DETERMINATION OF POTENTIAL BUSINESS LOCATIONS USING DATA MINING CLUSTERING

Dian Erdiansyah[1]; Indra Nugraha Abdullah[2]; Amandus Jong Tallo[3*]

Master of Computer Science Study Program[1, 2]
Budi Luhur University, Jakarta, Indonesia[1, 2]
www.budiluhur.ac.id[1, 2]
dian.erdiansyah.st@gmail.com[1], indra.nugraha@budiluhur.ac.id[2]

Irrigation Design & Coastal Management Engineering Study Program, Department of Civil Engineering[3]
Kupang State Polytechnic, Kupang, Indonesia[3]
pnk.ac.id[3]
mandustallo@gmail.com[3*]
(*) Corresponding Author

Abstract—*Potential locations for businesses are highly sought after by business people to set up, expand their business, or establish a new business. Limited information on potential business locations is still a problem faced by many business people in making business decisions. The purpose of this research is to overcome the limitations of potential business location information. The approach used is the K-Means data mining clustering method which is compared to the Gaussian Mixture Model. The dataset used is residential, road access data and business points that already exist around the location. Both clustering methods are compared to the model evaluation method to determine the model with the best performance. The results show that the clustering method with the K-Means algorithm is the clustering model with the best performance. The results of the clustering resulted in 2 clusters, one of which is a cluster of potential business locations of 1041 locations. The conclusion of this study is that data mining clustering can be used to determine the optimal business location cluster. The results of this study can be recommended for business people to look for potential business locations, and for local governments to publicize potential business locations in order to attract investors from outside.*

Keywords*: clustering, data mining, gaussian mixture model, k-means, optimal business location.*

Abstrak—*Lokasi yang potensial untuk bisnis sangat dicari oleh pelaku bisnis untuk mendirikan melakukan ekpansi bisnis atau mendirikan bisnis baru. Masih terbatasnya informasi lokasi potensial bisnis masih menjadi permasalahan yang banyak dihadapi oleh pelaku bisnis dalam mengambil keputusan bisnis. Tujuan penelitian ini adalah untuk mengatasi keterbatasan informasi lokasi bisnis yang potensial. Pendekatan yang digunakan adalah metode data mining clustering K-Means yang dibandingkan dengan Gaussian Mixture Model. Dataset yang digunakan yaitu pemukiman, data akses jalan dan titik bisnis yang sudah ada di sekitar lokasi. Kedua metode clustering dibandingkan dengan metode evaluasi model untuk menentukan model dengan kinerja terbaik. Hasil penelitian menunjukkan bahwa metode clustering dengan algoritma K-Means merupakan model clustering dengan kinerja terbaik. Hasil clustering dihasilkan 2 cluster, salah satu clusternya merupakan cluster lokasi bisnis potensial sebanyak 1041 lokasi. Kesimpulan penelitian ini, data mining clustering dapat digunakan untuk penentuan cluster lokasi bisnis optimal. Hasil penelitian ini dapat direkomendasikan bagi pelaku bisnis untuk mencari lokasi bisnis yang potensial, dan bagi pemerintah daerah mempublikasikan lokasi bisnis potensial dalam rangka menarik investor dari luar.*

Kata Kunci*: clustering, data mining, gaussian mixture model, k-means, lokasi bisnis optimal.*

INTRODUCTION

Business location is one of the aspects that influence the success of a business. Choosing the optimal business location is done by many business people to start a new business or expand their business. An analysis of the business location needs to be carried out for the feasibility of the business and the success of the business that will be built in the location (Puspitaningrum & Damanuri, 2022).

Business success due to the selection of the right location will be affected by several factors such as: accessibility, visibility, traffic, space, expansion, environment and competition (Sudiantini et al., 2023). Social and environmental factors affect business costs and profits (Mohammadi et al., 2023).

The improvement of nearby facilities and modes of transportation has a great influence on the selection of business locations, by analyzing data on the location of public facilities, roads for access, district boundaries, residents (Murad et al., 2024). The quality and comfort factor for customers is highly dependent on the facilities that support the delivery of goods (Yu, 2022). Research on optimal locations using K-Means with variations in data and data dimensions has been carried out in Angilar et al. (Aguilar & Barbosa, 2023).

The potential of customers who are close to the optimal business location will further increase the success rate of the business to be built. Analysis of consumer preferences for samgyeopsal Korean cuisine attributes and market segmentation with conjoint analysis and K-means method (Ong et al., 2023). Business people will experience obstacles in determining the optimal business location due to the limitation of valid information about the optimal business location. Business people to get a decent business location still use many conventional methods such as those commonly done in business feasibility studies. The determination of a feasible business location in a business feasibility study will begin by collecting data in the field to obtain data on several locations to be selected. Location data collection activities in the field take up a lot of time and also require a lot of money, as a result of which business people cannot make decisions quickly to carry out business strategies in fierce business competition. Studies of optimal locations in several areas have been carried out extensively (Ozkaya & Demirhan, 2022), (Manika et al., 2021), (Lin et al., 2022).

With advances in the field of information technology, business people are required to be able to make business decisions correctly and faster. So that the business feasibility study method is not the initial solution for business people, and business people to get business location information quickly need the help of a business information system in the form of *Location Intelligence* that uses valid and up-to-date data. Business information systems that suit the needs of business people have begun to emerge, one of which is *Location Intelligence* software which is a type of spatial-based business intelligence product, whether it is developed by a local company or a product from outside Indonesia.

The problem that exists in most of the *Location Intelligence* is the use of improper analysis methods so that it produces information that is not accurate and not optimal. Research is conducted to answer the problem of valid information for the needs of business people, and Location *Intelligence* applications that can help business people to determine better, faster and more valid business locations with a data mining clustering approach.

The basis of this research, previous research that uses a data mining approach such as in the study of determining the optimal location of distribution centers uses the clustering method in the Inner Mongolia Autonomous Region in China and provides the lowest transportation cost benefits (Wu et al., 2022). Previous research (Leenawong and Chaikajonwat, 2023), problems in determining the location of the distribution center for convenience store franchises in Thailand that are not optimal, by using the K-Means algorithm with various distance customizations in the K-Means method, namely: *Euclidean distance, Manhattan distance, Chebyshev distance, Weighted Euclidean, Weighted Manhattan*and *Weighted Chebyshev*, looking for clusters with the lowest distribution costs.

Previous research, the K-Means method was used to group the consumption level of students in Guangdong province, China (Yang et al., 2022). Research on the distribution of cluster system of traditional settlement types in Shaanxi based on K-meansclustering algorithm (Wang et al., 2022). Previous research (Lee et al., 2021), The problem is serious, more than 30% of the pollutants produced in the Seoul metropolitan area come from diesel-fueled vehicles. This study proposes a data mining approach in selecting the optimal location of electric vehicle charging stations, in order to expand charging facilities, to further increase the use of electric cars. The method used to determine the optimal location of SPKLU is Spatial Interpolation and algorithms *Gaussian Mixture Model* (GMM). In this study, the cluster formed was 3 according to the results of the analysis of 11 out of 26 criteria suggesting k = 3 as the optimal number of clusters. This research is useful in expanding the distribution of electric cars, considering the facilities in the location, high demand, and efficient resource allocation. Previous research (Lin et al., 2022), The retail business has not been optimal in delivering goods to several customers whose locations are scattered, so a distribution center is needed to serve all. A method to determine the optimal location of logistics activities by comparing the K-Means algorithm with additional constraints to compare the performance of the K-Means algorithm, *Ant Colony Optimization* (ACO) K-means, Particle Swarm Optimization (PSO) K-Means, Fruit-Fly Optimization (FOA) K-Keans which produces 3 clusters. The results of the research on determining the location of the distribution center can be completed with the FOA K-Means algorithm which has the best performance. The FOA-K-Means algorithm has the best performance with a total distance of 184365.3 (m), *Davies-Bouldin Index* amounting to 1.846223476. ACO is not suitable for cluster analysis with large samples, poor performance on 3 indicators. PSO-K-Means is more suitable for smaller area coverage than FOA. The disadvantage of this study is that the location of the distribution center is obtained using the average of the coordinates, considering customer segmentation. Retail businesses can apply FOA-K-means to determine the most optimal distribution center location and in the context of future distribution efficiency. Determination of the best clinic location using the K-Means clsutering method using parameters such as: number of population, *Point Of Interest*, detailed information of general hospitals, specialty hospitals, clinics, emergency departments, number of wards, number of train stations, number of bus stops, number of parking lots and number of clinics (Wang et al., 2020).

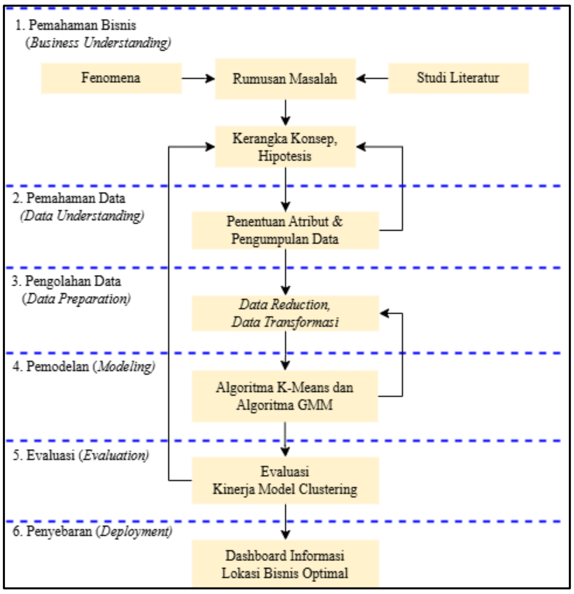
In another study that uses the K-Means clustering method for mapping the distribution of property company customers in Sidoarjo (Afifah et al., 2023), the K-Means clustering method is used to determine clustering due to the impact of natural disasters in Java (Revelation & Rushendra, 2022). In another study, using the K-Means clustering method used in the location search for public electric vehicle charging stations (SPKLU), which is the most optimal (Tambunan et al., 2023). Research on the creation of clustering to determine house prices using the K-Means and GMM clustering methods to find suitable housing, in the research Rahmattullah et al (2023).

Based on the research that has been conducted, this study will compare the clustering methods between K-Means and GMM for determining the optimal business location. The purpose of this study is to determine the potential business location by comparing the clustering methods of K-Means and GMM to get the model with the best performance.

MATERIALS AND METHODS

The method used in the research to determine the optimal business location is the CRoss Industry Standard Process for Data Mining (CRISP-DM) data mining methodology (Larose, 2014) (Figure 1).

The CRISP-DM methodology (Figure 1), consists of stages: starting with the business understanding phase which begins with the study of phenomena, problems and study studies; the data understanding phase which consists of determining dataset attributes, determining data sources and data collection; the data processing phase which consists of data reduction and data transformation, the modeling phase by comparing the K-Means and GMM methods, the evaluation phase which determines the performance of the model by evaluating the compared model, and the last phase of deployment which is presenting the clustering results in the map dashboard (Wamulkan A.S et al., 2024).


1. Pemahaman Bisnis
(Business Understanding)
Fenomena
Rumusan Masalah
Studi Literatur
Kerangka Konsep, Hipotesis
2. Pemahaman Data
(Data Understanding)
Penentuan Atribut & Pengumpulan Data
3. Pengolahan Data
(Data Preparation)
Data Reduction, Data Transformasi
4. Pemodelan (Modeling)
Algoritma K-Means dan Algoritma GMM
5. Evaluasi (Evaluation)
Evaluasi Kinerja Model Clustering
6. Penyebaran (Deployment)
Dashboard Informasi Lokasi Bisnis Optimal

Source: (Santiastry et al., 2024)

Figure 1. CRISP-DM Methodology

Business Intelligence Phase

The business understanding phase is to examine the problems that are used as the object of research. The determination of potential business locations at this time, the existing problems and their causes are studied, and the purpose of this research itself is to solve the problem so that the determination of potential business locations will be more optimal with the data mining clustering approach. Existing research is used as a direction to carry out this research, a conceptual framework is formed and a hypothesis is made that will be proven at the end of the research.

Data Understanding Phase

The data understanding phase is the activity of identifying data needs based on data needs in business understanding, identifying data sources that produce data, understanding datasets according to model needs, and ensuring that the data used is of guaranteed quality.

Data Preparation Phase

The data processing phase, starting with data collection, data processing until it becomes data ready for modeling use and with an accepted format for modeling. This stage consists of: 1) *data consolidation,* which consists of activities: collecting data from data sources, selecting data and integrating data. The data sources in the study are: POI data sourced from Googlemap, residential data and road data sourced from OSM. Other data obtained from the Geospatial Information Agency. 2) *Data Cleaning*, consisting of sub-activities such as: filling in values for data that is still blank, cleaning data for datasets, discarding inconsistent data. 3) *Data Transformation*, which consists of activities: data normalization, data conversion, data aggregate, and adding data attributes. 4) *Data Reduction*, which consists of activities: reducing the number of variables used, reducing the number of cases and balancing the data (Diana et al., 2023).

Modeling Phase

The modeling phase is carried out after the data is ready and according to the needs of the model. The model will be built by comparing the K-Means and GMM algorithms. The selection of the K-Means and GMM algorithms is based on previous research that approaches this research case, which examines locations based on area administration using the K-Means, GMM method and a comparison between K-Means and GMM (Handayanna & Sunarti, 2024), (Faidah et al., 2024), (Wahidah & Utari, 2023),

The purpose of modeling is to find clusters of potential and non-potential business locations.

The first model uses the K-Means algorithm, it searches for cluster clusters based on the closest distance and the centroid mean of the cluster does not change anymore (Santosa et al., 2020). The optimal number of clusters (K) is searched using the *Elbow* (Diana et al., 2023), optimal K selection using the *Silhoutte score* and *Calinski-Harabasz score*. After the data with the appropriate format is prepared, it is continued to select and calculate the centroid value, grouping the data based on the cluster (Wongoutong, 2024).

The second method uses the GMM algorithm, which is a model in the clustering method which is based on probability, by utilizing the weight value of combining several normal distributions (You et al., 2023). The GMM algorithm can identify and eliminate data that has a low probability in the component based on the outlier which can improve the robustness of the clustering results. Parameter determination in the GMM algorithm can use *BIC Score*, that is, to determine the optimal K.

The application of modeling with K-Means and GMM algorithms is carried out using Python enriched with *machine learning* libraries, namely pandas, and other libraries. The modeling results of each algorithm will form clusters of potential business locations.

Evaluation Phase

The model evaluation phase is carried out to get the model with the best performance. The first evaluation method used is: *Silhouette score*, one of the methods to evaluate the results of the clustering method (Rohman & Wibowo, 2024). The value range of the silhouette coefficient is between -1 to 1, and the data clustering system is said to be good when the value of the *Silhouette Index* close to 1 (Rohman & Wibowo, 2024). In the research Rahmattullah et al (2023), evaluation with *Silhouette score* in the K-Means model got better results than the GMM model. The second evaluation method is *Calinski-Harabasz score*. The comparison of the above evaluation methods will be carried out in clustering with K-Means, DBSCAN and GMM algorithms. The model comparison activity used was carried out to select the most optimal model for the K-Means and GMM models. For comparison, the model will use the results of the evaluation, namely: the *Silhouette score* and *Calinski Harabasz score*.

Deployment Phase

At the stage of dissemination of modeling results, namely potential business locations in the form of clustering potential business locations in a map data visualization. The benefit is that it is hoped that business people in decision-making can easily explore information and insights in the form of this deployment.

RESULTS AND DISCUSSION

The research on the location of potential businesses with the CRISP-DM process approach will look for the best algorithm to be used further. By comparing between the K-Means and GMM algorithms, and evaluating the model, one of the best models was selected.

Results of Business Understanding

Determining the best business location will provide *the most* ideal business location insights that business people need to make decisions in establishing a new place of business. The optimal business location is greatly influenced by the number of potential customers, socioeconomic status, and the existence of a number of established businesses. To measure the potential level of customers can be measured as mentioned above, a method will be used using datasets of residential areas, road access, POIs and business land. The determination of the parameters used will be determined by the evaluation method.

Data Understanding Results

The data understanding stage The first activity carried out is to determine the data needed to determine the best business location as explained in the *business understanding stage*. The data source for the entire dataset is as shown in Table 1.

Table 1. Data Source

It	Data Name	Data Source	Year
1	POI	Googlemap	2023
2	Road	OSM	2023
3	Settlement	BIG	2023

Source: (Research Result, 2024)

The data source used is a consideration for data quality. Quality data will be used in modeling, so that with good data quality it will produce quality modeling output. The final activity at this stage is to prepare the dataset attributes to be used in modeling. The data attributes for the study consist of village data attributes and location data attributes as shown in Table 2 and Table 3.

Table 2. Grid Data Attributes

It	Attribute	Description
1	ID Grid	Grid uniques ID
2	Broad	Grid size in units area

Source: (Research Result, 2024)

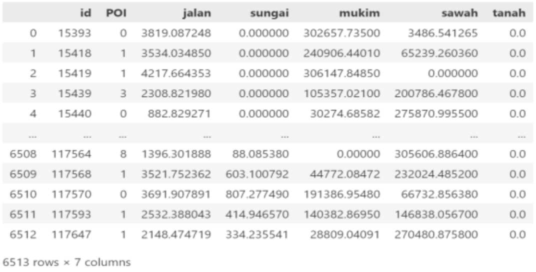
Table 3. Location Data Attributes

It	Attribute	Description
1	ID Grid	Grid Unique ID
2	POI	Point of Interest
3	Road	Length of road in meters
4	Settlement	Settlement area in meters
5	Paddy	Rice field area in meters
6	Vacant land	Area of vacant land in meters
7	River	The area of the river section in meters

Source: (Research Result, 2024)

Data Preparation Results

This stage begins with the collection of data from data sources, namely residential land data, road data and POI data. The initial data collected in GIS format is in the form of polygons like villages and in the form of points such as POIs. All data is integrated with the help of GIS software so that the dataset to be used becomes the following display (Figure 2).



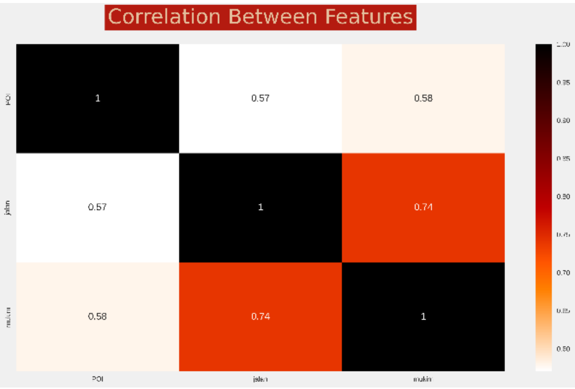
	id	POI	jalan	sungai	mukim	sawah	tanah
0	15393	0	3819.087248	0.000000	302657.73500	3486.541265	0.0
1	15418	1	3534.034850	0.000000	240906.44010	65239.260360	0.0
2	15419	1	4217.664353	0.000000	306147.84850	0.000000	0.0
3	15439	3	2308.821980	0.000000	105357.02100	200786.467800	0.0
4	15440	0	882.829271	0.000000	30274.68582	275870.995500	0.0
...	...	...	...	...	...	...	...
6508	117564	8	1396.301888	88.085380	0.00000	305606.886400	0.0
6509	117568	1	3521.752362	603.100792	44772.08472	232024.485200	0.0
6510	117570	0	3691.907891	807.277490	191386.95480	66732.856380	0.0
6511	117593	1	2532.388043	414.946570	140382.86950	146838.056700	0.0
6512	117647	1	2148.474719	334.235541	28809.04091	270480.875800	0.0

6513 rows × 7 columns

Source: (Research Result, 2024)

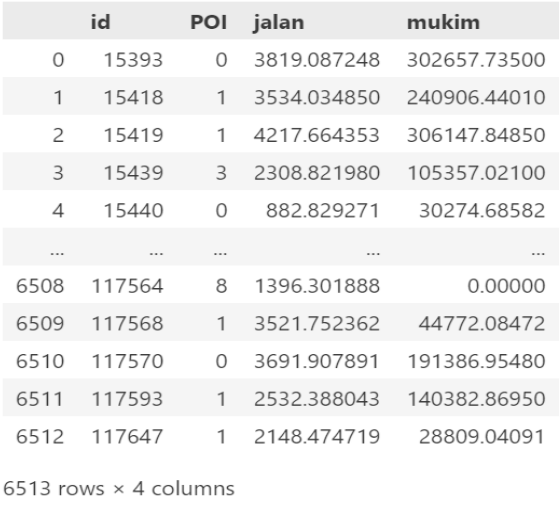
Figure 2. Initial dataset

From the data as shown in the table above, the results of feature selection and analysis are determined data parameters which are divided into: POI, roads and settlements (Figure 4). So that with the reduction of this parameter, the correlation between data becomes better.



Source: (Research Result, 2024)

Figure 3. Correlation between Features



	id	POI	jalan	mukim
0	15393	0	3819.087248	302657.73500
1	15418	1	3534.034850	240906.44010
2	15419	1	4217.664353	306147.84850
3	15439	3	2308.821980	105357.02100
4	15440	0	882.829271	30274.68582
...	...	...	...	...
6508	117564	8	1396.301888	0.00000
6509	117568	1	3521.752362	44772.08472
6510	117570	0	3691.907891	191386.95480
6511	117593	1	2532.388043	140382.86950
6512	117647	1	2148.474719	28809.04091

6513 rows × 4 columns

Source: (Research Result, 2024)

Figure 4. Dataset after Feature Selection

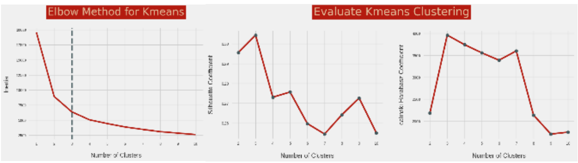
Data cleaning is carried out to clean data that has *missing values* or duplicate data. In the transformation data activity, standardization is carried out from the dataset so that the difference in data values is not too far. Data reduction is not done so the same dataset is still used. The *well-formed* data is then ready to be used for datasets in modeling.

Modeling Results

The first activity at this stage is the application of the model with the K-Means algorithm, the second with the GMM algorithm. The application of the K-Means model begins with determining the best number of clusters (K), starting with finding the optimal number of clusters

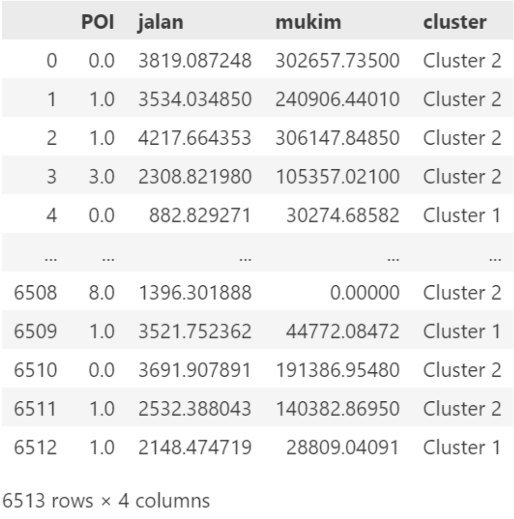
(K). Next determine the centroid, calculate the distance of all points to the centroid by using the Euclidean distance formula. The results of the distance calculation are then grouped according to the minimum distance or similarity. The repetition is done by calculating the average and making sure the new centroid is the same as the previous centroid. Repeating is carried out until the new centroid is obtained will be the same as the previous centroid, until the final cluster is formed.

The first step of the clustering method with the K-Means algorithm is to determine the ideal number of clusters (K), the best K value has been determined by the Elbow method, Silhouette score and Calinski Harabasz score. The following is a comparison of the K value obtained by the Elbow, Silhouette Index and Calinski Harabasz Index methods in the following graph (Figure 5).



Source: (Research Result, 2024)
Figure 5. Determination of Optimal K Cluster

Using the optimal number of clusters, clusters are calculated using the K-Means method (Figure 6).

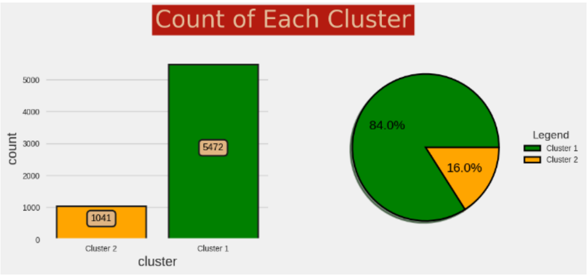


	POI	jalan	mukim	cluster
0	0.0	3819.087248	302657.73500	Cluster 2
1	1.0	3534.034850	240906.44010	Cluster 2
2	1.0	4217.664353	306147.84850	Cluster 2
3	3.0	2308.821980	105357.02100	Cluster 2
4	0.0	882.829271	30274.68582	Cluster 1
...	...	...	...	...
6508	8.0	1396.301888	0.00000	Cluster 2
6509	1.0	3521.752362	44772.08472	Cluster 1
6510	0.0	3691.907891	191386.95480	Cluster 2
6511	1.0	2532.388043	140382.86950	Cluster 2
6512	1.0	2148.474719	28809.04091	Cluster 1

6513 rows × 4 columns

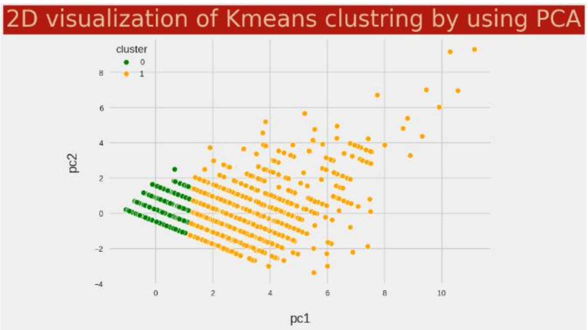
Source: (Research Result, 2024)
Figure 6. Best Business Location Cluster Visualization

Modeling with the K-Means algorithm resulted in two clusters consisting of cluster 1 as many as 5472 (84%) cluster members and cluster 2 (potential business location) as many as 1041 (16%) cluster members (Figure 7).



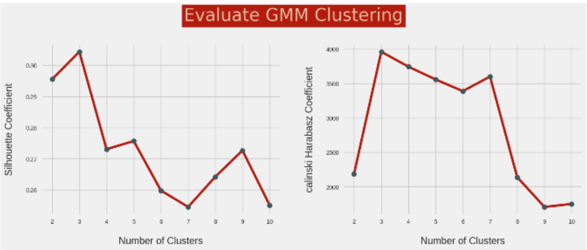
Source: (Research Result, 2024)
Figure 7. Potential Business Location Cluster

Furthermore, the results are visualized in 2D with simplification with the PCA variable (Figure 8).



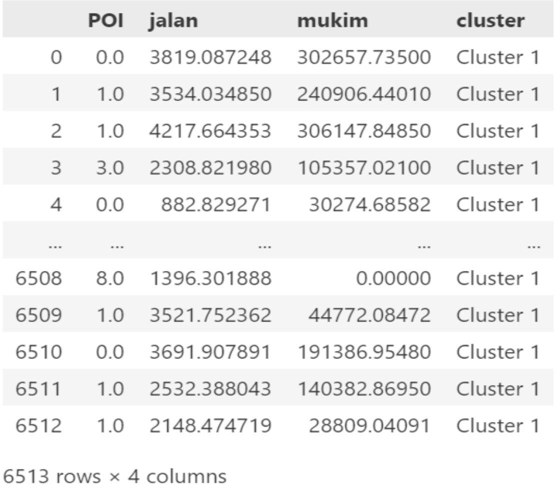
Source: (Research Result, 2024)
Figure 8. 2D Visualization of K-Means Clustering Results

The second modeling, using the GMM algorithm which starts with the selection of optimal parameters for the GMM model using the BIC score method. The GMM clustering method assumes that all points are a mixture of Gaussian opportunity distributions which are Gaussian distributions. Each distribution will have distribution parameters, the *Expectation Maximization* (EM) algorithm is used to model GMM. To determine the most optimal number of clusters through clustering modeling, the GMM method using BIC Score can be directly evaluated by the *Silhouete Index* and *Calinski-Harabasz Index* evaluation methods (Figure 9).



Source: (Research Result, 2024)
Figure 9. Evaluation of the Optimal K of the GMM Algorithm

From the graph above, it can be explained that the most optimal number of clusters (K) as a result of model evaluation with two evaluation methods is 2 with *a Silhouete score* of > 0.6 and a *Calinski-Harabasz score* of > 2000.



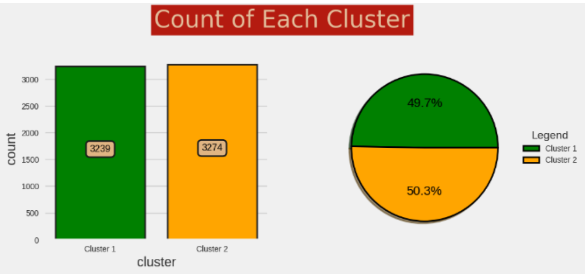
	POI	jalan	mukim	cluster
0	0.0	3819.087248	302657.73500	Cluster 1
1	1.0	3534.034850	240906.44010	Cluster 1
2	1.0	4217.664353	306147.84850	Cluster 1
3	3.0	2308.821980	105357.02100	Cluster 1
4	0.0	882.829271	30274.68582	Cluster 1
...	...	...	...	...
6508	8.0	1396.301888	0.00000	Cluster 1
6509	1.0	3521.752362	44772.08472	Cluster 1
6510	0.0	3691.907891	191386.95480	Cluster 1
6511	1.0	2532.388043	140382.86950	Cluster 1
6512	1.0	2148.474719	28809.04091	Cluster 1

6513 rows × 4 columns

Source: (Research Result, 2024)

Figure 10. GMM Potential Business Location Cluster

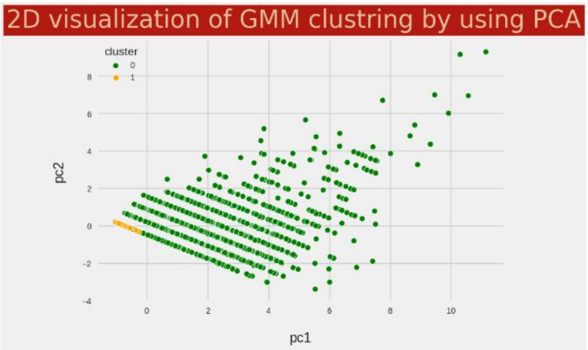
Modeling with the GMM clustering method resulted in two clusters, each cluster 1 with 3239 members or 50.3% and cluster 2 which is a cluster of potential business locations with 3274 members or 49.7% (Figure 11).



Source: (Research Result, 2024)

Figure 11. Number of GMM Cluster Members

Visualization of clustering potential business locations with 2D PCA parameters (Figure 12).

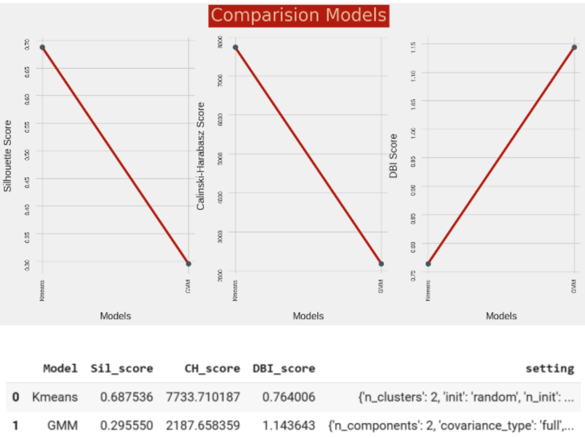


Source: (Research Result, 2024)

Figure 12. GMM Clustering Results

Evaluation Results

In this study, an evaluation of the model used was carried out, using the *Silhouette Coefficient* and *Calinski-Harabasz Index evaluation* methods on the modeling results with the K-Means and GMM algorithms.



	Model	Sil_score	CH_score	DBI_score	setting
0	Kmeans	0.687536	7733.710187	0.764006	{'n_clusters': 2, 'init': 'random', 'n_init': ...
1	GMM	0.295550	2187.658359	1.143643	{'n_components': 2, 'covariance_type': 'full',...

Source: (Research Result, 2024)

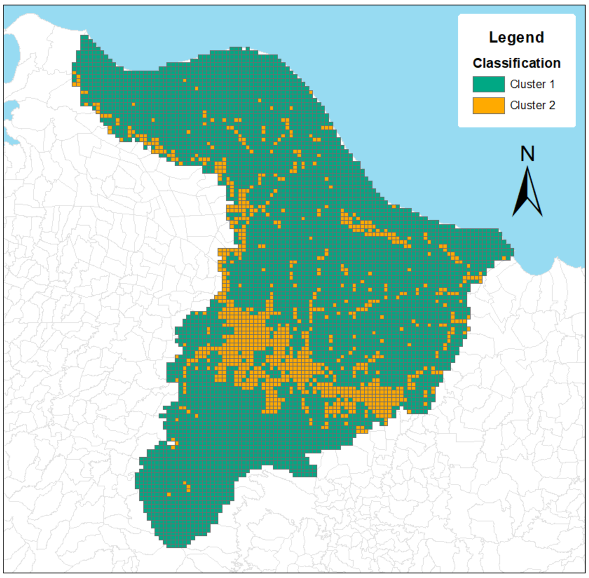
Figure 13. Model Evaluation Results

The results obtained (Figure 13), showing that the model with the K-Means algorithm obtained the value *Silhouette Coefficient* of 0.687536 and the value of *Calinski-Harabasz index* 7737.710187, while with the GMM algorithm the value of *Silhouette Coefficient* of 0.295550 and a value of *Calinski-Harabasz* score of 2187.658359. The results of the research, the K-Means method has the best performance, in line with the research conducted by Rahmattullah et al (2023), and is not in line with research that uses the same clustering method, namely the K-Means and GMM methods, which states that the GMM method has better performance than the K-Means method (Wahidah & Utari, 2023).

Deployment Results

Deployment from the results of the research on determining potential business locations, the method with the best performance was chosen, namely the K-Means method.

The results of clustering are presented in spatial-based information visualization so that it is easy for information users to understand, so that it will speed up business decision-making, especially determining potential business locations. The results of the optimal business location research are visualized in a map (Figure 14).



Source: (Research Result, 2024)

Figure 14. Visualization of Potential Business Locations

The results of this study (Figure 14) identified potential business locations (cluster 2), namely locations with orange color that spread in Karawang Regency.

Attribute information from the map in Figure 13, the results of analytical data mining in Karawang Regency detected a distribution of 1041 potential business locations. East Telukjambe and Klari districts are the sub-districts with the most potential locations, namely 98 and 94 locations. Based on Table 4, apart from the 2 sub-districts above, 6 sub-districts, namely: Cikampek, Ciampel, East Karawang, Kota Baru, Rengasdengklok and West Karawang are included in areas that have many potential locations for doing business. Complete sub-district profiling with number of villages, number of potential locations and names of villages with the most potential locations can be seen in Table 4 below.

Table 4. Potential Locations

Subdistrict	Count		Village
	Village	Locations	
Telukjambe Timur	9	98	Sirnabaya
Klari	13	94	Pancawati
Cikampek	10	67	Kalihurip
Ciampel	4	63	Kutanegara
Karawang Timur	8	57	Kondangjaya
Kota Baru	9	55	Wancimekar
Rengasdengklok	8	49	Amansari
Karawang Barat	8	46	Nagasari
Telukjambe Barat	9	42	Margakaya
Purwasari	8	36	Purwasari
Tempuran	11	34	Pancakarya
Batujaya	8	28	Batujaya
Cilamaya Kulon	10	26	Sumurgede
Cilamaya Wetan	9	26	Cilamaya
Majalaya	6	26	Bangle
Pedes	11	23	Jatimulya
Telagasari	13	23	Pasirtalaga
Tirtamulya	9	23	Citarik
Jatisari	11	21	Balonggandu
Tirtajaya	11	20	Kutamakmur
Banyusari	10	19	Jayamukti
Jayakerta	8	19	Medangasem
Kutawaluya	7	18	Kutakarya
Lemahabang	8	18	Karyamukti
Cibuaya	8	14	Cibuaya
Pakisjaya	6	13	Tanjungbungin, Tanahbaru
Cilebar	5	11	Ciptamargi
Rawamerta	6	8	Sukamerta
Pangkalan	2	5	Tamanmekar
Tegalwaru	3	4	Cintalaksana

Source: (Research Result, 2024)

Meanwhile, areas that have few potential locations only have under 10 potential locations, namely Rawamerta, Pangkalan and Tegalwaru sub-districts. Insights from this data mining can help business people in opening new business locations by considering the location of sub-districts which have many locations, the number of villages and the most dominant villages with their potential locations in each sub-district. Apart from that, this information can be used by those interested in local taxes or investment, it can be estimated where the potential for tax or income is the most, as well as where there is potential for investment.

CONCLUSION

In this study, based on the model evaluation method used and the model with two algorithms used, namely K-Means and GMM, two clusters were formed and the K-Means method is the most optimal and most reverse-performing model based on the results of the model evaluation. The selection of the parameters used was also analyzed with the same evaluation method so that the feature selection process was carried out. So it can be concluded that data mining can be used to determine potential business locations using the K-Means method. The results of this study help business people in making decisions in choosing potential business locations to start their business.

REFERENCE

Afifah, A. N., Nurcahyawati, V., & Hananto, V. R. (2023). Analisis Clustering dan Pemetaan Sebaran Pelanggan Perusahaan Properti di Sidoarjo. *Jurnal Edukasi Dan Penelitian Informatika*, *9*(3), 502–508. https://jurnal.untan.ac.id/index.php/jepin/article/view/67935

Aguilar, E. J., & Barbosa, V. C. (2023). Shape complexity in cluster analysis. *PLoS ONE*, *18*(5

May), 1–19. https://doi.org/10.1371/journal.pone.0286312

Diana, A., Ariesta, A., Wibowo, A., & Risaychi, D. A. B. (2023). New Student Clusterization Based on New Student Admission Using Data Mining Method. *Jurnal Pilar Nusa Mandiri*, *19*(1), 1–10. https://doi.org/10.33480/pilar.v19i1.4089

Faidah, D. Y., Hudzaifa, A. M., Theresia, N., & Widiantoro, C. E. (2024). Optimalisasi Strategi Pengelompokkan Potensi Padi Sebagai Solusi Efektif Kelangkaan Beras Di Jawa Barat. *Jurnal Lebesgue : Jurnal Ilmiah Pendidikan Matematika, Matematika Dan Statistika*, *5*(1), 529–537. https://doi.org/10.46306/lb.v5i1.592

Handayanna, F., & Sunarti, S. (2024). Penerapan Algoritma K-Means Untuk Mengelompokkan Kepadatan Penduduk Di Provinsi DKI Jakarta. *Journal of Applied Computer Science and Technology*, *5*(1), 50–55. https://doi.org/10.52158/jacost.v5i1.477

Larose, C. D. (2014). *Discovering Knowledge in data : An Introduction to Data Mining* (2nd ed.). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118874059

Lee, J., An, M., Kim, Y., & Seo, J. I. (2021). Optimal allocation for electric vehicle charging stations. *Energies*, *14*(18), 1–9. https://doi.org/https://doi.org/10.3390/en14185781

Leenawong, C., & Chaikajonwat, T. (2023). Modified K-Means Clustering for Demand-Weighted Locations : A Thailand ' s Convenience Store Franchise - Case Study. *Science and Technology*, *31*(2), 655–670.

Lin, T. X., Wu, Z. H., & Pan, W. T. (2022). Optimal location of logistics distribution centres with swarm intelligent clustering algorithms. *PLoS ONE*, *17*(8 August), 1–16. https://doi.org/10.1371/journal.pone.0271928

Manika, S., Karalidis, K., & Gospodini, A. (2021). Mechanism for the Optimal Location of a Business as a Lever for the Development of the Economic Strength and Resilience of a City. *Urban Science*, *5*(4). https://doi.org/10.3390/urbansci5040070

Mohammadi, Z., Barzinpour, F., & Teimoury, E. (2023). A location-inventory model for the sustainable supply chain of perishable products based on pricing and replenishment decisions: A case study. *PLoS ONE*, *18*(7 July), 1–29. https://doi.org/10.1371/journal.pone.0288915

Murad, A., Faruque, F., Naji, A., Tiwari, A., Qurnfulah, E., Rahman, M., & Dewan, A. (2024). Optimizing health service location in a highly urbanized city: Multi criteria decision making and P-Median problem models for public hospitals in Jeddah City, KSA. *PLoS ONE*, *19*(1 January), 1–14. https://doi.org/10.1371/journal.pone.0294819

Ong, A. K. S., Prasetyo, Y. T., Esteller, A. J. D., Bruno, J. E., Lagorza, K. C. O., Oli, L. E. T., Chuenyindee, T., Thana, K., Persada, S. F., & Nadlifatin, R. (2023). Consumer preference analysis on the attributes of samgyeopsal Korean cuisine and its market segmentation: Integrating conjoint analysis and K-means clustering. *PLoS ONE*, *18*(2 February), 1–23. https://doi.org/10.1371/journal.pone.0281948

Ozkaya, G., & Demirhan, A. (2022). Multi-Criteria Analysis of Sustainable Travel and Tourism Competitiveness in Europe and Eurasia. *Sustainability (Switzerland)*, *14*(22). https://doi.org/10.3390/su142215396

Puspitaningrum, Y., & Damanuri, A. (2022). Analisis Lokasi Usaha Dalam Meningkatkan Keberhasilan Bisnis Pada Grosir Berkah Doho Dolopo Madiun. *Niqosiya: Journal of Economics and Business Research*, *2*(2), 289–304. https://doi.org/https://doi.org/10.21154/niqosiya.v2i2.977

Rahmattullah, R., Indwiarti, I., & Rohmawati, A. A. (2023). Clustering Harga Rumah: Perbandingan Model K-Means dan Gaussian Mixture Model. *E-Proceeding Of Engineering*, *10*(3), 3441–3449. https://openlibrary.telkomuniversity.ac.id/pustaka/files/185889/jurnal_eproc/clustering-harga-rumah-perbandingan-model-k-means-dan-gaussian-mixture-model.pdf

Rohman, N., & Wibowo, A. (2024). Clustering of Popular Spotify Songs in 2023 Using K-Means Method and Silhouette Coefficient. *Jurnal Pilar Nusa Mandiri*, *20*(1), 18–24. https://doi.org/10.33480/pilar.v20i1.4937

Santiastry, S., Apriandari, W., Informatika, T., Sukabumi, U. M., Sukabumi, K., Bayes, N., Inggris, T. B., & Sukabumi, U. M. (2024). *PENERAPAN ALGORITMA NAIVE BAYES DAN METODE CRISP-DM DALAM*. *8*(5), 10432–10439.

Santosa, R. G., Chrismanto, A. R., & Kurniawan, E. (2020). JEPIN (Jurnal Edukasi dan Penelitian Informatika) Analisis Cluster Terhadap Karakteristik Mahasiswa Jalur Prestasi FTI UKDW. *JEPIN*, *6*(1), 13–22. https://jurnal.untan.ac.id/index.php/jepin/article/view/37216

Sudiantini, D., Febrianti, A. S., Nugroho, A. A., Jannah,

N. A., Candra, M., Edwar, R. A., & Aliyanti, T. (2023). Pengaruh Pengambilan Lokasi Usaha Terhadap Kesuksesan Berbisnis UMKM. *Jurnal Ilmiah Multidisipline*, *1*(9), 334–338. https://doi.org/https://doi.org/10.5281/zenodo.10044851

Tambunan, H. B., Sitanggang, R. B., Mafruddin, M. M., Prasetyawan, O., Kensianesi, Istiqomah, Cahyo, N., & Tanbar, F. (2023). Initial location selection of electric vehicles charging infrastructure in urban city through clustering algorithm. *International Journal of Electrical and Computer Engineering*, *13*(3), 3266–3280. https://doi.org/10.11591/ijece.v13i3.pp3266-3280

Wahidah, Z., & Utari, D. T. (2023). Comparison of K-Means and Gaussian Mixture Model in Profiling Areas By Poverty Indicators. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, *17*(2), 0717–0726. https://doi.org/10.30598/barekengvol17iss2pp0717-0726

Wahyu, A., & Rushendra, R. (2022). Klasterisasi Dampak Bencana Gempa Bumi Menggunakan Algoritma K-Means di Pulau Jawa. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, *8*(1), 174. https://doi.org/10.26418/jp.v8i1.52260

Wamulkan A.S, U. A. H., Utami, N. W., & Anggara, I. N. Y. (2024). Bali Tourist Visits Clustered Via Tripadvisor Reviews Using K-Means Algorithm. *Jurnal Pilar Nusa Mandiri*, *19*(2), 117–124. https://doi.org/10.33480/pilar.v19i2.4571

Wang, X., Shao, C., Xu, S., Zhang, S., Xu, W., & Guan, Y. (2020). Study on the location of private clinics based on K-means clustering method and an integrated evaluation model. *IEEE Access*, *8*(1), 23069–23081. https://doi.org/10.1109/ACCESS.2020.2967797

Wang, X., Shen, A., Hou, X., & Tan, L. (2022). Research on cluster system distribution of traditional fort-type settlements in Shaanxi based on K-means clustering algorithm. *PLoS ONE*, *17*(3 March). https://doi.org/10.1371/journal.pone.0264238

Wongoutong, C. (2024). The impact of neglecting feature scaling in k-means clustering. *PLoS ONE*, *19*(12), 1–19. https://doi.org/10.1371/journal.pone.0310839

Wu, J., Liu, X., Li, Y., Yang, L., Yuan, W., & Ba, Y. (2022). A Two-Stage Model with an Improved Clustering Algorithm for a Distribution Center Location Problem under Uncertainty. *Mathematics*, *10*(14). https://doi.org/10.3390/math10142519

Yang, C., Wen, H., Jiang, D., Xu, L., & Hong, S. (2022). Analysis of college students' canteen consumption by broad learning clustering: A case study in Guangdong Province, China. *PLoS ONE*, *17*(10 October), 1–18. https://doi.org/10.1371/journal.pone.0276006

You, J., Li, Z., & Du, J. (2023). A new iterative initialization of EM algorithm for Gaussian mixture models. *PLoS ONE*, *18*(4 April), 1–17. https://doi.org/10.1371/journal.pone.0284114

Yu, W. (2022). Robust competitive facility location model with uncertain demand types. *PLoS ONE*, *17*(8 August), 1–22. https://doi.org/10.1371/journal.pone.0273123