.com/blog/create-data-rich-customer-profile/

Ghosal, S., Bhattacharyya, R., & Majumder, M. (2020). Impact of complete lockdown on total infection and death rates: A hierarchical cluster analysis. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *14*(4), 707–711. https://doi.org/10.1016/j.dsx.2020.05.026

Huang, Z., Zheng, H., Li, C., & Che, C. (2024). Application of Machine Learning-Based K-means Clustering for Financial Fraud Detection. *Academic Journal of Science and Technology*, *10*(1), 33–39. https://doi.org/10.54097/74414c90

Jagannathan, P., Rajkumar, S., Frnda, J., Divakarachari, P. B., & Subramani, P. (2021). Moving Vehicle Detection and Classification Using Gaussian Mixture Model and Ensemble Deep Learning Technique. *Wireless Communications and Mobile Computing*, *2021*, 1–15. https://doi.org/10.1155/2021/5590894

Ji, X., Zhang, X., Zhang, Y., Yin, Z., Yang, M., & Han, X. (2021). Three-Phase Symmetric Distribution Network Fast Dynamic Reconfiguration Based on Timing-Constrained Hierarchical Clustering Algorithm. *Symmetry*, *13*(8), 1479. https://doi.org/10.3390/sym13081479

M. Ghazal, T., Zahid Hussain, M., A. Said, R., Nadeem, A., Kamrul Hasan, M., Ahmad, M., Adnan Khan, M., & Tahir Naseem, M. (2021). Performances of K-Means Clustering Algorithm with Different Distance Metrics. *Intelligent Automation & Soft Computing*, *29*(3), 735–742. https://doi.org/10.32604/iasc.2021.019067

Patel, E., & Kushwaha, D. S. (2020). Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model. *Procedia Computer Science*, *171*, 158–167. https://doi.org/10.1016/j.procs.2020.04.017

Sarkar, M., Puja, A. R., & Chowdhury, F. R. (2024). Optimizing Marketing Strategies with RFM Method and K-Means Clustering-Based AI Customer Segmentation Analysis. *Journal of Business and Management Studies*, *6*(2), 54–60. https://doi.org/10.32996/jbms.2024.6.2.5

*name of corresponding author

Shetty, P., & Singh, S. (2021). Hierarchical Clustering: A Survey. *International Journal of Applied Research*, *7*(4), 178–181. https://doi.org/10.22271/allresearch.2021.v7.i4c.8484

Tabianan, K., Velu, S., & Ravi, V. (2022). K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. *Sustainability*, *14*(12), 7243. https://doi.org/10.3390/su14127243

Zhao, B., Wen, X., & Han, K. (2024). *Learning Semi-supervised Gaussian Mixture Models for Generalized Category Discovery*. https://github.com/DTennant/GPC.

*name of corresponding author