Sainmatika: Jurnal Ilmiah Matematika dan Ilmu Pengetahuan Alam
Volume 22 No. 2, December 2025 DOI 10.31851/sainmatika.v22i2.20168
https://jurnal.univpgri-palembang.ac.id/index.php/sainmatika

Clustering Optimization Using K-Means with Principal Component Analysis and Mean-Median Based Centroid Initialization

Aisyah Rahma Putri, Lita Wulandari Aeli*, Ramdhan Fazrianto Suwarman, Andi Daniah Pahrany
*e-mail: lita.wulandariaeli.fmipa@um.ac.id

Mathematics Study Program, Faculty of Mathematics and Natural Sciences, State University of Malang

ABSTRACT

East Java is the province with the second largest population in Indonesia, which results in major challenges in solving poverty and unemployment, so clustering of districts/cities is needed to identify areas that need more attention, especially in certain aspects of welfare. This study uses data on community welfare indicators in East Java Province and the K-Means clustering algorithm combined with PCA and mean-median centroid initialization. From this method, six clusters were formed, with the main focus on increasing economic growth, decreasing unemployment, and improving the quality of life, which in turn can reduce poverty, thus improving overall welfare. The results show that the application of PCA to K-Means is able to improve the quality of clustering results, while centroid initialization with the mean median value can accelerate the convergence process of K-Means, where the PCA formed produces two principal components with a cumulative percentage of variance of 58%. The clustering evaluation resulted in a Silhouette Coefficient value of 0.505 and DBI of 0.601 with 7 iterations.

Keywords: Initial Centroid, K-Means, Mean, Median, PCA, Welfare

INTRODUCTION

Community welfare refers to a condition in which the basic needs of the population, such as clothing, food, shelter, education and healthcare-are fulfilled. This is a condition that must be realized for all citizens, because including one of mandates outlined in the preamble of the 1945 Constitution of the Republic of Indonesia. Community welfare has always been a primary responsibility of the state, aimed at reducing economic disparities among the population. Therefore, targeted, integrated and sustainable efforts from the central government, local governments, and the community are essential to achieving this goal.

East Java is province with the second largest population in Indonesia, following West Java. According to a survey in East Java conducted by *Litbang Kompas* in June 2024, revealed that public perceptions of the local government's performance in addressing poverty and creating employment opportunities tend to be negative. The survey also revealed that welfare issues, especially in poverty and unemployment levels, are among the top priorities set agendas by the East Java Provincial Government for the next five years. Given its status as the second most populous province, improving community welfare in East Java is expected to have a significant impact on national welfare.

Community welfare is a complex and multidimensional concept which includes some aspects of life. Consequently, a clustering algorithm is required to identify different groups

p-ISSN 1829 586X
e-ISSN 2581-0170

within the population based on this welfare indicators. Clustering is used to group data based on similarity (Lestari et al., 2020). Several research has explored districts/cities clustering based on welfare indicators, but most have covered only five to six dimensions (Pratama, 2020). Therefore, to provide more comprehensive insights into the state of community welfare, this study incorporates 13 welfare dimensions, including poverty levels, unemployment, economy, health, education conditions, social, and basic needs. Area with low economic and social conditions are at high risk in community welfare (Aeli et al., 2022). This approach allows for the identification of priority areas, particularly those requiring improvement in specific welfare aspects.

This study applies K-Means clustering to districts/cities in East Java. K-Means clustering algorithm is an unsupervised learning algorithm that partitions and maps each data to some different clusters based on similarity. Historically, K-Means has been widely applied due to its ease of implementation, efficiency in handling large datasets, and its ability to produce high-quality clustering results within a relatively short time (Meiriza et al., 2023). However, the standard K-Means algorithm has weaknesses, as mentioned by Zubair et al. (2024), namely that the main weakness of this algorithm lies in the random initialization of the centroid. Inefficient centroid selection can result in a high number of iterations, which ultimately increases the time required to achieve convergent results.

Previous research by Rosyada and Utari (2024), applying dimensionality reduction through principal component analysis (PCA) in K-Means can address multicollinearity and reduce data complexity, thereby enhancing clustering quality. In addition, Umargono et al. (2020) found that initializing centroids using the mean values effectively represents the overall data distribution, as the mean reflects the average of all values. This makes the mean a suitable choice for depicting the data's center of gravity, enabling the initial centroid position to be close to the ideal final position. Furthermore, Zubair et al. (2022) reported that combining PCA-based dimensionality reduction with centroid initialization can further improve clustering performance.

Research by Rosyada & Utari (2024) shows that applying dimension reduction with PCA on K-Means can overcome multicollinearity and reduce data complexity. However, this study did not consider other weaknesses of K-Means, such as sensitivity to initial centroid selection and the presence of outliers. On the other hand, research by Umargono et al. (2020) combined the Elbow method and the K-Means algorithm with centroid initialization based on the mean and median. The results revealed that determining the centroid based on the mean and median values was proven to reduce the number of iterations by up to 22.58% compared to random initialization, as well as producing more stable and faster converging clusters.

This research combines PCA with the K-Means algorithm, employing centroid initialization based on mean-median average values. PCA is used to mitigate multicollinearity among variables, as multicollinearity can affect clustering accuracy and quality (Budiman et al., 2024). Meanwhile, centroid initialization aims to reduce K-Means' sensitivity to initial centroid selection. The mean-median average values are chosen as initial centroids for their respective strengths: the median is resistant to outliers as it depends only on the middle value, while the mean reflects the entire data distribution, requiring fewer adjustments to achieve the ideal

p-ISSN 1829 586X
e-ISSN 2581-0170

cluster centroid (Umargono et al., 2020). K-Means, PCA, and centroid initialization methods each have their own advantages and disadvantages, so combining them is expected to complement each other. PCA plays a role in reducing dimensions and overcoming multicollinearity, while the mean median approach to centroid initialization helps improve stability and resistance to outliers.

This research aims to assess the impact of applying PCA to K-Means and using mean-median average initial centroids on improving cluster quality and the efficiency of convergence in K-Means. By producing high-quality clusters, the mapping of districts/cities in East Java based on their welfare characteristics will enable the identification of areas requiring greater attention, particularly in specific welfare aspects. The resulting map is expected to provide more targeted support to the East Java Provincial Government in its welfare development planning.

MATERIAL AND METHOD

1. Descriptive Statistics

a. Mean (μ)

The mean (average) was calculated by summing the data points and dividing by the number of observations (Bluman, 2019).

$$\mu_i = \frac{\sum_{l=1}^{n} x_{i,l}}{N}$$

b. Median (MD)

The median represents the midpoint of the ordered dataset (Bluman, 2019).

For odd number of data

$$MD_i = x_{\frac{N+1}{2}}$$

For even number of data

$$MD_i = \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2}$$

c. Standard Deviation (σ)

Standard Deviation was a measure how far the data in a set is spread from the mean (Bluman, 2019).

$$\sigma_i = \sqrt{\sigma_i{}^2} = \sqrt{\frac{\sum_{l=1}^{n}(x_{i,l} - \mu_i)^2}{N}}$$

2. Outlier Detection

Outliers were detected using boxplots and the interquartile range (IQR) method.

$$IQR = Q_3 - Q_1$$

3. Assumption Test

a. Test data feasibility with Kaiser-Meyer-Olkin (KMO) and Measure Sampling Adequacy (MSA)

Kaiser-Meyer-Olkin (KMO) were used to assess data sufficiency in overall for PCA analysis (Pendi, 2021). The KMO values ranges from 0 to 1, with values closer to 1 indicate sufficient data for PCA. The hypothesis (Haumahu & Lewaherilla, 2020):

H_0: The data are sufficient for PCA analysis

H_1: The data are not sufficient for PCA analysis

KMO formula is as following (Rosyada & Utari, 2024):

$$KMO = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r^2{}_{ij}}{\sum_{i=1}^{p}\sum_{j=1}^{p} r^2{}_{ij} + \sum_{i=1}^{p}\sum_{j=1}^{p} a^2{}_{ij}}$$

Reject H_0 if the resulting KMO value is less than 0.5 (Haumahu & Lewaherilla, 2020).

Measure Sampling Adequacy (MSA) test data adequacy on each variable. The hypothesis is (Haumahu & Lewaherilla, 2020):

H_0: The variables are sufficient for PCA analysis

H_1: The variables are not sufficient for PCA analysis

MSA statistical test is as following:

$$MSA_i = \frac{\sum_{j=1}^{p} r^2{}_{ij}}{\sum_{j=1}^{p} r^2{}_{ij} + \sum_{j=1}^{p} a^2{}_{ij}}$$

Reject H_0 if the resulting MSA value is less than 0.5 (Haumahu & Lewaherilla, 2020).

b. Correlation test between variable with Bartlett's Test of Sphericity

Bartlett's Test of Sphericity whether there is correlation among variable in the data. The hypothesis between variable with Bartlett used (Taufik et al., 2023):

H_0: The correlation matrix is an identity matrix

H_1: Correlation matrix is not an identity matrix

If the value significance (p-value) < 0.05, then H_0 rejected, meaning the variables are correlated and suitable for PCA (Taufik et al., 2023).

c. Multicollinearity test with Variance Inflation Factor (VIF)

Multicollinearity refers to a condition which two or more independent variables exhibit a high degree of correlation (Rosyada & Utari, 2024). For detect existence symptom multicollinearity can use VIF value. A VIF value greater than 10 indicates a serious multicollinearity serious, which requires appropriate corrective measures (Ann et al., 2013). The VIF value is calculated with following equation (Ann et al., 2013):

$$VIF_i = \frac{1}{1 - R^2{}_i}, \qquad i = 1, \dots, p$$

4. Principal Component Analysis (PCA)

PCA is a method for reducing the dimensions original variables that are highly correlated into principal components. This PCA method tries to summarize variation value or difference value from high correlated of variables mutual origin, to some interrelated components independent. The components formed are expected to represent important information from data. Reduction Dimensions using PCA with stages as following (Rais et al., 2021):

a. Standardize data with this equation:

$$\boldsymbol{Z}_{i,l} = \frac{x_{i,l} - \mu_i}{\sigma_i}$$

b. Covariance matrix

Covariance matrix is used to examine the direction and strength of the linear relationship between two variables. However, the size covariance did not give information about how much strong connection between two variables, because covariance value got from results calculations involving variable in different scales. Here is the formula for count covariance matrix (Johnson & Wichern, 2007):

$$Cov(X_i, X_j) = \frac{\sum(x_{i,l} - \mu_i)(x_{j,l} - \mu_j)}{N}$$

c. Matrix correlation

Coefficient correlation show direction connection between two variables and how much strong connection said. If the value correlation approaching 1, this indicates that There is connection very strong positive between second variable. This means that when one variable increases, other variables also tend to increase. And if value correlation approaching -1, this indicates that there is connection very strong negative. Subsequent calculations using matrix correlation, because with Correlation matrix will remove effect from different scales in variables and allows analysis focus on relationships between variables, not on the scale of each variable. Here is the formula to count matrix based on covariance matrix:

$$r_{i,j} = \frac{Cov(X_i, X_j)}{\sigma_i \sigma_j}$$

d. Eigenvalues

Eigenvalues are non - negative scalar values indicating the proportion of variance from principal component. The bigger eigenvalues, the more variance captured by the principal components. The eigenvalues are calculated with this equation:

$$\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$$

e. Eigenvectors

Every eigen vectors pointing toward certain in room variables and each the direction represent One principal component.

$$(\boldsymbol{A} - \lambda \boldsymbol{I})\vec{v} = \vec{0}$$

f. Loading component

Loading is values that indicate correlation variable original to principal component (Hays, 1983). Loadings components formed from this equation (Rais et al., 2021):

$$r_{X_i,PC_t} = \overrightarrow{v_{\iota,t}}\sqrt{\lambda_t}$$

g. Principal Components Equation

Principal Components Equation is representation base How principal components formed as linear combination of variables original with elements eigenvectors.

$$PC_t = \overrightarrow{v_{1,t}}\, Z_1 + \overrightarrow{v_{1,t}}\, Z_2 + \cdots + \overrightarrow{v_{\iota,t}}\, Z_i$$

h. New data transformation results reduction with PCA

Data transformation involves original data projection to room new ones formed by principal components. This involves multiplying the standardized data by eigenvector. This produces a dataset with more dimensions small, but still maintain information important from the original data.

$$PC_{l,t} = \overrightarrow{v_{1,t}}\, z_{l,1} + \overrightarrow{v_{2,t}} z_{l,2} + \cdots + \overrightarrow{v_{\iota,t}}\, z_{l,i}$$

5. Algorithm K-Means Clustering

Clustering is method data grouping with analyze data object based on distance or similarity (Gupta & Chandra, 2019). Principles This clustering is maximize similarity objects intra cluster and minimize similarity objects inter cluster (Williams, 2022). Clustering can grouped become a number of types, namely Hierarchical clustering and Non-hierarchical clustering (Butar Butar, 2023). K-Means is one of the Non-hierarchical clustering method, which is a grouping method based on partition that divides data objects to in a number of cluster, so that similarity between object in One cluster high, than with objects in other clusters (Norshahlan et al., 2023). Stages data grouping with algorithm K-Means covering (Rais et al., 2021):

a. Determine the number of cluster (k) which will used with method elbow

One of step basic and important in unsupervised learning uses K-Means is determine cluster value (k). The elbow method is one of the method best in determination kand is more suitable for application to the selection of k relatively small values. (Cui, 2020). Elbow method is visual way to determine k with count the value of Within Cluster Sum of Squares (WCSS). The following is WCSS:

$$WCSS_k = \sum_{c=1}^{k} \sum_{P_l \in Cluster_c} distance(P_l, Centroid_c)^2$$

For count distance between two points, used formula distance Euclidean with equation as following:

$$d(P_l, Centroid_c) = \sqrt{\sum_{u=1}^{t}(P_l(u) - Centroid_c(u))^2}$$

$$d(P_l, Centroid_c) = \sqrt{\sum_{u=1}^{t}(P_l(u) - Centroid_c(u))^2}$$

b. Conducting validity testing cluster use the value of Silhouette Coefficient and Davies-Bouldin Index with this following equation:

1) Silhouette Coefficient

Silhouette Coefficient was used for see quality results cluster formed with evaluate how much Good an object or observation placed in a cluster (Rosyada & Utari, 2024). This method combines method cohesive with method separation. Cohessian method functioning to measure how much near distance between observation in the same cluster, whereas method separation used For measure how much Far a separate cluster with other clusters. As for the stages in count Silhouette Coefficient covering:

i) Count average observation distance to - lwith all observations in the same cluster.

$$a(l) = \frac{1}{|A|-1}\sum_{m\in A, m\neq l} d(l,m)$$

ii) Count the value of b(l)which is the minimum value of the average distance of the observation to lthe cluster Awith all observations in the cluster different.

$$s(l) = \min\left(\frac{1}{|C|}\sum_{m\in C} d(l, C_m)\right),\ C \neq \text{A}$$

The values *s*(*l*) are between -1 and 1, where each value is interpreted as follows:

s(*l*) ≈ 1, the observation is very precisely l located in its cluster

s(*l*) ≈ 0, the observation lis between the two clusters (not clear placement cluster his)

s(*l*) ≈ -1, the observation is not lexactly in its cluster

iii) Count the average value of the Silhouette Coefficient of every observation in a cluster certain.

$$SI_A = \frac{1}{|A|}\sum_{l=1}^{|A|} s(l)$$

iv) Count Silhouette Coefficient global with the equation:

$$SC_k = \frac{\sum_{A=1}^{k}(|A|\times SI_A)}{\sum_{A=1}^{k}|A|}$$

Measuring value from Silhouette Coefficient global as following:

Table 1. Measurement Values Silhouette Coefficient Global

Silhouette Coefficient Value	Interpretation
0,71 – 1,00	Generated clusters strong
0,51 – 0,70	Generated clusters good
0,26 – 0,50	Generated clusters weak
≤ 0,25	Can not categorized as cluster

2) Davies-Bouldin Index

Davies-Bouldin Index (DBI) is one of the metrics used for evaluate quality results clustering in data analysis. The purpose of this metric is for measure how much good results clustering in separate group data into different clusters and minimize distance between observation in One same cluster. As for the stages in count DBI value is:

i) Count cluster centroid for each cluster with this following equation:

$$Centroid_c = \left(\frac{1}{|C_c|}\sum_{x_l\in C_c} x_l, \frac{1}{|C_c|}\sum_{y_l\in C_c} y_l\right),\ c = 1,\dots,k$$

ii) Count size dispersion / spread in A cluster

$$SSW_c = \frac{1}{|C_c|}\sum_{x_l\in C_c} d(l, Centroid_c),\ c = 1,\dots,k$$

iii) Count separation / distance inter – cluster

$$SSB_{c,a} = d(Centroid_c, Centroid_a)$$

iv) Count the ratio of inter – cluster

$$R_{c,a} = \frac{SSW_c + SSW_a}{SSB_{c,a}}$$

v) Count the value of Davies-Bouldin Index

$$DBI_k = \frac{1}{k}\sum_{c=1}^{k} max_{c\neq a}\left(R_{c,a}\right)$$

c. Choose initial centroid with use this following equation

$$Centroid_{initial} = \frac{(mean+median)}{2}$$

The stages that are needed carried out including:

1) For every data point, count the distance from point origin O(0,0)using the Euclidean distance formula.

$$(P_l, O) = \sqrt{\sum_{u=1}^{t}(P_l(u) - O(u))^2}$$

2) Sorting the distance that has been counted from the lowest to highest value. According to order distance, also sort the original data points.
3) Divide the data points equally into k cluster.
4) Count the mean and median from data points in every cluster. Then count initial centroid.
   i) Mean

$$\mu_c(u) = \frac{\sum_{l=1}^{n_c} P_{l,c}(u)}{n_c},\ u = 1,\dots,t$$

   ii) Median

   Median for a cluster with an odd number of members:

$$MD_c(u) = P_{\frac{n_c+1}{2}}(u)$$

Median for a cluster with the even number of members:

$$MD_c(u) = \frac{P_{\frac{n_c}{2}}(u) + P_{\frac{n_c}{2}+1}(u)}{2}$$

iii) Initial Centroid

$$Centroid_{awal,c} = \frac{(\mu_c + MD_c)}{2}$$

$$c = 1, \dots, k$$

5) The value Centroid_(initial,c)of each cluster will used as initial centroid in stage K-Means.

d. Counting distance all over observation with initial centroid use Euclidean equation:

$$d(P_l, Centroid_{awal,c}) = \sqrt{\sum_{u=1}^{t}(P_l(u) - Centroid_{awal,c}(u))^2}$$

e. Allocate observation study to centroid nearest cluster based on results calculation distance The smallest Euclidean.

f. Count centroid new based on membership from each cluster with this following equation:

$$Centroid_{baru,c}(u) = \frac{1}{n_c}\sum_{l=1}^{n_c} P_{l,c}(u)\,,$$

$$u = 1, \dots, t$$

g. Counting distance all over observation with new centroid use Euclidean equation:

$$d(P_l, Centroid_{baru,c}) = \sqrt{\sum_{u=1}^{t}(P_l(u) - Centroid_{baru,c}(u))^2}$$

h. Allocate observation study to centroid nearest cluster based on results calculation distance The smallest Euclidean.

i. Repeat steps f, g, and h if still there is observation data that moves cluster and stop if the data in cluster Already stable or reach condition convergent.

RESULT AND DISCUSSION

1. Research Data

Data used in this research is secondary data sourced from publication of East Java Province data on the Central Statistics Agency website East Java Province, with object study consists of from 38 districts or cities in the province The variables used in This research is an indicator welfare society based on research previously used indicator similar. Variables the covering number of poor population (X_1), poverty line (X_2), poverty depth index (P1) (X_3), poverty severity index (P2) (X_4), open unemployment rate (X_5), minimum wage (X_6), Gini ratio (X_7), economic growth (X_8), life expectancy (X_9), not yet fully achieved (X_{10}), per capita monthly expenditure on food expenditure (X_{11}), households with access to clean drinking water sources (X_{12}), and households that have received assistance/social assistance/local government subsidy programs (X_{13}). These thirteen variables are used as indicators of community welfare in East Java Province.

Table 2. Statistics Descriptive

Variables	N	Mean	Maximum	Minimum	Std. Deviation	Unit
X_1	38	110,232.6	251,360	7.100	67,414,5166678	Soul
X_2	38	487,879,080	718,370	352,606	87,857,8522006	Rupiah per capita per month
X_3	38	1,4934	4.50	0.35	0.79768	Index number
X_4	38	0.3324	1.42	0.06	0.24052	Index number
X_5	38	4,6629	8.05	1.71	1.40990	Message
X_6	38	2,694,768,489	4,525,479.19	2,114,335.27	778,694,385	Rupiah per month
X_7	38	0.34587	0.423	0.254	0.036989	Index number
X_8	38	4,7082	6.19	1.20	1.11939	Percent
X_9	38	72,4179	74.91	67.60	1.95596	Year
X_{10}	38	8,3755	11.82	5.07	1.63643	Year
X_{11}	38	50,8203	62.52	37.34	6,19011	Percent
X_{12}	38	96,1195	100.00	79.26	4.74754	Percent
X_{13}	38	10,0553	36.03	1.81	6,32539	Percent

2. Detection Outlier

Outlier detection results with the boxplot method shown in Figure 1 shows that there are some variables, such as index depth poverty, index severity poverty, minimum wage, and some variable other own outlier values. This can be seen from the value observations that are outside the lower limit and upper limits. The existence of these outliers becomes consideration in election method initial centroid based on average mean-median, with objective for reduce influence outlier values and the centroid value more representative.

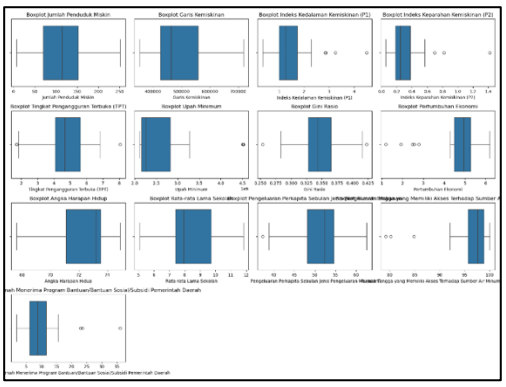

Figure 1. Boxplot

3. Assumption Test

Before operating PCA and K-Means stages, the first is do assumption tests of dataset. The assumption test includes, data feasibility test, correlation test between variables, and multicollinearity test.

a. Test data feasibility with KMO and MSA

KMO is used to see data sufficiency in overall to perform PCA analysis. While MSA is used to see data adequacy on each variable. The KMO value obtained with the assistance of SPSS and R Studio software is 0.73, while MSA value is as following:

Table 3. MSA Test Results

Variables	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}
MSA Values	0.78	0.60	0.72	0.67	0.85	0.68	0.86	0.63	0.61	0.80	0.82	0.55	0.74

A KMO value of 0.73 indicates that acceptance criteria were met H_0, so it means that the data size is sufficient to perform PCA analysis. And in Table 3, the MSA value for each variable is above 0.5, which also means that acceptance occurs H_0, so that all variables used in this study are sufficient to perform PCA analysis.

b. Correlation test between variable with Bartlett's Test of Sphericity

For ensure that the data has required characteristics for get meaningful results in research, correlation test was conducted between variable with Bartlett's Test of Sphericity. The value of Bartlett 's Test of Sphericity obtained with help SPSS software is as following:

Table 4. Bartlett's Test of Sphericity

Bartlett's Test	
Chi-Square	413,530
df	78
Sig.	0,000

In Table 4 it can be seen seen that the sig. or p-value produced is 0.000, so p-value<α=0,05, meaning that rejection is carried out H_0. So, it can be concluded that there is a correlation between the variables, so that the calculation can be continued.

c. Multicollinearity test with Variance Inflation Factor (VIF)

After done checking multicollinearity in variables study with use VIF value. A VIF value greater than 10 indicates a serious multicollinearity problem. VIF value obtained with help Jupyter using the statsmodels library, shown in the following table:

Table 5. VIF Values

VIF Value			
$\boldsymbol{X_1}$	3.93	X_8	1.84
$\boldsymbol{X_2}$	7.26	X_9	4.18
$\boldsymbol{X_3}$	67.27	X_{10}	13.83
$\boldsymbol{X_4}$	46.77	X_{11}	12.97
$\boldsymbol{X_5}$	2.55	X_{12}	1.49
$\boldsymbol{X_6}$	2.49	X_{13}	1.66
$\boldsymbol{X_7}$	2.95		

The Table 5 above show there is several variables that have VIF value is above 10, so can concluded that the data occurs symptom multicollinearity. Thus, PCA is used to handle symptoms of multicollinearity.

4. Principal Component Analysis (PCA)

a. Data Standardization

Data standardization is carried out use ensure that every feature in data contributed in a way proportional to analysis, or in other words every feature own balanced weight. This involves centralization to the average and scaling of the data, so that the data has zero mean and deviation standard one. The following is shown results data standardization:

Table 6. Standardization Result Data

Data	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	**…**	$\boldsymbol{X_{13}}$
1	-0.50482645	-1.539681154	…	-0.60316649
2	-0.39342612	-1.056366356	…	0.05291956
…	…	…	…	…
38	-1.52982824	1.435340362	…	2.14290932

b. Matrix Covariance

Matrix the covariance below is only can used For see direction linear relationship between variable Because is results calculation of data that has not been standardized.

$$Cov = \begin{pmatrix} 4.544717e+03 & -2.232114e+06 & \dots & -5.203515e+01 & -1.865091e+02 \\ -2.232114e+06 & 7.719002e+09 & \dots & 1.540772e+05 & 1.906338e+05 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -5.203515e+01 & 1.540772e+05 & \dots & 2.253918e+01 & 6.767708e+00 \\ -1.865091e+02 & 1.906338e+05 & \dots & 6.767708e+00 & 4.001056e+01 \end{pmatrix}$$

c. Coefficient Correlation

Coefficient The correlation below provides outlook about strength and direction linear relationship between variable.

Table 7. Coefficients Correlation

	X_1	X_2	…	X_{12}	X_{13}
X_1	1,0000000	-0.3768620	…	-0.16258272	-0.4373804432
X_2	-0.3768620	1,0000000	…	0.36939297	0.3430299785
…	…	…	…	…	…
X_{12}	-0.1625827	0.3693930	…	1,0000000	0.2253643803
X_{13}	-0.4373804	0.3430300	…	0.22536438	1,0000000

The following is a form matrix from coefficient correlation.

$$\boldsymbol{r} = \begin{pmatrix} 1,0000000 & -0,3768620 & \dots & -0,16258272 & -0,4373804432 \\ -0,3768620 & 1,0000000 & \dots & 0,36939297 & 0,3430299785 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -0,1625827 & 0,3693930 & \dots & 1,0000000 & 0,2253643803 \\ -0,4373804 & 0,3430300 & \dots & 0,22536438 & 1,0000000 \end{pmatrix}$$

The next calculation were carried out using the correlation matrix, as it removes the effect of different variable scales and allows the analysis to fokus on relationships among variables.

d. Eigenvalues and Eigenvectors

Eigenvalues (λ)

The following is an eigenvalues represent amount variance explained by each principal components.

This are the Table 8 Total Variance Explained, which is the percentage column variance containing percentage variance explained by each principal component.

$$\lambda = \begin{pmatrix} 6,015465701 \\ 1,549948636 \\ 1,330472627 \\ 1,083903522 \\ 0,884909028 \\ 0,633712585 \\ 0,503464565 \\ 0,351205808 \\ 0,317907312 \\ 0,209274882 \\ 0,068181166 \\ 0,042721984 \\ 0,008832185 \end{pmatrix}$$

And column percentage cumulative containing percentage cumulative variance explained by principal components now and all principal components previously.

Table 8. Variance Explained Total

Component	Initial Eigenvalues		
	Total	Variance (%)	Cumulative (%)
1	6.015465701	46%	46%
2	1.549948636	12%	58%
3	1.330472627	10%	68%
4	1.083903522	8%	77%
5	0.884909028	7%	84%
6	0.633712585	5%	88%
7	0.503464565	4%	92%
8	0.351205808	3%	95%
9	0.317907312	2%	97%
10	0.209274882	2%	99%
11	0.068181166	1%	100%
12	0.042721984	0%	100%
13	0.008832185	0%	100%

Eigenvectors ($\vec{v}$)

Below are the which eigenvectors represent the direction in which the data varies the most. Because eigenvectors formed from matrix correlation, so that long from eigenvector is 1.

$$\vec{v} = \begin{pmatrix} 0{,}2803064 & 0{,}42295015 & \dots & 0{,}186862185 & -0{,}003302354 \\ -0{,}2634501 & 0{,}32667215 & \dots & -0{,}495422838 & -0{,}025776259 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -0{,}1207303 & 0{,}13693688 & \dots & 0{,}008340910 & 0{,}043575996 \\ -0{,}1909837 & -0{,}18367533 & \dots & 0{,}060154851 & 0{,}028199571 \end{pmatrix}$$

Furthermore is determination the number of principal components that will maintained. Determination the number of principal components that will be maintained depends on the purpose analysis conducted . There are several guides that can followed for decide How many principal components that will be maintained, among others :

1) Maintain sufficient principal components for explain percentage certain of the total variance, for example 80%.
2) Maintain principal components that have eigenvalue bigger from the average.
3) Use scree graph, which is a plot of eigenvalues to order principal component. In this graph it will be searching for elbow point, the point where there is difference significant between eigenvalue large and eigenvalue small.

In this study, the aim he did reduction dimensions using PCA is optimize results clustering K-Means, so that in this is the usage method screen graph more chosen for determine the number of principal components that will maintained Because produce results more optimal clusters.

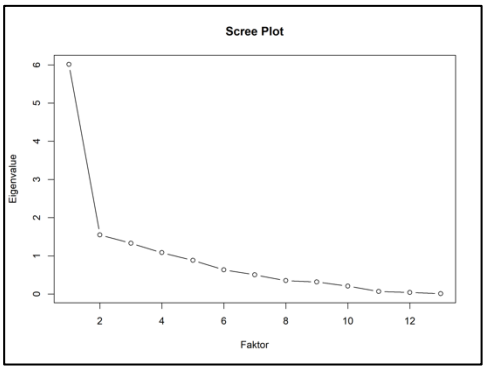

Figure 2. Scree Plot

Based on Figure 1, the elbow point is clearly visible on the second component. Viewed from eigenvalues and Table 8, components 1has eigenvalue as big as 6,015465701and percent of variance as big as 46%. Where it means the component 1 was able to explain 46%the variance or pattern in the original data. Furthermore, the component 2has an eigenvalue of 1.549948636 and percent of variance as big as 12%. Where it means that 12%the variance or pattern in the original data can be explained by the components 2. Of the two principal components First obtained percentage cumulative variance 58%, where it is assessed Enough explain variability in the original data.

e. Loading Component

Loading components show the magnitude correlation variable to score components that are formed. With R Studio help found:

Table 9. Component Loadings

	PC_1				PC_2		
X_1	-0.69	X_8	0.47	X_1	0.53	X_8	0.04
X_2	0.65	X_9	0.62	X_2	0.41	X_9	-0.01
X_3	-0.85	X_{10}	0.93	X_3	0.23	X_{10}	0.04
X_4	-0.74	X_{11}	-0.94	X_4	0.30	X_{11}	0.05
X_5	0.63	X_{12}	0.30	X_5	0.51	X_{12}	0.17
X_6	0.42	X_{13}	0.47	X_6	0.78	X_{13}	-0.23
X_7	0.77			X_7	-0.09		

Based on information about the value of loadings in the Table 9 above, obtained conclusion that variable X_1, X_2, X_3, X_4, X_5, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, and X_{13} is variables dominant or contribute big to PC_1. PC_1 captures the largest share of variability in the data with welfare economy in a way comprehensive, such as condition poverty and inequality socio-economic.

Besides that, for variable X_6 is contributing variables big to PC_2. Where is PC_2 catch variability or the second largest information and explain variation additional that is not caught entirely by PC_1. In this case with consider X_6 as variable contributor contribution big as well as variables other in PC_2, then this PC_2 reflects aspect addition in related data with welfare economy other like inequality income that becomes the specific information, where the information contained not to full caught by PC_1.

f. Principal Component Equation

Principal component equation formed based on eigenvectors, then obtained equation as following:

$$PC_1 = 0{,}28Z_1 - 0{,}26Z_2 + 0{,}35Z_3 + 0{,}30Z_4 - 0{,}26Z_5 - 0{,}17Z_6 - 0{,}32Z_7 - 0{,}19Z_8 - 0{,}25Z_9 - 0{,}38Z_{10} + 0{,}38Z_{11} - 0{,}12Z_{12} - 0{,}19Z_{13}$$

$$PC_2 = 0{,}42Z_1 + 0{,}33Z_2 + 0{,}19Z_3 + 0{,}24Z_4 + 0{,}41Z_5 + 0{,}62Z_6 - 0{,}08Z_7 + 0{,}03Z_8 - 0{,}01Z_9 + 0{,}03Z_{10} + 0{,}04Z_{11} + 0{,}14Z_{12} - 0{,}18Z_{13}$$

Principal component equation above will form new data from results linear combination between eigenvectors with initial data that has been standardized. Every principal component nature upright straight one each other, which means that each principal component is not will correlated one each other. And every next principal component will explain the remaining variations that are not captured by the previous principal components.

g. Data Transformation

Data transformation using equation principal component with multiply the data that has been standardized and eigenvectors so that new data obtained as following:

Table 10. Data Transformation

Data to -	PC_1	PC_2
1	1.63	-2.30
2	-0.16	-1.24
3	0.19	-1.51
4	-0.92	-0.87
…	…	…
38	-3.10	-0.74

With PCA producing data with more dimensions small However still maintain part big information important things contained in the initial data.

5. Algorithm K-Means Clustering

a. WCSS values for each *k*

Table 11a. WCSS Values

k	**1**	**2**	**3**	**4**	**5**
WCSS	287.47	145.90	96.36	74.67	51.20

Table 12b. WCSS Values

k	**6**	**7**	**8**	**9**	**10**
WCSS	26.38	22.25	19.45	18.01	14.40

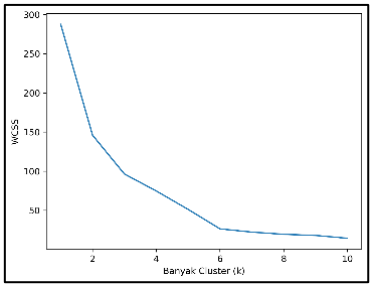

Figure 3. WCSS Graph

In the Figure 2 above, its seen that k=6 shows the turning point of the curve or in other words the “elbow” of the curve. This shows that k=6 it is the sum of k optimum cluster.

b. Validity test results cluster using the value of Silhouette Coefficient and Davies-Bouldin Index (DBI)

1) Silhouette Coefficient value for each cluster (*k*)

Table 13a.Silhouette Coefficient Values

k	**2**	**3**	**4**	**5**
SC	0.403	0.352	0.369	0.399

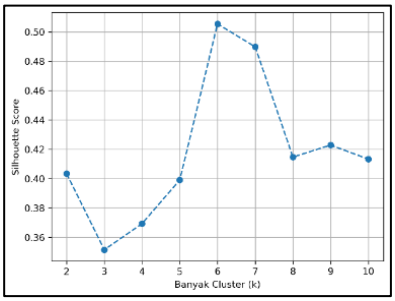

Figure 4. Silhouette Coefficient Value Graph

p-ISSN 1829 586X
e-ISSN 2581-0170

Table 14b.Silhouette Coefficient Values

k	**6**	**7**	**8**	**9**	**10**
SC	0.505	0.490	0.415	0.423	0.413

Figure 4 shows that chart of the highest value is in $k = 6$, then Table 12 showing the Silhouette Coefficient value 0,505. This value shows the quality of the clustering results. is moderate. The value 0,505 indicates that each data tends to be closer to other data in its respective cluster.

2) DBI Value for each cluster (k)

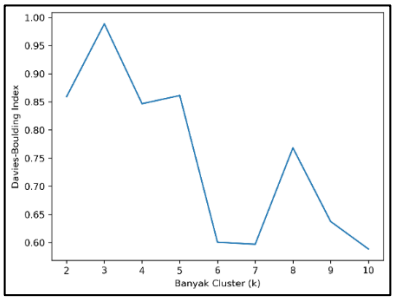

Figure 5. DBI Value Graph

Table 15a. DBI Value

k	**2**	**3**	**4**	**5**
DBI	0.859	0.989	0.847	0.861

Table 16b. DBI Value

k	**6**	**7**	**8**	**9**	**10**
DBI	0.601	0.597	0.768	0.637	0.589

In the image of DBI and the table 13 show DBI shows that k=6 is not k with the lowest value. However, $k = 6$ it has a DBI value 0,601 which is classified as low, indicating clustering results can be considered relatively good. Low DBI value below 1 indicates that each cluster is separate well.

Table 17. Results of SC and DBI Value Evaluation

Number of PC	**SC**	**DBI**
2	0.505	0.601
4	0.343	0.981
5	0.278	1,205

Based on results evaluation clustering use The Silhouette Coefficient and DBI presented in Table 14 reveal that quality clustering best obtained with using two principal components, although variance cumulative covered only by 58%. PCA was designed for identify the most contributing component to total variance of data at once allow for eliminate redundant variables in analysis (Jolliffe, 2014). Thus, the K-Means algorithm focuses more on the most informative features (Shlens, 2014). Selecting two PCs according to with objective analysis, namely produces optimal clustering, although must sacrifice part variance cumulative.

c. Stage to choose the initial centroid

1) The distance of each data point from the origin was calculated using the Euclidean formula from $O(0,0)$

Table 18. Data Distance to Point of Origin

Data to -	PC_1	PC_2	**Distance from $O(0,0)$**
1	1.63	-2.30	2.82
2	-0.16	-1.24	1.25
3	0.19	-1.51	1.52
4	-0.92	-0.87	1.27
5	-0.01	-0.75	0.75
…	…	…	…
38	-3.10	-0.74	3.19

After getting the score of distance from each data point to point origin $O(0,0)$, the next stage is to sort the data based on the results of the distance calculation from the lowest to the highest value.

2) The results of sorting data based on distance value

Table 19. Results of data sorting based on distance value

Data to -	PC_1	PC_2	Distance from $O(0,0)$
18	0.16	-0.19	0.25
17	-0.46	0.03	0.46
19	0.20	-0.67	0.70
5	-0.01	-0.75	0.75
10	-0.63	-0.85	1.06
…	…	…	…
29	5.23	1.11	5.35

As has been known previously, that the value of k with optimal cluster is at $k = 6$, so six centroids will be needed with still -each cluster has One centroid. With thus, to get six centroid, data that has been sorted above will share into the six parts, where the divisions are represent One cluster.

3) The results of data division to in six cluster

Table 20. Results of data division into in six cluster

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
(0.16; -0.19)	(-0.16; -1.24)	(0.19; -1.51)	(2.04; -0.66)	(1.63; -2.30)	(-3.57; 2.46)
(-0.46; 0.03)	(-0.92; -0.87)	(-0.09; 1.67)	(-2.15; -0.27)	(-1.01; 2.86)	(-4.26; 1.02)
(0.20; -0.67)	(1.30; 0.26)	(-0.25; 1.80)	(2.26; 0.17)	(-3.10; -0.74)	(4.20; 1.72)
(-0.01; -0.75)	(-0.68; -1.18)	(-2.01; -0.04)	(2.03; -1.22)	(3.24; 1.18)	(-4.41; 2.61)
(-0.63; -0.85)	(1.31; 0.44)	(1.64; -1.19)	(-2.34; -0.60)	(-3.96; -0.78)	(5.31; 0.44)
(-0.06; 1.15)	(0.90; -1.16)	(2.02; -0.25)	(-2.38; -1.06)	(-3.93; -1.67)	(5.23; 1.11)
(1.24; 0.05)	(1.49; 0.25)				

4) Count the mean and median from each data in each cluster

Table 21. Initial of Each Cluster

Cluster to -	Mean	Median	$Centroid_{awal}$
1	(0.06; -0.17)	(-0.01; -0.75)	(0.03; -0.46)
2	(0.46; -0.50)	(-0.68; -1.18)	(-0.11; -0.84)
3	(0.25; 0.08)	(-1.13; 0.88)	(-0.44; 0.48)
4	(-0.09; 0.61)	(2.14; -0.53)	(1.03; -0.57)
5	(-1.19; -0.24)	(0.07; 0.22)	(-0.56; -0.01)
6	(0.41; 1.56)	(-0.11; 2.16)	(0.15; 1.86)

d. PCA clustering results with K-Means for koptimum, $k = 6$

Table 22. Clustering Results

Cluster	Data to -	Member Cluster
		Regency /City
1	1, 6, 8, 9, 11, 12, 21, 22, 23, 24, and 28	The cities of Pacitan, Kediri, Lumajang, Jember, Bondowoso, Situbondo, Ngawi, Bojonegoro, Tuban, Lamongan and Pamekasan.
2	2, 3, 4, 5, 10, 17, 18, 19, and 20	The cities of Ponorogo, Trenggalek, Tulungagung, Blitar, Banyuwangi, Jombang, Nganjuk, Madiun and Magetan
3	15, 32, and 37	Sidoarjo, Malang City, and Surabaya City
4	13, 26, 27, and 29	Probolinggo, Bangkalan, Sampang, and Sumenep
5	30, 31, 33, 34, 35, 36, and 38	The cities of Kediri, Blitar, Probolinggo, Pasuruan, Mojokerto, Madiun and Batu.
6	7, 14, 16, and 25	Malang, Pasuruan, Mojokerto, and Gresik

Based on the clustering results above, are made visualization distribution district / city in form coordinates and maps using Python and ArcGIS software, as shown in figure 6 and figure 7.

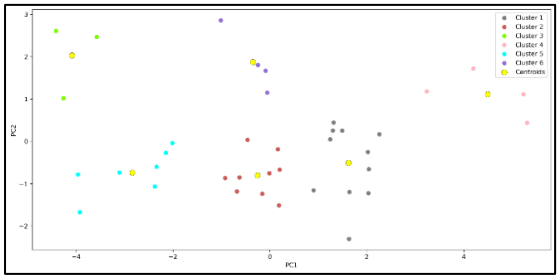

Figure 6. Cluster Distribution

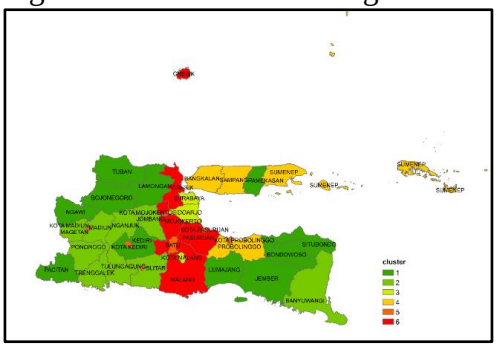

Figure 7. Cluster Map

e. Aerage value every variable in each cluster

Table 23. Average Value of Each Variable by Cluster

Variables	Cluster to -					
	1	2	3	4	5	6
X_1	135.78	91.13	97.77	207.37	12.04	167.02
X_2	424,958.91	437,550	654,908.67	486,014.25	587,716.71	476,026.75
X_3	1.72	1.18	0.73	3.37	0.87	1.39
X_4	0.37	0.23	0.17	0.88	0.19	0.30
X_5	3.76	4.79	7.20	3.46	4.94	5.67
X_6	2,314,705.74	2,287,875.67	4,079,401.67	2,299,218.00	2,586,179.31	4,202,556.56
X_7	0.33	0.34	0.40	0.29	0.37	0.35
X_8	4.45	4.94	5.98	3.46	4.81	5.00
X_9	71.30	73.32	74.52	70.01	73.35	72.66
X_{10}	7.38	8.17	10.81	5.82	10.50	8.58
X_{11}	54.51	50.65	39.53	60.71	43.54	52.38
X_{12}	93.14	95.99	98.13	96.61	99.43	96.81
X_{13}	6.73	10.84	8.97	7.58	17.77	7.21

Referring to the Table 20 above, there are several analyses that can displayed. First, cluster 1 illustrates area with level relative well- being low. This is indicated by several indicators, such as index depth (X_3)

and severity poverty (X_4) are above average, which indicates that poor people in this area live far below the poverty line. Average length of schooling (X_{10}) and access to drinking water worthy (X_{12}) below average also becomes sign that need base public Not yet fulfilled in a way maximum, where this also applies influence quality life its population. In addition, the economic growth (X_8) which is below average indicates limitations opportunity for increase income and community welfare in a way overall.

In cluster 2 it reflects aspect poverty and inequality more income low. This looks from the poverty line (X_2), the index depth (X_3) and severity poverty (X_4), level unemployment open (X_5), and Gini ratio (X_7) is below average. However, what is challenge in This cluster is economic growth (X_8) which is below average. This indicates that although inequality and poverty can more controlled, without existence growth adequate economy, opportunities for increase community welfare still limited. Therefore that, attention need given for increase economic growth to be able to support repair more carry on in quality life and creation opportunity more economy Good for public.

Next, cluster 3 has value high on the poverty line variable (X_2), the level unemployment open (X_5), Gini ratio (X_7), economic growth (X_8), life expectancy (X_9), and not yet fully achieved (X_{10}). On the other hand, this cluster has value low on variables index depth poverty (X_3), index severity poverty (X_4), and expenditure per capita a month type expenditure food (X_{11}). This shows that cluster 3 has condition economy, quality life, and education is relatively Good.

Expenditure per capita a month type expenditure low food (X_{11}) indicates existence improvement income resident in This cluster. Phenomenon the in accordance with Engel's law, which states that when income increase and preference consumer still, then percentage expenditure for food tend decrease. In addition, the high economic growth (X_8) in This cluster strengthens conclusion the.

However thus, in the middle good condition said, cluster 3 is still show existence group trapped residents in valley poverty. This is reflected from height inequality income, such as seen in the value like this high ratio (X_7). Inequality the possibility big due to the height level unemployment open (X_5). Although thus, income the poor population in this cluster is relatively not significantly from the poverty line (X_2), and the gap between they also don't severe. This is seen from the value index depth poverty (X_3) and index severity low poverty (X_4). In general, cluster 3 shows condition relative socio- economic good. However, some group poor people in This cluster still need attention specifically for welfare in a way overall Can improved.

In cluster 4 high on variables number of poor population (X_1), index depth poverty (X_3), and the index severity poverty (X_4), then low on variables economic growth (X_8), life expectancy (X_9) and not yet fully achieved (X_{10}). However, in cluster 4, the variables level unemployment open (X_5) and the Gini ratio (X_7) is the lowest, and variable expenditure per capita a month type expenditure food (X_{11}) is the highest. This shows that distribution income in cluster 4 is classified as even, not There is inequality extreme among its population.

However, the level high poverty, quality low living and education, as well inequality expenditure for need

p-ISSN 1829 586X
e-ISSN 2581-0170

life severe base among poor group, dominated by cluster 4. Expenditure per capita a month type expenditure high food (X_{11}) show that resident more focus for fulfil need base type food compared to non- food like education and health. In general, distribution residents in the city / district included in cluster 4 is classified evenly, will but condition economy in a way overall not significantly different and at a low level. This is seen from level high poverty as well as quality poor living and education adequate.

Next, cluster 5 shows conditions that are not significantly different with cluster 2. Visible from index depth (X_3) and severity poverty (X_4), and like this ratio (X_7) which is below the average, this indicates that cluster 5 has condition sufficient welfare good. However, economic growth (X_8) and per capita expenditure for type food (X_{11}) which is lower from the average need attention.

Lastly in cluster 6 shows relative condition better from aspect welfare social. Some indicators important like life expectancy (X_9), not yet fully achieved (X_{10}), access to drinking water feasible (X_{12}), and per capita expenditure for food (X_{11}) is above average, which indicates quality enough life good. In addition, the economic growth (X_8) is also above average. However, what becomes attention is level unemployment open (X_5) which is higher, which shows existence challenge in create field enough work for population. Although many aspects greater welfare ok, problem employment still need get attention serious for the economy this area can develop more evenly and increase welfare in a way overall.

f. Comparison method K-Means

Table 24. Comparison of K -Means Methods

PCA	Initial Centroid	Many Iteration	Optimal k value	Validity Test k Optimum	
				Silhouette Coefficient	DBI
–	Random	9	3	0.252	1,289
–	$\frac{(\text{Mean} + \text{Median})}{2}$	2	2	0.280	1,289
✓	Random	10	6	0.505	0.601
✓	$\frac{(\text{Mean} + \text{Median})}{2}$	7	6	0.505	0.601
✓	Mean	8	3	0.455	0.728
✓	Median	9	6	0.505	0.985

The Table 21 above show results comparison from several methods. Based on table 21 the obtained several conclusions. First, compared with K-Means without PCA, K-Means with PCA providing contribution in improvement quality results cluster. Where the value Silhouette Coefficient obtained from method K-Means with PCA, giving more value tall compared to with K-Means without PCA. This is means clustering K-Means with PCA providing results that indicate that every data tends to more near or similar with other data in their respective clusters. Likewise, with DBI value, K-Means with PCA produces low DBI value below 1, which means each cluster is well separate. Second, initial centroid with mean, median, and combination mean median contributes to speed up convergence, with initial combination centroid mean median yields more convergence fast compared to mean or median only. In other words, the initial

centroid combination Mean Median increases efficiency achievement convergence. So, in general overall, in this case is modified K-Means with PCA and initial centroid combination mean median can increase results quality cluster and improve efficiency achievement convergence in clustering with algorithm k-means.

CONCLUSION

From the proposed approachsix optimal clusters were formed. Based on he cluster profiling, to support government programs aimed at improving regional welfare, the main focus should be on enhancing economic growth and reducing the open unemployment rate. In addition, improvements in access education and quality live in the areas with life expectancy and a low and low average length of schooling also requires attention. Reducing income and social inequality will help create a more equitable distribution of welfare, so that improvements in welfare can be realized more effectively.

Furthermore, the results of this study revealed that there was an improvement performance clustering through implementation Principal Component Analysis (PCA) on K-Means and initial centroid with average mean median. Experiment shows that Application of PCA on K-Means capable give improvement quality results clustering. This is indicated by the value Silhouette Coefficient is 0.505 and value Davies-Bouldin Index (DBI) of 0.60, where both indicates more quality clustering has higher compared to with K-Means without PCA. This result also indicates that each cluster is well-separated and that the data points within the same cluster are closely related.

The study also revealed that initialization centroid with the value average mean median can accelerate the convergence process K-Means, where in this research was achieved in seven iterations. Ased on the experimental results, his approach has been proven to be more efficient compared to centroid initialization that uses only mean or median value separately.

For future research, t is recommended to test the performance of the proposed approach with other dimensionality reduction methods such as the Gini Index or Linear Discriminant Analysis (LDA) to deepen and enrich the insights and findings.

Acknowledgments

The authors would like to thank Universitas Negeri Malang (UM) for providing financial support for the publication of this research.

REFERENCES

Aeli, L. W., Pahrany, A. D., & Indratno, S. W. (2022). Life insurance model with regression Cox proportional hazard affected by areal spatial factor. *In Mathematics, Substance and Surmise: Views on the Meaning and Ontology of Mathematics.* AIP Publishing. https://doi.org/10.1007/978-3-319-21473-3_6

Ann G. Ryan, Douglas C. Montgomery, Elizabeth A. Peck, G. G. V. (2013). *Solutions manual to accompany introduction to linear regression analysis. Technometrics*, *49*(2). https://doi.org/10.1198/tech.2007.s499

Badan Pusat Statistik. (2023). *Hasil long form sensus penduduk 2020 Provinsi Jawa Timur*. Badan Pusat Statistik Provinsi Jawa Timur.

Bluman, A. G. (2019). *Elementary statistics: A step by step approach: A brief version* (8th ed.). McGraw Hill Education.

Budiman, A. V. Y., Permai, S. D., & Irwansyah, E. (2024). Criminality

mapping in Java Island using clustering large applications based on randomized search (CLARANS). *2024 4th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET 2024)*. https://doi.org/10.1109/IRASET60544.2024.10548525

Butar Butar, R. P. (2023). Analisis hierarchical dan non-hierarchical clustering untuk pengelompokkan potensi ekonomi kelautan Indonesia 2021. *Jurnal Sistem dan Teknologi Informasi (JustIN)*, *11*(3), 543. https://doi.org/10.26418/justin.v11i3.67283

Cui, M. (2020). On the elbow method. *5–8*. https://doi.org/10.23977/accaf.2020.010102

Gupta, M. K., & Chandra, P. (2019). P-k-means: K-means using partition based cluster initialization method. *SSRN Electronic Journal*, 567–573. https://doi.org/10.2139/ssrn.3462549

Haumahu, G., & Lewaherilla, N. (2020). Penerapan analisis komponen utama dalam mereduksi faktor-faktor penyebab diare di Provinsi Maluku. *Mathematics & Applications (MAP Journal)*, *2*(1), 41–46.

Hays, W. L. (1983). Review of *Using multivariate statistics. Contemporary Psychology: A Journal of Reviews*, *28*(8). https://doi.org/10.1037/022267

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed., pp. 671–757). Pearson.

Jolliffe, I. T. (1998). Principal components. *Data Handling in Science and Technology*, *20*(Part A), 519–556. https://doi.org/10.1016/S0922-3487(97)80047-0

Lestari, T. E., Permadi, H., & Susilowati, S. (2020). Data mining pada faktor-faktor potensi daerah di Kabupaten Sidoarjo Provinsi Jawa Timur. *Jurnal Matematika*, *10*(2), 67. https://doi.org/10.24843/jmat.2020.v10.i02.p124

Meiriza, A., Ali, E., Rahmiati, & Agustin. (2023). Perbandingan algoritma K-Means dan K-Medoids untuk pengelompokan program BPJS Ketenagakerjaan. *The Indonesian Journal of Computer Science*, *12*(2), 714–728. https://doi.org/10.33022/ijcs.v12i2.3184

Norshahlan, M., Jaya, H., & Kustini, R. (2023). Penerapan metode clustering dengan algoritma K-Means pada pengelompokan data calon siswa baru. *Jurnal Sistem Informasi Triguna Dharma (JURSI TGD)*, *2*(6), 1042. https://doi.org/10.53513/jursi.v2i6.9148

Pendi, P. (2021). Analisis regresi dengan metode komponen utama dalam mengatasi masalah multikolinearitas. *Bimaster: Buletin Ilmiah Matematika, Statistika dan Terapannya*, *10*(1), 131–138.

Pratama, R. C. (2020). Pengelompokan kabupaten/kota di Provinsi Papua berdasarkan indikator kesejahteraan rakyat 2020. *Seminar Nasional Official Statistics, 2020*, 853–862.

Rais, M., Goejantoro, R., & Prangga, S. (2021). Optimalisasi K-Means cluster dengan principal component analysis pada pengelompokan kabupaten/kota di Pulau Kalimantan berdasarkan indikator tingkat pengangguran terbuka. *Eksponensial*, *12*(2), 129. https://doi.org/10.30872/eksponensial.v12i2.805

Rosyada, I. A., & Utari, D. T. (2024). Penerapan principal component analysis untuk reduksi variabel pada algoritma K-Means clustering. *Jambura Journal of Probability and Statistics*, *5*(1), 6–13. https://doi.org/10.37905/jjps.v5i1.18733

Taufik, A., Novita, E., Eva, M., Ar-Rosid, D., Istighfarin, & Putri, A. R. (2023). Penerepan principal component analysis (PCA) untuk mereduksi variabel-variabel seputar pertanian yang saling berkorelasi di Provinsi Jawa Timur.

Umargono, E., Suseno, J. E., & S. K., V. G. (2020). K-Means clustering optimization using the elbow method and early centroid determination based on mean and median. *Proceedings of ISSTEC 2019*, *474*, 234–240. https://doi.org/10.5220/0009908402340240

Williams, P. (2022). Smart devices. *Cossma*, *23*(12). https://doi.org/10.1016/b978-0-08-100741-9.00012-7

Wira, B., Budianto, A. E., & Wiguna, A. S. (2019). Implementasi metode K-Medoids clustering untuk mengetahui pola pemilihan program studi mahasiswa baru tahun 2018 di Universitas Kanjuruhan Malang. *RAINSTEK: Jurnal Terapan Sains & Teknologi*, *1*(3), 53–68. https://doi.org/10.21067/jtst.v1i3.3046

Zubair, M., Iqbal, M. A., Shil, A., Chowdhury, M. J. M., Moni, M. A., & Sarker, I. H. (2022). An improved K-Means clustering algorithm towards an efficient data-driven modeling. *Annals of Data Science*, *11*(5), 1525–1544. https://doi.org/10.1007/s40745-022-00428-2