Clustering Analysis of Socio-Economic Districts/Cities In East Java Province Using PCA And Hierarchical Clustering Methods

Rifqi Hilal Bhahari[1], Kusnawi[2]*
[1)2)]Universitas Amikom Yogyakarta, Yogyakarta, Indonesia
rifqihilal@students.amikom.ac.id , khusnawi@amikom.ac.id

Submitted :Sept 3, 2024 | **Accepted** : Sept 24, 2024 | **Published** : Oct 3, 2024

Abstract: This study aims to analyze the socio-economic conditions of districts/cities in East Java using Principal Component Analysis (PCA) and Hierarchical Clustering. Socio-economic data for 2023 from 38 districts/cities includes the percentage of poor people, regional GDP, life expectancy, average years of schooling, per capita expenditure, and unemployment rate. PCA was used to reduce the dimensionality of the data, facilitating analysis and visualization. The reduced data was then analyzed using Hierarchical Clustering to group districts based on similar socio-economic characteristics. The clustering results were evaluated with the Silhouette Index and Davies-Bouldin Index. This study identified four main clusters with different socio-economic characteristics. The best clusters have high regional GDP, life expectancy, average years of schooling, and high per capita expenditure and low unemployment rates. The worst clusters show a high percentage of poor people and high unemployment rates. These results assist the government in designing more effective policies to improve welfare in East Java.

Keywords: PCA; Hierarchical Clustering; Socio-Economic; East Java.

INTRODUCTION

Indonesia is a densely populated country with a very high population growth. The total population of Indonesia has increased every year. In line with the distribution of the population, the most densely populated area is Java Island (Rohsulina et al., 2020). East Java is one of the large provinces in Indonesia with a population of more than 39.69 million people in 2019 and in 2020 it increased to more than 40.67 million people (Fauzia et al., 2019). Having a large population, the population of East Java can be used as a capital so that it can advance development for the welfare of the community. East Java Province has the most regencies and cities on the island of Java, with 29 regencies and 9 cities. The large number of cities and regencies creates problems for the government to increase equitable development in each region for the welfare of the community. Other problems such as poverty are also still a serious challenge in some areas, adding to the complexity of efforts to improve welfare. The success of development depends on the population, quality, and availability of resources that include various socioeconomic factors (Syaputri et al., 2021). Socio-economic indicators involve aspects such as demography, health, education, housing, crime, socio-culture, and household welfare.

The East Java Provincial Government categorizes its regions by various socioeconomic indicators such as GRDP (Gross Regional Income), life expectancy, per capita expenditure, and unemployment rate using descriptive statistics methods. However, this method is not optimal because it only measures one economic indicator per region, without considering the interaction between indicators. For example, the unemployment rate, per capita income, and GRDP affect each other. Therefore, a more comprehensive and integrated approach is needed that not only looks at one indicator but also takes into account various other interrelated factors. Such an approach will provide a deeper and more accurate understanding of economic development between regions, as well as assist in the formulation of more effective and efficient policies.

In line with the Industrial Revolution 4.0, various statistical methods have been developed, one of which is cluster analysis. Cluster analysis is a statistical method that aims to group objects based on the similarity of their characteristics (Afifi et al., 2019). The clustering method is used to group regencies and cities based on the socioeconomic conditions of the region. Researchers used two clustering models, namely Principal Component Analysis (PCA) and Hierarchical Clustering. PCA is a technique used to reduce the dimensions of complex data into simpler dimensions. This will facilitate data analysis and make the data easier to understand. After performing dimension reduction using PCA, the data will be classified using Hierarchical Clustering. Hierarchical Clustering is a method used to categorize data into different groups using certain criteria. This method will group regencies and cities based on socioeconomic factors that determine the overall development of regencies and cities. 20 With

*name of corresponding author

this research, it is hoped that it can provide benefits for the government to conduct equitable development of regencies and cities that have been grouped. In addition, by detailing these differences, the government and stakeholders can more effectively allocate resources and develop programs that suit the specific needs of each region.

LITERATURE REVIEW

In this study, researchers explored various sources of information to find theories that were in accordance with the research title. In addition, researchers also refer to previous research as a basis for consideration to identify existing strengths and weaknesses. Some previous studies that were used as literature reviews in this study also helped researchers in developing appropriate analysis methods. Some previous studies that were used as literature reviews in this study include:

Research that aims to group customers based on segmentation. This research uses the Agglomerative Hierarchical Clustering (AHC) method to group customers based on RFM (Recency, Frequency, Monetary) characteristics and determine the right promotional strategy for each customer group. The results show that 7 clusters are the best results, with different promotional strategies for each type of customer. AHC can be used for customer grouping and appropriate promotional strategies. This research contributes to the development of marketing strategies based on customer segmentation and provides insight into the use of clustering methods in customer data analysis (Widyawati et al., 2020).

The research discusses the application of the Hierarchical Clustering method on time series data for clustering districts / cities in East Kalimantan Province based on population. The clustering method used is Hierarchical Clustering with all the algorithms in the agglomerative method. This study aims to obtain optimal clustering results in the process of clustering districts / cities in East Kalimantan Province based on population with time series data. The results showed that the number of representative clusters in grouping districts / cities in East Kalimantan Province is 2 clusters with the largest silhouette coefficient value of 0.68, which is included in the good cluster category. In addition, this study also displays descriptive statistics in the form of time series plots for each district / city in East Kalimantan Province, which shows the development of population every year (Tri et al., 2019).

Another study aims to cluster districts/cities based on socioeconomic factors that include aspects of demographics, health, education, housing, crime, socio-culture, and household welfare. The methods used in this research include descriptive statistics, which includes measures of data centering (mean) and measures of data dispersion (variance). In addition, this study also used the K-Means method for clustering analysis. The results of this study provide a visualization of cluster results based on a map of East Java after obtaining the optimum group and group members (Ferdiana et al., 2023).

Research using the K-Means Cluster method with Principal Component Analysis (PCA) to group districts / cities in Kalimantan based on Human Development Index indicators. The results showed the formation of 4 clusters with a Silhouette Coefficient value of 0.540, which indicates a medium cluster structure. The optimal cluster has different characteristics, where Cluster 1 has a higher economic growth rate, Cluster 2 has higher other variables, Cluster 3 has a higher percentage of poor people, and Cluster 4 has a higher open unemployment rate (Anwar et al., 2022).

Another study aimed to categorize social media users based on their duration of use using Hierarchical Clustering and Non-Hierarchical Clustering algorithms. The results showed that there were two groups of social media users, namely a group with usage within reasonable limits (Old Cluster) and a group with excessive usage (Millennial Cluster). The Complete Linkage method proved to be the best method with the optimal number of clusters as many as two clusters, with the smallest Davies-Bouldin Index (DBI) value, which is 0.42595. Social media users can be grouped based on the duration of their use, and the characteristics of each group can be identified (Alamtaha et al., 2023).

Research by Apriliana and Widodo (2023) used hierarchical cluster analysis to group provinces in Indonesia based on the number of Base Transceiver Stations (BTS), telephone signal strength, and 4G internet signal strength. The results showed the existence of 3 groups based on the three indicators used. Ward's method was chosen as the best method in hierarchical cluster analysis, with the dendrogram showing the division of objects very well (Apriliana & Widodo, 2023).

METHOD

This research was conducted by following the stages shown in the flowchart in Figure 1. The following is a flowchart for Clustering Analysis of Social-Economic District/City in East Java Province Using PCA and Hierarchical Clustering Methods.

*name of corresponding author

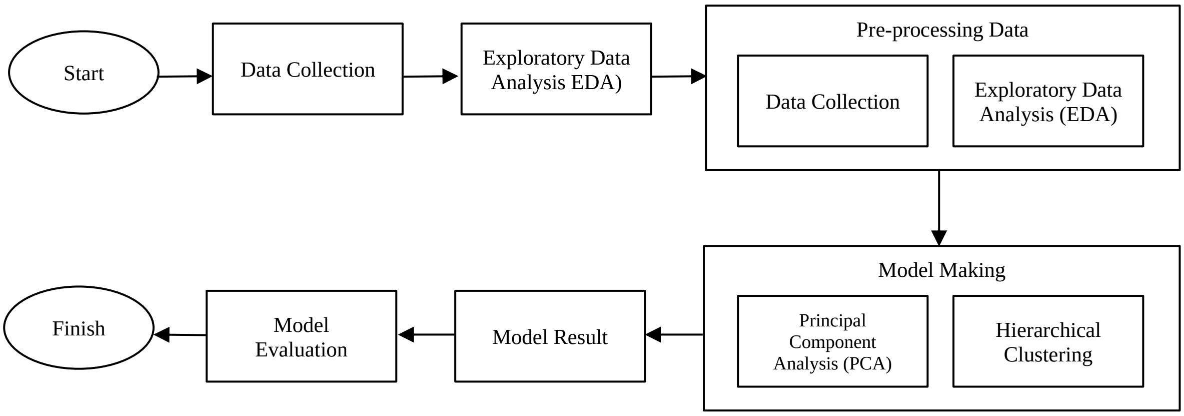


Fig 1. Research Flow

In the research listed in Figure 1, several stages will be carried out, starting with collecting secondary data obtained from the website of the East Java Central Bureau of Statistics (BPS). Then Exploratory Data Analysis (EDA) is carried out to see the distribution of data. Next, the collected data will go through a preprocessing stage, including handling outliers and scaling the data. After that, the PCA method will be used to reduce the data dimension. Then, clustering is done using the Hierarchical Clustering method by displaying the cluster results, visualization of the clustering results is also done to get a better understanding. evaluate the clustering model using Silhouette Score and DBI.

Data Collection

Data was collected using indicators that affect socioeconomic conditions in cities and districts in East Java. This study uses secondary data from the East Java Central Bureau of Statistics (BPS), covering 38 datasets consisting of 29 districts and 9 cities. This study uses five variables based on the main factors affecting socioeconomic conditions, which are shown in Table 1.

Table 1. Research Variables

Variable	Description
poorpeople_percentage	Percentage of Poor Population
reg_gdp	Gross Domestic Product (in billion)
life_exp	Life Expectancy Rate
avg_schooltime	Average Years of School
exp_percap	Expenditure per Capita
tpt	Open Unemployment Rate

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an important step in data analysis and processing that aims to understand the structure, patterns, and relationships in the data before applying more complex analysis techniques. The following is the EDA conducted in this study:

1. Basic descriptive statistics for each socio economic indicator
   a. Count: This indicates the amount of data present in each column. The dataset has 38 data points
   b. Mean: Measures the average value for each indicator. Where x is the data value and N is the number of observations.
   c. Standard Deviation: Measures how far the values in the dataset are spread out from the mean.
   d. Minimum and Maximum: Determines the smallest and largest values in the dataset.
   e. Quartiles: Divides the data into three equal parts to see the distribution of the data.
2. Use various visualization tools to understand distributions and relationships in data. Heatmap, For visualizing correlations between variables. Heatmaps help identify strong relationships (positive or negative) between variables.

Preprocessing

Preprocessing is the first step in data processing that aims to clean and prepare raw data before further analysis. This step aims to eliminate disturbances and errors in the data and ensure that the data is in a ready-to-use condition. Through preprocessing, the data will be more structured, cleaner, and ready to be processed efficiently and accurately at the next stage of analysis. The following are the stages in preprocessing carried out in this study:

*name of corresponding author

A. Outlier Handling

At this stage, outliers are handled by winsorizing, which limits extreme values to certain percentiles (Dash et al., 2023). Researchers used the 5th and 95th percentiles because they are common standards in various fields. The data was sorted first, and interpolation was used to determine the lower and upper limits, with the formula (1).

$$Px = x_i + (P - i) \times (x_{i+1} - x_i) \qquad (1)$$

Description:

Px = percentile value.
x_i = the value of i in the sorted data.
x_{i+1} = i +1 value in the sorted data.
P = calculated percentile position.
i = the rounded part of the percentile position.

B. Data Scaling (normalization)

Transforming data into a specific range or scale. The aim is to ensure that each variable has an equal contribution to the analysis. The standardization process using Standard Scaler from sklearn.preprocessing is based on calculating the mean and standard deviation of each feature. The formula for transforming each feature is shown in (2).

$$z = \frac{x - \mu}{\sigma} \qquad (2)$$

Description:

z = standardized feature values
x = original value of the feature
μ = the average (mean) of the feature
σ = the standard deviation of the feature

PCA (Principal Component Analysis)

Principal Component Analysis (PCA) is an analytical method used to simplify data through linear transformation. The main goal is to reduce the dimensionality of variables that are interrelated and have a large number, so that the data can be interpreted more easily (Johnson & Wichern, 2007). The PCA implementation steps in the research include (Mishra et al., 2017):

1. Data standardization
2. Creating a standardized data matrix
3. Calculating the covariance matrix
4. Calculating eigenvalues and eigenvectors
5. Transformation of data to principal component space

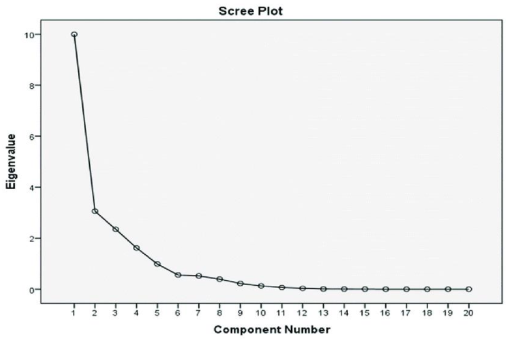


Fig 2. Scree plot example

The use of the Scree Plot in figure 2 helps in determining the number of principal components that need to be retained. The inflection point on this plot indicates the optimal number of components. Using PCA, researchers can extract structure from a set of data that has considerable dimensions, allowing for easier interpretation of the data without significantly reducing the characteristics of the original data.

Hierarchical Clustering

Hierarchical clustering is a method of grouping data that forms a tree structure with a clear hierarchy between objects. Some common methods for determining the distance between clusters are Single linkage, Complete linkage, Average linkage, Ward. In this study, focus is given to the agglomerative approach with the complete linkage method. The agglomerative approach used starts with each object as a separate cluster and gradually merges the clusters. The Complete Linkage method, also known as the furthest neighbor technique, focuses on the pair with the furthest distance from the observation value (Tri et al., 2019). The distance between clusters in this method is calculated using a formula (3).

$$d_{(UV)W} = \max\left(d_{UW}, d_{VW}\right) \qquad (3)$$

Description:
$d_{(UV)W}$ = distance between the combined cluster (UV) and cluster W
d_{UV} = distance between cluster U and cluster W
d_{VW} = distance between cluster V and cluster W

The clustering results are visualized in the form of a tree diagram known as a dendrogram. This dendrogram helps in determining the optimal number of clusters by observing the distance between merging clusters.

Cluster Result Interpretation

The results of Hierarchical Clustering are interpreted by determining the optimal number of clusters based on the dendrogram. Furthermore, interpreting the results into optimal clusters of regions in East Java and identifying each socio-economy from the resulting clusters.

Model Evaluation

In this study, two main evaluation methods are used to assess the quality of clustering results, namely the Silhouette Index (SI) and the Davies-Bouldin Index (DBI).

A. Silhouette Index (SI)
The Silhouette Index (SI) is used to validate single data, single clusters, or entire clusters. SI values range from -1 to +1, with values close to 1 indicating good clustering (Kaufman & Rousseeuw, 1990) The SI calculation involves several steps, starting with calculating the average distance of objects to other objects in the same cluster ($a(i)$), followed by calculating the average distance of objects to objects in different nearby clusters ($b(i)$). Then, the silhouette value for each point is calculated using formula (4).

$$s(i) = \frac{b(i)-a(i)}{\max\left(a(i), b(i)\right)} \qquad (4)$$

The last step is to calculate the average of all silhouette values to get the overall Silhouette Index.

B. Davies-Bouldin Index (DBI)
The Davies-Bouldin Index evaluates clusters based on the ratio between the distance within the cluster and the distance between clusters. Smaller DBI values indicate better clustering results, with a range of non-negative values (Sitompul et al., 2019) DBI is effective for comparing clustering results of different models or parameters. The formula for calculating the DBI value can be found in the equation(5).

$$\text{DBI} = \frac{1}{K}\sum_{i=1}^{K} \underset{i \neq j}{max}\left(R_{ij}\right) \qquad (5)$$

Description:
K = the number of clusters in the clustering result.
R_{ij} = the ratio of the distribution within clusters i and j to the distance between the centers of clusters i and j.
$max_{i \neq j}(R_{ij})$ = For each cluster i, find the maximum value of R_ij for all clusters j that are not equal to i.
$\sum_{i=1}^{K} \underset{i \neq j}{max}(R_{ij})$ = calculate the sum of all maximum R_ij values for each cluster i.
$\frac{1}{K}$ = divide the sum by the total number of clusters K to get the average.

RESULT

The EDA performed in this study is to calculate basic statistics for each socio-economic indicator and the following is an example of application to the poor people percentage variable:

a) Count: This indicates the amount of data present in each column. The dataset has 38 data
b) Mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (6)$$

$$\mu = \frac{13.65+9.53+10.63+6.53+8.69+\cdots+4.74+4.65+3.31}{38}$$

$$\mu = \frac{421.49}{38} = 11.09$$

c) Standard Deviation:

*name of corresponding author

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \qquad (7)$$

$$\sigma = \sqrt{\frac{(13.65-11.09)^2+(9.53-11.09)^2+\cdots+(3.31.65-11.09)^2}{38}}$$

$$\sigma = 4.50$$

d) Min = 3.31, Max = 21.76

e) Kuartil: Q1(Nilai di posisi 25%) = 7.1875, Q2 (Nilai di posisi 50%) = 9.665, Q3 (Nilai di posisi 75%) = 12.36

At this stage EDA performs baseline statistics for each socio-economic indicator, using function describe in python, results in table 2. Descriptive statistical analysis shows significant variations in socioeconomic indicators across different regions of East Java.

Table 2. Descriptive statistics results

	poorpeople_percentage	reg_gdp	life_exp	avg_schooltime	exp_percap	tpt
count	38	38	38	38	38	38
mean	10,29263158	77767,52	72,41789	8,375526316	12287,0526	4,662895
std	4,321289025	121406,2	1,982218	1,658403168	2263,2568	1,428828
min	3,31	8038,7	67,6	5,07	9363	1,71
25%	7,1875	23149,98	71,0875	7,41	10720,5	4,0825
50%	9,665	41224,85	73,21	7,925	11924,5	4,665
75%	12,36	94730,7	73,5225	9,725	13419,25	5,6
max	21,76	715294,7	74,91	11,82	18977	8,05

To obtain correlation values between variables, the corr() function from Pandas was used. This function calculates the Pearson correlation matrix which measures the linear relationship between two variables (the result ranges from -1 to 1). Visualization of correlations between variables. Heatmaps help identify strong relationships (positive or negative) between variables. The color scale on the right in figure 3 shows that dark blue represents negative correlation (close to -1), light green represents high positive correlation (close to 1), and light blue indicates low correlation or no correlation (close to 0).

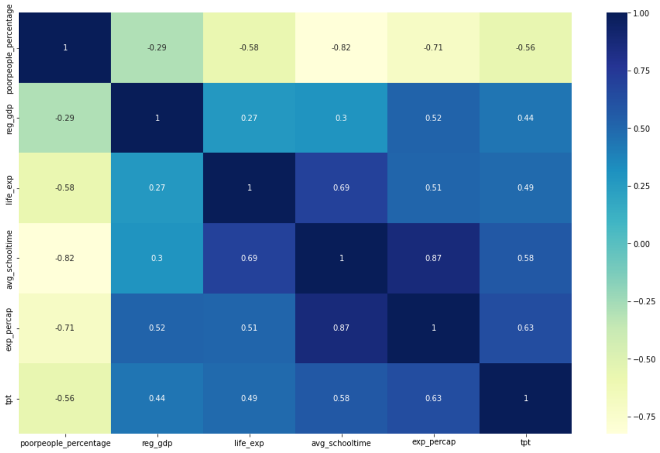


Fig 3. Heatmap of correlations between variables

After pre-processing the data, the next step is to implement PCA with machine learning. The main objective of PCA analysis is to reduce the number of variables without significantly reducing the characteristics of the original data. Determining the optimal number of components in PCA using the visualization scree plot in python

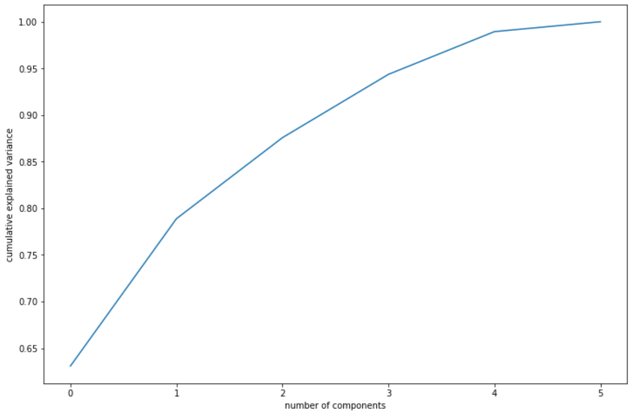


Fig 4. Result scree plot visualization to determine components

Based on Figure 4, the "elbow" point appears to be around the 4th component. This shows that after the 4th component, the addition of the main component does not provide significant improvement, so we chose the 4th component for modeling. The original data which had many variables is now reduced to only 4 variables (principal components) based on the scree plot, which covers most of the variance in the original data seen in table 3.

Table 3. PCA result frame data

	0	1	2	3
0	-1.883194	-1.327468	-0.079269	1.279451
1	-0.250056	-0.054726	0.091262	0.815876
2	-0.286557	-0.410084	-0.325206	1.191596
3	1.035111	0.258917	-0.368033	0.770489

After reducing the dimensions of the data using PCA, then continue using the Hierarchical Clustering. Using the linkage() function, derived from the scipy library and used to perform hierarchical clustering with the complete method and then displayed in the form of a denogram in Figure 5.

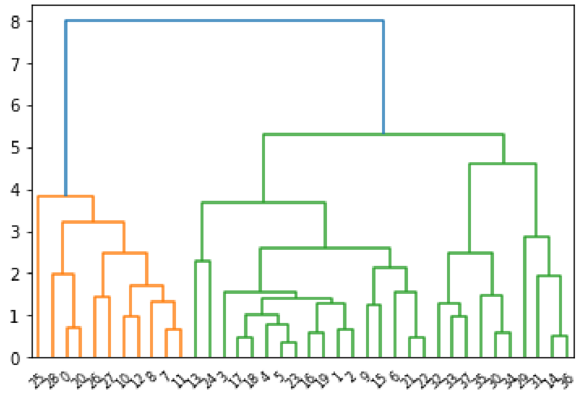


Fig 5 Denogram results for Hierarchical Clustering

Looking at the dedrogram in Figure 5, it can be seen that the cut-off point at n = 4 is the most optimal. Dendrogram analysis shows optimal division into 4 clusters.

Table 4. Results of cluster division using Hierarchical Clustering

cities_reg	**ClusterID**	**poorpeople_percentage**	**reg_gdp**	**life_exp**	**Avg_schooltime**	**exp_percap**	**tpt**
Pacitan	0	0,856224378	-0,76016	0,225063	-0,320775071	-1,2208278	-2,10448
Ponorogo	1	-0,179058431	-0,65808	0,588906	-0,385516162	-0,7544342	0,019175
Trenggalek	1	0,097352028	-0,70254	1,163672	-0,307826853	-0,8465672	-0,08588
Tulungagung	1	-0,932905136	-0,25529	1,194783	0,184205442	-0,3214568	0,762079
Blitar	1	-0,390135508	-0,30977	1,005479	-0,353145617	-0,3529634	0,206777

*name of corresponding author

In table 4 is the top 5 data of cluster division results from various cities and districts in the East Java region. There are 4 clusters from the results of denogram analysis on hierarchical clustering, Less Prosperous, Moderately Prosperous, Prosperous, and Very Prosperous. The clustering results divide East Java districts/cities into four groups with the following characteristics in Table 5.

Table 5. Clustering results and characteristics of each cluster

Cluster 0 (Less Prosperous Area Cluster)	
Cluster Member	Pacitan, Lumajang, Jember, Bondowoso, Situbondo, Probolinggo, Ngawi, Bangkalan, Sampang, Pamekasan, Sumenep
Characteristics	- These regions have relatively high poverty rates compared to other clusters. - Low GRDP, the economies of these areas are less developed compared to others. - Life Expectancy is relatively low, the health and welfare of the population in these areas may be less secure. - Average Years of Schooling is relatively low, access to and quality of education may be inadequate. - The level of individual economic well-being is lower in these areas. - Unemployment rate is low, although the unemployment rate is low, the jobs available may not be enough to provide a decent income.
Cluster 1(Moderately Prosperous Regional Cluster)	
Cluster Member	Ponorogo, Trenggalek, Tulung Agung, Blitar, Kediri, Malang, Banyuwangi, Pasuruan, Mojokerto, Jombang, Nganjuk, Madiun, Magetan, Bojonegoro, Tuban, Lamongan, Gresik
Characteristics	- Lower percentage of poor compared to Cluster 0 GRDP is quite high, the economy of these regions is quite developed. - Life Expectancy is quite high, the health and welfare of the population is quite good. - Average Years of Schooling is quite high - Expenditure per Capita is moderately high, the economic welfare of individuals is better than Cluster 0. - Unemployment rate moderate, unemployment rate more controlled than Cluster 3.
Cluster 2(Highly Prosperous Regional Cluster)	
Cluster Member	Sidoarjo, Kota Kediri, Kota Malang, Kota Surabaya
Characteristics	- The percentage of poor people is very low, these regions have the lowest poverty rates. - The GRDP is very high, the economies of these regions are highly developed and strong. - Life Expectancy is highest, the health and well-being of the population is very secure. - Highest Average Years of Schooling, access to and quality of education is excellent. - Spending per Capita is highest, the economic well-being of individuals is very high. - Low Unemployment Rate
Cluster 3(Prosperous Region Cluster)	
Cluster Member	Kota Blitar, Kota Probolinggo, Kota Pasuruan, Kota Mojokerto, Kota Madiun, Kota Batu
Characteristics	- The percentage of poor people is relatively low, these regions have a lower poverty rate than Cluster 0. - GRDP is relatively low but higher than Cluster 0, these regions have a moderately developed economy. - Life Expectancy is higher than Cluster 0 - Average Years of Schooling is relatively high, access to and quality of education is adequate. - Expenditure per Capita is quite high, the level of economic welfare of individuals is quite good. - Highest Unemployment Rate, the unemployment rate is quite high compared to other clusters.

To make it easier to visualize the distribution of clusters in districts/municipalities in East Java, a mapping was conducted, which can be seen in Figure 6.

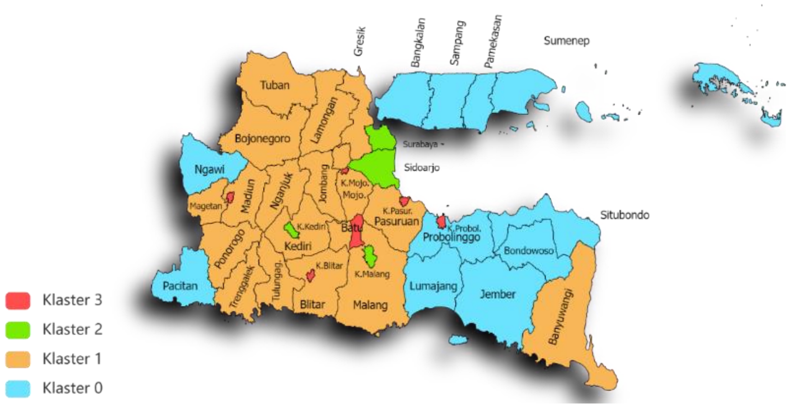
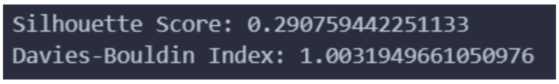


Fig 6. Visualization of cluster results based on East Java map

Using data that has been done Hierarchical Clustering before on table 4. The evaluation of the clustering model was conducted using two main metrics, namely the Silhouette Index and the Davies-Bouldin Index (DBI). These two metrics are used to measure the quality of the clusters generated by the Hierarchical Clustering model, as shown in Figure 7.



```
Silhouette Score: 0.290759442251133
Davies-Bouldin Index: 1.0031949661050976
```

Fig 7. Model evaluation results

The model evaluation used two main metrics. The Silhouette Index, with a range of -1 to 1, yielded a value of 0.2908, indicating fairly good clustering despite some overlap between clusters. This value indicates that objects within the same cluster are quite similar, but some objects may be at the boundary between clusters. The Davies-Bouldin Index, with a range ≥ 0 (no fixed upper limit), yielded a value of 1.0032. Lower values indicate better clustering, and this result indicates that the resulting clusters are reasonably good although not perfect. These two metrics show that although there are some similarities between different clusters, the clusters are still relatively well separated and compact.

DISCUSSIONS

Based on the clustering results using PCA and hierarchical clustering on socio-economic data for districts/cities in East Java in 2023, the clustering model research results are as follows:

1. Cluster Characteristics: The analysis revealed four distinct clusters representing varying levels of socio-economic development: Less Prosperous, Moderately Prosperous, Prosperous, and Highly Prosperous. This categorization provides valuable insights into regional disparities within East Java.
2. Socio-Economic Indicators: The clustering was based on six key indicators: poverty percentage, GDP, life expectancy, average school duration, per capita expenditure, and unemployment rate. These indicators effectively captured the multifaceted nature of socio-economic development in the region.
3. Regional Disparities: The results highlight significant socio-economic disparities among East Java's districts/cities. Urban areas like Surabaya, Malang, and Kediri emerged as highly prosperous, while several rural districts showed lower socio-economic indicators.
4. Model Evaluation: The Silhouette Index (0.2908) and Davies-Bouldin Index (1.0032) suggest that the clustering model performed reasonably well, though there is some overlap between clusters. This indicates that while distinct groups were identified, some regions may share characteristics across cluster boundaries.
5. Policy Implications: These findings can inform targeted policy interventions. Less prosperous regions may require focused efforts in poverty alleviation, education, and healthcare, while more prosperous areas might benefit from strategies to maintain and enhance their economic advantages.

This clustering approach offers a data-driven framework for understanding and addressing regional socio-economic disparities in East Java, potentially serving as a model for similar analyses in other regions.

CONCLUSION

This research successfully categorizes regencies and cities in East Java Province based on socio-economic indicators using PCA and Hierarchical Clustering methods. PCA effectively reduces the dimensions of complex data, while Hierarchical Clustering produces regional groupings with different socioeconomic characteristics. Model evaluation using the Silhouette Index (0.2908) and Davies-Bouldin Index (1.0032) showed reasonably good clustering results, although there is room for improvement. These results provide valuable insights for the government in formulating more targeted policies. For further development, it is recommended to conduct research with more complete data and a longer time period, as well as considering alternative clustering methods. Improving the quality and accuracy of socioeconomic data is also important to produce more reliable analysis as a basis for decision-making.

REFERENCES

Afifi, A., May, S., Donatello, R. A., & Clark, V. A. (2019). *Practical Multivariate Analysis*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315203737

Alamtaha, Z., Djakaria, I., Yahya, N. I., Matematika, J., & Mipa, F. (2023). Implementasi Algoritma Hierarchical Clustering dan Non-Hierarchical Clustering untuk Pengelompokkan Pengguna Media Sosial. *Estimasi: Journal of Statistics and Its Application*, *4*(1), 2721–379. https://doi.org/10.20956/ejsa.vi.24830

Anwar, K., Goejantoro, R., & Prangga, S. (2022). Pengelompokan Kabupaten/Kota Di Pulau Kalimantan Berdasarkan Indikator Indeks Pembangunan Manusia Tahun 2020 Menggunakan Optimasi K-Means Cluster Dengan Principle Component Analysis (PCA). *EKSPONENSIAL*, *13*(2), 131. https://doi.org/10.30872/eksponensial.v13i2.1053

Apriliana, T., & Widodo, E. (2023). Analisis Cluster Hierarki untuk Pengelompokan Provinsi di Indonesia berdasarkan Jumlah Base Transceiver Station dan Kekuatan Sinyal. *KONSTELASI: Konvergensi Teknologi Dan Sistem Informasi*, *3*(2), 286–296. https://doi.org/10.24002/konstelasi.v3i2.7143

Dash, Ch. S. K., Behera, A. K., Dehuri, S., & Ghosh, A. (2023). An outliers detection and elimination framework in classification task of data mining. *Decision Analytics Journal*, *6*, 100164. https://doi.org/10.1016/j.dajour.2023.100164

Fauzia, A. N., Muslim, I., & Karimi, K. (2019). PENGARUH FAKTOR SOSIAL EKONOMI TERHADAP FERTILITAS DI DESA SINAR GADING KECAMATAN TABIR SELATAN KABUPATEN MERANGIN JAMBI. *Abstract of Undergraduate Research, Faculty of Economics, Bung Hatta University*, *Vol. 15 No. 3 (2019): KUMPULAN SUMMARY EXECUTIVE MAHASISWA PRODI EP WISUDA KE 72 OKTOBER 2019*.

Ferdiana, K., Agam Saputri, V., & Irhamah. (2023). *Nomor 2, Februari 2023 Analisis Clustering Kabupaten… | Ferdiana, K; Saputri, VA* (Vol. 2).

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis* (6th, illustrated ed., Vol. 0131877151). Pearson Prentice Hall.

Kaufman, L., & Rousseeuw, P. (1990). Finding Groups in Data: An Introduction To Cluster Analysis. In *Wiley, New York. ISBN 0-471-87876-6*. https://doi.org/10.2307/2532178

Mishra, S., Sarkar, U., Taraphder, S., Datta, S., Swain, D., Saikhom, R., Panda, S., & Laishram, M. (2017). Principal Component Analysis. *International Journal of Livestock Research*, 1. https://doi.org/10.5455/ijlr.20170415115235

Rohsulina, P., Rahman, K., & Hidayat, A. (2020). CARRYING CAPACITY OF AGRICULTURAL LAND IN MOJOLABAN SUBDISTRICT, SUKOHARJO-CENTRAL JAVA. In *Journal of Geography Science and Education* (Vol. 2, Issue 1).

Sitompul, B., Sitompul, O., & Sihombing, P. (2019). Enhancement Clustering Evaluation Result of Davies-Bouldin Index with Determining Initial Centroid of K-Means Algorithm. *Journal of Physics: Conference Series*, *1235*, 012015. https://doi.org/10.1088/1742-6596/1235/1/012015

Syaputri, D., Noprita, P. H., & Romelah, S. (2021). Implementasi Algoritma K-Means untuk Pengelompokan Distribusi Sosial Ekonomi Masyarakat Berdasarkan Demografi Kependudukan. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *1*(1), 1–6. https://doi.org/10.57152/malcom.v1i1.5

Tri, A., Dani, R., Wahyuningsih, S., & Rizki, N. A. (2019). Penerapan Hierarchical Clustering Metode Agglomerative pada Data Runtun Waktu. *Jambura Journal of Mathematics*, *1*. http://ejurnal.ung.ac.id/index.php/jjom,P-

Widyawati, W., Saptomo, W. L. Y., & Utami, Y. R. W. (2020). Penerapan Agglomerative Hierarchical Clustering Untuk Segmentasi Pelanggan. *Jurnal Ilmiah SINUS*, *18*(1), 75. https://doi.org/10.30646/sinus.v18i1.448