Jurnal Aplikasi Sains Data
Journal homepage: https://jasid.upnjatim.ac.id/index.php/jasid

Application of K-Means Clustering for Regency/City Clustering in East Java Based on 2024 Human Development Index Indicators

Kholidatus Emilia[1], Ayu Sri Rahayu[2], Devina Putri Yuliani[3], Dwi Arman Prasetya[4], Prismahardi Aji Riyantoko[5]

[1,2,3,4]Department of Data Science, Universitas Pembangunan Nasional "Veteran" Jawa Timur, [5]Department of Information and Communication Systems, Okayama University

[1]22083010103@student.upnjatim.ac.id, [2]22083010107@student.upnjatim.ac.id, [3]22083010028@student.upnjatim.ac.id, [4]arman.prasetya.sada@upnjatim.ac.id, [5]pnai2m3s@s.okayama-u.ac.jp

ABSTRACT

Abstract— This study applies the K-Means clustering algorithm to group 38 regencies and cities in East Java Province based on five Human Development Index (HDI) indicators for the year 2024. These indicators include Life Expectancy (UHH), Expected Years of Schooling (HLS), Mean Years of Schooling (RLS), and Real Expenditure Per Capita (PPK). The aim of this research is to uncover hidden patterns and disparities in regional development, which can be used as a basis for more targeted and data-driven policy interventions.The optimal number of clusters was determined using three evaluation metrics: the Elbow Method, Silhouette Score, and Davies-Bouldin Index. These evaluations collectively identified three distinct clusters. Cluster 0 represents regions with high levels of development across all indicators. Cluster 1 consists of regions with moderate development levels and potential for improvement, while Cluster 2 contains regions with significantly lower values, particularly in education and income metrics.In addition to clustering, a correlation analysis was conducted to examine the relationship between HDI and its supporting indicators. The results show that Mean Years of Schooling (RLS) and Real Expenditure Per Capita (PPK) have the strongest positive correlation with HDI across all clusters. This highlights the key role of education and economic well-being in improving human development. The findings emphasize the importance of clustering analysis in shaping equitable and region-specific development strategies.

Keywords: Human Development Index, K-Means Clustering, East Java, Correlation Analysis

I. INTRODUCTION

National development in Indonesia aims to improve the overall welfare of society, including economic, educational, and health aspects. In order to achieve this goal, indicators are needed that can represent the quality of human life comprehensively. The Human Development Index (HDI) is one of the measurement tools used for this purpose, covering three main dimensions. These dimensions include longevity and healthy living, as measured by life expectancy at birth; access to education, as measured by expected years of schooling and average years of schooling; and a decent standard of living, as measured by real expenditure per capita [1]. Every year the HDI becomes an important indicator in evaluating human development achievements, both at the national and regional levels, including districts/cities.

East Java Province, as one of the provinces with the largest number of districts/cities in Indonesia, shows a development diversity. Variations in geographical, social, and economic conditions between

* *Corresponding author.*
E-mail address: arman.prasetya.sada@upnjatim.ac.id
Doi: 10.33005/jasid.v1i2.21

regions are factors that influence differences in HDI achievements in each region. Therefore, an analysis is needed that can group the regions in this province based on their similar human development characteristics. This grouping is expected to be capable of providing better insight into development conditions and support more efficient and appropriate program planning. For this purpose, one of the methods that can be used is cluster analysis.

Cluster analysis is a statistical method used to group objects based on the level of similarity between these objects. One of the commonly used algorithms in cluster analysis is K-Means. This algorithm divides data into several groups or clusters based on the distance of the data to the cluster centre (centroid). By using K-Means, patterns and interrelationships between districts/cities in East Java based on HDI indicators can be identified more systematically. The clustering results can also be used to map the distribution of human development in the region.

The clustering process using the K-Means algorithm provides a more systematic overview of human development conditions in each district/city in East Java Province. By utilizing the four main indicators of the Human Development Index—namely Life Expectancy (AHH), Literacy Rate (AMH), Average Years of Schooling (RLS), and Per Capita Expenditure (PPK)—it is possible to identify groups of regions with similar development characteristics. The clustering results are highly useful in helping local governments set development priorities and allocate resources more effectively and efficiently. Moreover, the output of this analysis can serve as a foundation for formulating evidence-based policies, allowing development programs in East Java to be more targeted and tailored to the specific needs of each group of regions. Thus, the cluster analysis approach using the K-Means algorithm becomes an important tool in supporting more focused and sustainable regional development planning.

II. RESEARCH METHODOLOGY

A. Data Collection

Data collection is the initial stage in the research process that focuses on obtaining information relevant to the topic under study, with the aim of obtaining a thorough and accurate understanding of the phenomenon being analyzed [2]. The data used in this study is secondary data obtained from the official publication of the Central Bureau of Statistics (BPS) of East Java Province. The data collected includes the values of five socioeconomic indicators from 38 districts/cities in East Java. These indicators include the Human Development Index (HDI), Life Expectancy (UHH), Average Years of Schooling (RLS), Expected Years of Schooling (HLS) and Real Expenditure per Capita (PPK). All data were collected within the same year so that the results of the cluster analysis would not be biased by differences in measurement time.

B. Data Preprocessing

The data preparation process is a data treatment stage aimed at producing a useful and high-quality model. This stage is the most resource-intensive for the analysis team. A good and accurate model begins with good data preparation [3] The steps carried out include checking for missing values, checking for duplicate data, identifying rows containing outliers, removing unnecessary columns, and finally, normalizing the data. The results showed that there were no missing values and no duplicate data. However, outliers were found in the Expected Years of Schooling (HLS) and Per Capita Expenditure (PPK) columns.

Based on a brief review of the data, there is one column that is not needed.One column in the dataset was removed to simplify the analysis process. The column that was removed is "Kabupaten/Kota". Finally, data normalization was performed with the aim of increasing the accuracy of the algorithm and enhancing the interpretability of visualizations.

C. Data Modelling

The main analysis was conducted using the K-Means Clustering method, which is an unsupervised learning algorithm for dividing data into a number of clusters based on the similarity of values.This algorithm works by randomly determining the cluster center (centroid), and then grouping the data based on the closeness of the Euclidean distance. The process proceeds iteratively, where the centroid is updated until it converges. Determination of the optimal number of clusters is done using the Elbow method, by calculating Within-Cluster Sum of Squares (WCSS) values at various values of k, and selecting the elbow point as the best number of clusters.

D. Clustering

In this study, the determination of the optimal number of clusters was carried out using the Elbow Method, which is a common approach in cluster analysis to identify the most representative number of clusters that fit the structure of the data.

The analytical process began by applying the K-Means algorithm to a range of cluster numbers, specifically from k = 1 to k = 10. For each value of k, the inertia or Within-Cluster Sum of Squares (WCSS) was calculated, which represents the total squared distance between each data point and its corresponding cluster centroid. A lower WCSS value indicates higher homogeneity within a cluster and thus better clustering quality.

The WCSS is calculated using the following formula:

$$WCSS = \left|\left|x - \mu_i\right|\right|^2 \quad (1)$$

Here's a breakdown of this formula:

- k represents the total number of clusters.
- C_i is the set of point in cluster i
- μ_i is the centroid of cluster i
- x is a point within cluster Ci.
- $\left|\left|x - \mu_i\right|\right|^2$ calculates the squared Euclidean distance between a point x and the centroid μ_i, which quantifies how far the point is from the centroid.

The WCSS values were then visualized through an Elbow Curve, which illustrates the relationship between the number of clusters (k) and the inertia value. Typically, the WCSS decreases significantly up to a certain point and then begins to level off. The "elbow point" on this curve indicates the optimal number of clusters—beyond this point, adding more clusters yields minimal improvement in WCSS.

To objectively determine the elbow point, the KneeLocator algorithm was employed, which automatically detects the point of significant change on the curve. The results of the analysis indicated that the optimal number of clusters lies at the identified elbow point, and this value is used as the basis for the clustering process in the subsequent stage.

In addition to using the Elbow Method, this study also applies an evaluative approach based on internal validation metrics to determine the optimal number of clusters, namely through the Silhouette Score and the Davies-Bouldin Index. These two metrics are used to assess the quality of clustering results produced by the K-Means algorithm.

The analysis is carried out by applying the K-Means algorithm to a varying number of clusters, specifically from k = 2 to k = 10. For each value of k, two metrics are calculated as follows:

1. Silhouette Score, which measures how similar a data point is to its own cluster compared to other clusters. The Silhouette Score ranges from -1 to 1, with higher values indicating better clustering

$$SC = \frac{1}{n}\sum_{i=1}^{n} \quad s(i) \quad (2)$$

performance. The formula for the Silhouette Score is [4]:

- With, $s(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}}$, $b(i) = min\ d(i,v), a(i) = \frac{1}{n_{k-1}} \Sigma_{h\ \epsilon\ Cl_k, h \neq i} d(i,h)$

- $b(i)$: minimum value of the average distance of object *i* with all objects on the cluster to *v*-th
- $a(i)$: average *i*-th object distance with all objects in a cluster

The optimal grouping occurs when the maximum silhouette coefficient (SC) minimizes the distance within the group ($a(i)$) while maximizing the distance between different groups ($b(i)$). A higher silhouette coefficient value indicates better group quality [5].

2. Davies-Bouldin Index (DBI), which calculates the average ratio of intra-cluster distances to inter-cluster distances. Lower DBI values indicate better clustering, as they reflect more compact and

$$DBI = \frac{1}{K}\sum_{k=1}^{K} \quad R_k \tag{3}$$

well-separated clusters. The formula for the Davies-Bouldin Index is [6]

- With, $R_k = max_{k\neq v}\left(\frac{S_k + S_v}{M_{k,v}}\right)$, $M_{k,v} = d(C_k, C_v), k \neq v$, $S_k = \frac{1}{n_k}\sum_{i=1}^{n_k} \quad d(x_i, C_k), and \ S_v = \frac{1}{n_v}\sum_{i=1}^{n_v} \quad d(x_i, C_v), k \neq v$
- S_k : average of the *i*-th object distance with *k*-th centroid cluster
- S_v: average of the *i*-th object distance with *v*-th centroid cluster
- $M_{k,v}$: *k*-th cluster centroid distance and *v*-th cluster centroid distance

The evaluation results of these two metrics are stored in a DataFrame for further analysis. From the results, the optimal number of clusters based on the Silhouette Score is determined by the highest score, while the optimal number of clusters based on the Davies-Bouldin Index is determined by the lowest score.

The evaluation is visualized in two separate plots using Plotly: the first plot shows the trend of the Silhouette Score relative to the number of clusters, and the second plot displays the trend of the Davies-Bouldin Index. This visualization aims to facilitate the observation of clustering quality trends as the number of clusters increases, while also helping to identify the optimal point based on each metric.

The final outcome of this stage is a recommendation for the optimal number of clusters based on the evaluation of both internal metrics, which then serves as the basis for the final clustering process.

After the clustering process using the K-Means algorithm is completed, the next step is to identify the centroid values or cluster centers of each formed group. A centroid represents the average of all objects within a cluster based on the analyzed variables, namely HDI, Life Expectancy (UHH), Expected Years of Schooling (HLS), Mean Years of Schooling (RLS), and Real Expenditure per Capita (PPK).

The steps of K-Means are [7] :

a. Determining the number of *K*-clusters to be formed;
b. Randomly determine the initial cluster center (centroid);
c. Calculate the distance of each object with each centroid;
d. Grouping each object into the closest centroid, an object will become a member of the *k*-th cluster if the distance of that object to the *k*-th centroid is of the least value when compared to the distance to other centroids;
e. Determine the new centroid by calculating the average of the objects on each cluster using the following Equation:

$$C_{kj} = \frac{1}{n_k}\sum_{i=1}^{n_k} \quad x_{ij} \tag{4}$$

With $k = 1,2,3,\ldots,K; j = 1,2,3,\ldots,p$; C_{kj} is the centroid of the *k*-th cluster of the *j*-th variable; n_k is the number of objects on the *k*-th cluster; and x_{ij} is the value on the *i*-th object on the *j*-th variable;

f. Repeat steps c through e until none of the members of each cluster have changed.

However, in this study, the centroid values were calculated automatically by the K-Means algorithm and then displayed in a table for each cluster. The table reflects the general characteristics or average profiles of each cluster, making it useful for interpreting the differences in human development levels among the grouped regions.

Subsequently, the distance between each data point and the centroid of its assigned cluster was calculated using the Euclidean metric. This calculation indicates how close or far a data point is from its cluster center. Formula equation for calculating Euclidean distance in Equation 5 [8].

$$d_{euc}(x_i, C_k) = \sqrt{\sum_{j=1}^{p} (x_{ij}, C_{kj})^2}, j = 1,2,3,\ldots,p \quad (5)$$
$$K = 2,3,4,\ldots$$

where $d_{euc}(x_i, C_k)$ is the Euclidean distance between the i-th object, the j-th variable to the center of the cluster (centroid) k-th on the j-th variable; $k = 1,2,\ldots,K$; x_{ij} is the value of the i-th object on the j-th variable; C_{kj} is the center of the k-th centroid on the j-th variable; p is the number of observed variables; and K is the number of clusters.

The minimum distance was then added as a new column in the dataset, which can be used to assess the degree of closeness (homogeneity) of each region to the characteristics of its respective cluster.

The next step aims to identify the data point that is closest to the centroid of each cluster. For every cluster formed, one data point with the smallest distance to its centroid is selected. This data point is considered the most representative or the one that best reflects the average characteristics of its cluster. The results are then presented as a profile of the regions that most accurately represent each cluster.

III. RESULT & DISCUSSION

A. Determination of the Optimal Number of Clusters

The first step is to determine the optimal number of clusters (k) to be applied to the K-Means algorithm. The evaluation is performed using three popular methods, namely Elbow Method, Silhouette Coefficient, and Davies-Bouldin Index.

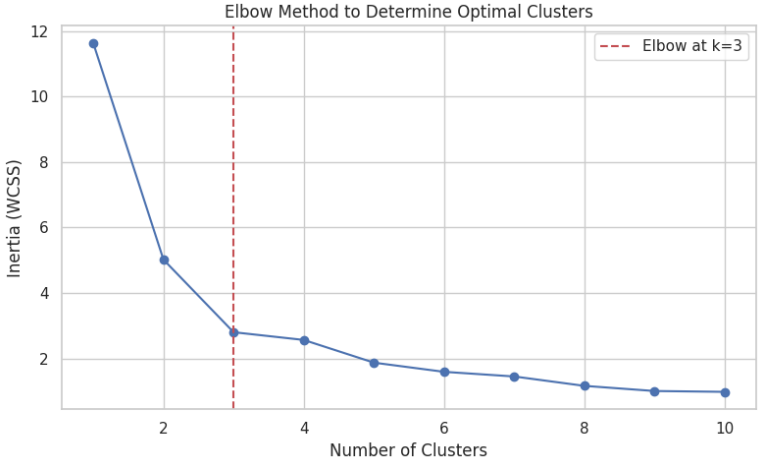

Figure 1. Elbow Method to Determine Optimal Clusters

The results of the Elbow method showed a significant change in the inertia value (within-cluster sum of squares) when the number of clusters increased from 2 to 3. After k = 3, the decrease in the inertia value becomes sloping, which indicates that the optimal number of clusters is at the elbow point.

Evaluation of the optimal number of clusters is done by comparing two evaluation metrics, namely Silhouette Score and Davies-Bouldin Index (DBI). The visualization of these two metrics is shown in the following Figure.

Figure 2. Optimal Number of Clusters Evaluation Chart

TABLE 1. Optimal Number of Clusters Evaluation

Number of Cluster (k)	Silhouette Score	Davies-Bouldin Index	Optimal (Silhouette)	Optimal (DB Index)
2	0.4680	0.7686	True	False
3	0.4436	0.7556	False	True
4	0.3971	0.7862	False	False
5	0.3402	0.8472	False	False
6	0.2996	0.9490	False	False
7	0.2823	1.0159	False	False
8	0.2829	0.9568	False	False
9	0.2923	0.9218	False	False
10	0.2595	0.9690	False	False

Silhouette Score is used to measure how well each object is in its own cluster. Higher values indicate that the objects in the cluster have strong similarities and are well separated from the rest of the cluster. From the table, it can be seen that the highest Silhouette Score value is at cluster number 2, which is 0.4680. Thereafter, the Silhouette Score consistently decreases, indicating poorer cohesion and separation between clusters. Therefore, from the Silhouette Score point of view, k = 2 is the most optimal number of clusters.

Otherwise, DBI evaluates the average ratio of intra-cluster to inter-cluster distances, where lower values are better as they indicate more compact and well-separated clusters. Based on the graph, the lowest DBI value is at cluster number 3, which is 0.7556. This means that in terms of cluster compactness and separability, the selection of k = 3 is more optimal than the other number of clusters.

Although the Silhouette Score value is highest at k = 2, this approach results in clustering that is too general and does not capture enough variation in the data. In contrast, the number of clusters k = 3 provides a balance between cluster performance (lowest DBI) and a more representative data structure. Therefore, the number of clusters used in this analysis is 3, as it provides structurally optimal clustering results and is more informative for interpretation of regional development policies.

TABLE 2. Centroid Value

IPM	UHH	HLS	RLS	PPK	Cluster
0.887312	0.870387	0.781168	0.842817	0.633941	0
0.450884	0.593217	0.349831	0.467181	0.262435	1
0.177642	0.170562	0.264509	0.193141	0.084750	2

After determining the most optimal number of clusters using evaluation methods such as the Elbow Method or Silhouette Score, the next step is to calculate the centroid value of each cluster. In the K-Means algorithm, the initial centroid is determined randomly, but it will continue to be updated based on the average value of all cluster members until convergence is achieved. Table 2 displays the final centroid values of each cluster formed based on five regional development indicators.

Each value in the table represents the standardized average of each indicator in each cluster. Cluster

0 shows the highest values across all indicators, including HDI (0.887), UHH (0.870), HLS (0.781), RLS (0.842), and PPK (0.633). The higher the cluster centroid value, the stronger the relationship or correlation.

TABLE 3. Closest Data to the Centroid of each Cluster

Regency/City	IPM	UHH	HLS	RLS	PPK	Cluster	Distance_to_Centroid
Sidoarjo	82.67	75.63	15.22	10.91	15710	0	0.079674
Madiun	74.81	74.79	13.27	8.20	12668	1	0.061352
Sumenep	69.78	73.86	13.59	6.10	10156	2	0.174919

Furthermore, calculating the distance between each data point and its cluster centroid yields information about the representative of each cluster. The data point with the closest distance to the centroid is considered the ideal or typical characteristic of the cluster. For example, Sidoarjo district is the center of cluster 0, Madiun district is the center of cluster 1, and Sumenep district represents cluster 2.

B. Clustering Results

The clustering process was conducted on data from 38 districts/cities in East Java using the K-Means algorithm with five main variables: Human Development Index (HDI), Life Expectancy (UHH), Expected Years of Schooling (HLS), Average Years of Schooling (RLS), and Real Expenditure Per Capita (PPK). Based on the evaluation results of the Silhouette Score and Davies-Bouldin Index metrics, the most optimal number of clusters was determined to be 3 clusters. The clustering results produced three main groups with significantly different characteristics.

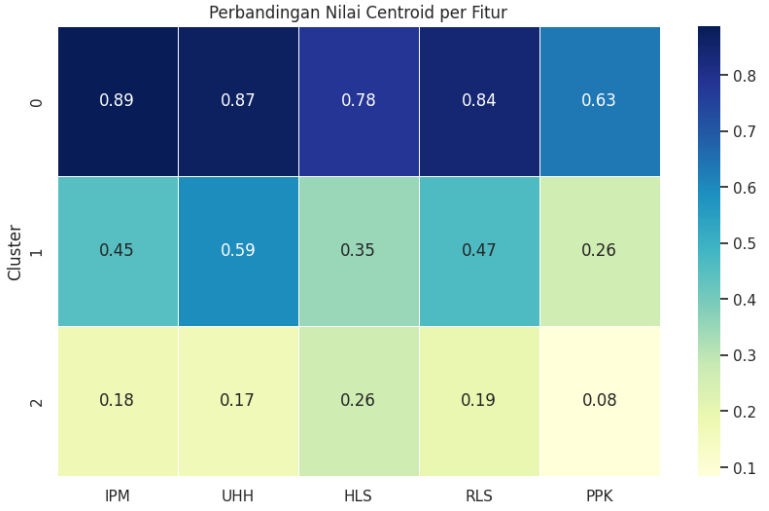

Figure 3. Comparison of Centroid Values per Feature

Visualizing the clustering results through centroid heatmaps illustrates the stark differences between clusters in terms of the five indicators. The heatmap clearly illustrates the distinct characteristics of each cluster. These clear differences in centroid values demonstrate that the clustering method has effectively grouped the regions into distinct social, economic, and educational profiles. For example, cluster 0 shows high dominance on all variables, while cluster 2 tends to consistently score low on almost all dimensions.

TABLE 4. Cluster 0

Kabupaten/Kota	IPM	UHH	HLS	RLS	PPK	Cluster	Distance_to_Centroid
Kabupaten Sidoarjo	82.67	75.63	15.22	10.91	15710	0	0.079674
Kota Kediri	81.88	75.94	15.71	10.92	13670	0	0.330287
Kota Blitar	81.44	75.20	14.81	10.82	14933	0	0.222384
Kota Malang	84.68	75.54	15.79	11.14	17791	0	0.306902
Kota Mojokerto	81.76	75.99	14.13	11.38	14842	0	0.285217

Kota Madiun	84.51	75.67	14.54	12.11	17518	0	0.263263
Kota Surabaya	84.69	76.02	14.87	10.89	19666	0	0.405311
Kota Batu	79.69	75.36	14.58	9.87	14253	0	0.330429

Cluster 0 contains cities with very high HDI values, such as Surabaya City, Malang City, and Madiun City. This cluster is characterized by a combination of high HLS and RLS values, as well as a PPK level that is far above the provincial average. This means that areas in this cluster are already at a more mature stage of human development, with the support of education and relatively equitable economic welfare.

TABLE 5. Cluster 1

Kabupaten/Kota	IPM	UHH	HLS	RLS	PPK	Cluster	Distance_to_Centroid
Kabupaten Pacitan	71.49	74.74	12.69	7.90	10099	1	0.350483
Kabupaten Ponorogo	73.70	75.28	13.78	7.80	11065	1	0.246675
Kabupaten Trenggalek	72.47	75.35	12.63	7.92	10872	1	0.319156
Kabupaten Tulungagung	75.13	75.20	13.36	8.68	11966	1	0.122640
Kabupaten Blitar	73.44	75.32	12.67	7.87	12020	1	0.250347
Kabupaten Kediri	75.18	75.07	13.63	8.26	12388	1	0.103480
Kabupaten Malang	73.53	75.34	13.49	7.80	11190	1	0.229053
Kabupaten Banyuwangi	74.30	74.13	13.14	7.78	13320	1	0.321564
Kabupaten Pasuruan	72.36	74.61	12.78	7.46	11617	1	0.271410
Kabupaten Mojokerto	76.69	74.95	12.99	9.13	13903	1	0.232077
Kabupaten Jombang	75.67	74.64	13.61	8.78	12454	1	0.149536
Kabupaten Nganjuk	75.24	74.64	13.18	8.25	13376	1	0.150840
Kabupaten Madiun	74.81	74.79	13.27	8.20	12668	1	0.061352
Kabupaten Magetan	76.77	75.42	14.08	8.69	12915	1	0.302945
Kabupaten Ngawi	73.91	75.21	12.89	7.84	12414	1	0.179200
Kabupaten Bojonegoro	72.75	74.91	13.18	7.59	11204	1	0.201773
Kabupaten Tuban	72.31	74.95	12.54	7.53	11579	1	0.285345
Kabupaten Lamongan	75.90	75.07	14.03	8.48	12419	1	0.206099
Kabupaten Gresik	78.93	74.48	13.98	10.03	14356	1	0.453060
Kota Probolinggo	77.79	74.31	13.98	9.72	13405	1	0.395266
Kota Pasuruan	78.90	74.86	13.67	9.94	14664	1	0.405657

Cluster 1 includes cities/districts with medium human development performance. Some of these include Kediri District, Jombang District, and Magetan District. The average HDI in this cluster is in the range of 74-76, with UHH and PPK being quite stable but not as high as cluster 0. This cluster can be considered as an area that is heading towards an advanced stage of development and has the potential to develop better with the right policy interventions.

TABLE 6. Cluster 2

Kabupaten/Kota	IPM	UHH	HLS	RLS	PPK	Cluster	Distance_to_Centroid
Kabupaten Lumajang	70.31	74.57	12.41	7.27	10124	2	0.355918
Kabupaten Jember	70.93	74.17	13.50	6.54	10700	2	0.207617
Kabupaten Bondowoso	71.22	73.31	13.33	6.53	11689	2	0.233093
Kabupaten Situbondo	71.22	73.36	13.20	6.93	11216	2	0.200198
Kabupaten Probolinggo	70.85	73.93	12.64	6.31	12258	2	0.205549
Kabupaten Bangkalan	67.33	73.43	11.98	6.01	9841	2	0.341276
Kabupaten Sampang	66.72	73.66	12.55	5.08	9782	2	0.301595
Kabupaten Pamekasan	70.85	73.66	13.69	7.17	9811	2	0.236539
Kabupaten Sumenep	69.78	73.86	13.59	6.10	10156	2	0.174919

Cluster 2 includes areas with relatively low HDI values, such as Sampang, Bangkalan and Sumenep districts. This cluster is characterized by low HLS and RLS, and smaller per capita expenditure than the other two clusters. This situation shows that there is still a development gap in the horseshoe area and Madura islands that requires more intensive attention and development interventions

C. Correlation between the Human Development Index (HDI) and other indicators

The correlation between the Human Development Index (HDI) and supporting indicators such as Life Expectancy (LE), Expected Years of Education (EYE), Average Years of Education (AYE), and Real Per Capita Expenditure (RPE) is the main focus of the analysis to understand the relationship between human development variables. This analysis helps identify the most influential factors affecting the HDI in each regional cluster, thereby providing a foundation for formulating more effective and targeted development policies.

TABLE 7. Correlation of HDI with Another Indicators

UHH	HLS	RLS	PPK	Cluster
0.375253	0.295495	0.677105	0.877364	0
-0.255590	0.744576	0.937618	0.890847	1
0.151840	0.637073	0.787336	0.611912	2

The correlation table shows that the relationship between HDI and each indicator varies across clusters. For example, the correlation between HDI and UHH appears positive in Clusters 0 and 2 (0.375 and 0.152), but negative in Cluster 1 (-0.256), indicating that in Cluster 1, life expectancy is not in line with HDI improvement. Meanwhile, the correlation between HDI and HLS, RLS, and PPK tends to be positive and quite strong, especially for RLS and PPK, which show correlations above 0.6 for all clusters, indicating a significant contribution of education and per capita expenditure to HDI improvement.

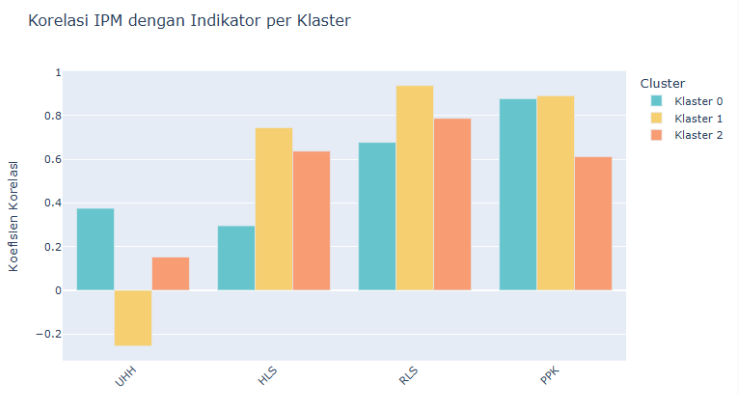

Figure 4. Correlation of Human Development Index (HDI) with Other Indicators per Cluster

The bar chart visualization reinforces the findings in the previous correlation table by presenting a visual comparison between clusters, where PPK and RLS are the most dominant indicators contributing to the Human Development Index (HDI) in all clusters. Overall, this chart makes it easy to quickly and intuitively identify the most influential indicators in each cluster.

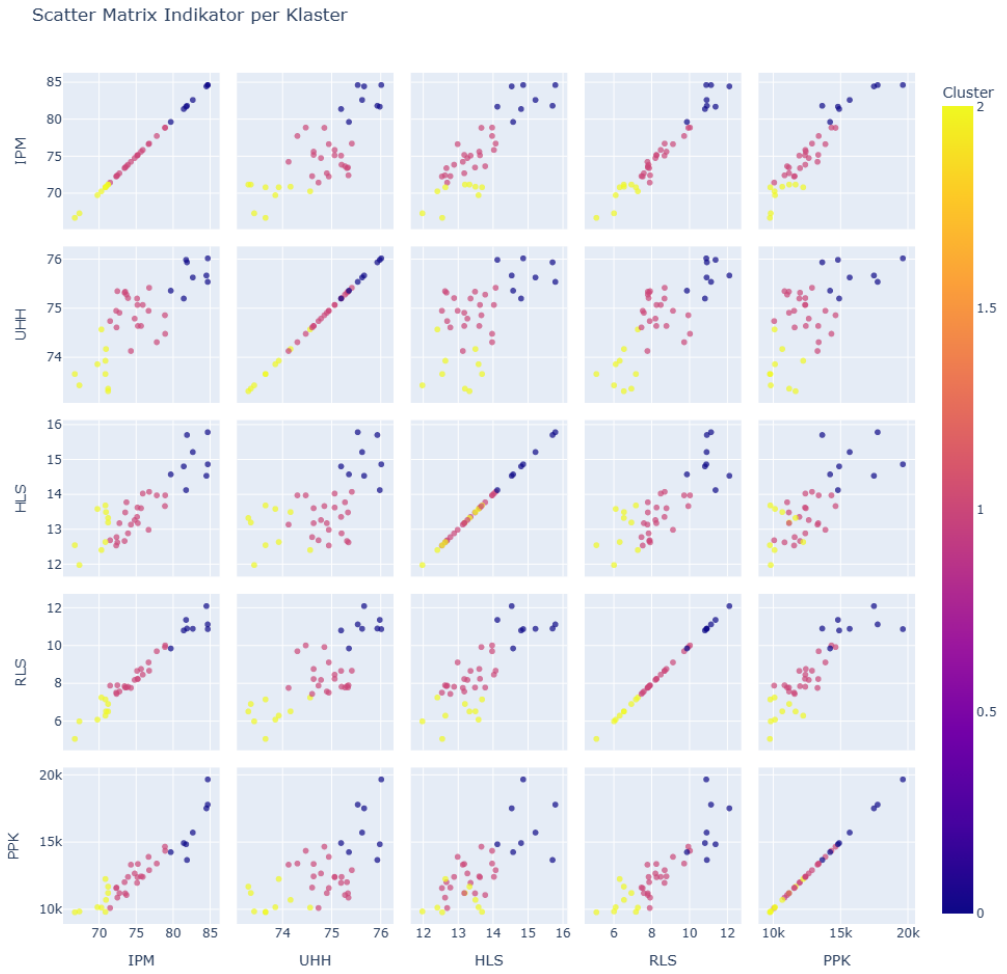

Figure 5. Scatter Matrix Indicators per Cluster

Figure 5. shows the results of visualization using a scatter matrix, which displays the relationship between indicators simultaneously for each cluster, with colored dots representing Cluster 0 (dark blue), Cluster 1 (pink), and Cluster 2 (yellow). In the scatter diagram mapping the relationship between HDI and RLS and PPK, a very clear upward trend is observed across all clusters, indicating a strong positive correlation between HDI and average years of education and per capita expenditure. Cluster 1 (pink) shows a data distribution with higher HDI and education indicator values compared to other clusters, reinforcing the dominance of this positive relationship.

Conversely, in the scatter diagram of HDI and UHH, a more complex and varied pattern is observed across all clusters. Cluster 1 shows a more scattered distribution without a clear upward trend, consistent with the previously identified negative correlation. Meanwhile, clusters 0 and 2 show a moderate upward trend. The relationship between education indicators (HLS and RLS) also shows a strong positive correlation across clusters, with points tending to form a diagonal line.

This scatter matrix also visually illustrates the differences in characteristics between each cluster, where each cluster has a different range of values and patterns of relationships between indicators. This reinforces the finding that education and per capita expenditure are the main variables positively correlated with HDI, while the relationship with life expectancy (LE) is more dynamic and dependent on cluster characteristics.

CONCLUSION

This study has successfully demonstrated the application of the K-Means clustering algorithm in

grouping districts and cities in East Java based on their 2024 Human Development Index indicators. By using supporting indicators such as Life Expectancy, Expected and Mean Years of Schooling, and Real Expenditure Per Capita, the analysis identified three clusters with significantly different development profiles. The clustering results revealed regional disparities, highlighting the need for differentiated development strategies. Cluster 0 includes highly developed urban regions, Cluster 1 represents regions with moderate development, and Cluster 2 consists of less developed areas requiring more focused intervention. The correlation analysis further indicated that education and economic indicators are the most influential factors affecting HDI. These findings can serve as a valuable input for policymakers in designing more equitable and efficient development programs tailored to each region's unique characteristics.

REFERENCES

[1] Badan Pusat Statistik, INDEKS PEMBANGUNAN MANUSIA 2024, Jakarta, 2025.
[2] A. Hanifah, A. Munawaroh, N. Husainah, S. Jamilah, S. Hartinah, S. H. Harun and M. Annas, Pengantar Ilmu Statistik, Duta Sains Indonesia, 2025.
[3] I. K. Sukesa, "CRISP DM Sebagai Salah Satu Standard untuk Menghasilkan Data Driven Decision Making yang Berkualitas", 22 Juni 2022, [Online]. Tersedia: https://www.djkn.kemenkeu.go.id/artikel/baca/15134/CRISP-DM-Sebagai-Salah-Satu-Standarduntuk-Menghasilkan-Data-Driven-Decision-Making-yang-Berkualitas.html [Diakses: 16 Juni 2025].
[4] A. I. T. Utami, F. Suryaningrum, D. Ispriyanti, "K-Means Cluster Count Optimization With Silhouette Index Validation And Davies Bouldin Index (Case Study: Coverage Of Pregnant Women, Childbirth, And Postpartum Health Services In Indonesia In 2020)", BAREKENG: Jurnal Ilmu Matematika dan Terapan, vol. 17, no. 2, pp. 0707-0716, 2023.
[5] L. Vendramin, R. Campello, and E. R. Hruschka, "On the Comparison of Relative Clustering Validity Criteria", Proceedings of the SIAM International Conference on Data Mining, vol. 3, no. 4, pp. 733-744, 2009.
[6] M. D. Kartikasari, "Self-Organizing Map Menggunakan Davies Bouldin Index dalam Pengelompokan Wilayah Indonesia Berdasarkan Konsumsi Pangan", Jambura J.Math, vol. 3, no. 2, pp. 187-196, 2021.
[7] D. T. Larose, and C. D. Larose, Discovering Knowledge in Data An Introduction to Data Mining Second Edition Wiley Series on Methods and Applications in Data Mining. New Jersey: John Wiley and Sons, Inc, 2014.
[8] R. A. Johnson, and D. W. Wichern, "Applied Multivariate Statistical Analysis", Prentice Hall, 2002.