Journal of Mathematics and Scientific Computing with Applications

E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 6, Number 2, 2025
DOI: 10.53806/jmscowa.v6i2.1313
Page: 57–63

IMPLEMENTATION OF PRINCIPAL COMPONENT ANALYSIS (PCA) IN DIMENSION REDUCTION BASED ON INDONESIAN HEALTH DATA

Siti Rafiah Rangkuti[1], **Nurul Fadhillah**[2], **Rita Novita Sari** [3], **Klause Roder**[4]

[1,2]Department of Mathematics, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia
[3]Department of computer science, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia
[4]Department of Applied Mathematics, Universty of Ausburg, Ausburg, Germany

Article Info

Article history:
Received : June 11, 2025
Revised : July 15, 2025
Accepted : August 20, 2025

Keywords:
Data Visualization;
Dimensionality Reduction;
Health Data;
Indonesia;
Principal Component Analysis.

ABSTRACT

Indonesian health data for 2024 has multidimensional characteristics with a large number of interconnected variables, leading to high complexity in the analysis and visualization process. This complexity poses a challenge in generating information that is easy to understand and can support data-driven decision-making. This research aims to implement the Principal Component Analysis (PCA) method as a technique for dimension reduction and visualization of Indonesian health data. The research method used is a quantitative approach with descriptive-exploratory secondary data analysis. The research stages include data pre-processing, PCA implementation, principal component determination, variable contribution analysis, and data visualization using scatter plots and biplots. The research results show that PCA is able to significantly reduce the number of variables while still retaining most of the main information contained in the data. Principal component analysis-based visualization produces clearer and more easily interpretable patterns and structures in health data. Thus, PCA has proven effective in simplifying the complexity of national health data and supporting the presentation of more informative and actionable information for decision-making in the health sector.

Corresponding Author:
Siti Rafiah Rangkuti,
Department Of Mathematics,
Universitas Islam Negeri Sumatera Utara Medan
Email: rafiahsiti@uinsu.ac.id

1. INTRODUCTION

Many factors affect Indonesia's public health. The many markers can make it hard to show and understand health in Indonesia [1] [2]. Too many variables can sometimes cause multicollinearity problems, which means that there are a lot of variables that are related to each other [3] [4]. Life Expectancy, Health Center Ratio, Hospital Ratio, Percentage of PHBS RT, Percentage of RT with Adequate Sanitation, Percentage of Low Birth Weight Infants, Percentage of Infants Receiving Exclusive Breastfeeding, and Diarrhea Morbidity Rate (per 1000 population) are all factors that are looked at when figuring out health characteristics [5]. Taking these factors into account, we can identify key indicators to evaluate the health characteristics in Indonesia [6].

The issue can be solved by reducing the factors that influence the health characteristics of the Indonesian people [7]. The Principal Component Analysis (PCA) method will be applied to this issue [6]. Principal Component Analysis is an analysis of statistics designed for minimizing the dimension of a dataset while conserving its essential qualities [8]. Alongside reduction of dimensionality Principal Component Analysis can also mitigate multicollinearity concerns in Multiple Linear Regression Analysis [9]. Principal Component Analysis continues to be refined for factor reduction [10].

2. RESEARCH METHOD

The research object of this study is Indonesian health data from 2024, obtained from secondary data sourced from official government agencies, such as the Ministry of Health of the Republic of Indonesia or other relevant institutions [11]. The data includes various health indicators representing the health conditions of the Indonesian population, such as healthcare service indicators, public health status, demographic factors, and environmental factors. The dataset used is multivariate, with a relatively large number of variables and the potential for correlation between variables [12].

Principal component analysis explains the structure of the variance-covariance matrix of a set of variables in linear combinations of those variables. Principal components are used for variable reduction and interpretation [13]. Assume there are p variables comprising n objects. Consider that from these p variables, k principal components (where $k \leq p$) are generated, which are linear combinations of the p variables. The k principal components can substitute the p variables from in which they originate without a substantial reduction of information regarding the overall variables [14]. principle component analysis is typically an intermediate analysis, indicating that the outcomes of the principal components can be utilized for later analyses. In this case using software R studio to analysis the data to get the information representing the health condition [15].

2.1 Calculating value of the Bartlett Test of Sphericity and the Kaiser-Meyer-Olkin (KMO)

Before conducting the principal component analysis, it utilizes a correlation matrix. The first stage in principal component analysis is the construction of the correlation matrix [16]. This matrix is utilized to derive the proximity values of the correlations among the study variables. The proximity value can be utilized to perform various tests to assess the alignment with the correlation values derived from the principal component analysis.

2.2 Bartlett's Test

The Bartlett's test is used to determine whether the correlation matrix is an identity matrix. This test is used when most of the correlation coefficients are less than 0.5 [17].

Hypothesis:

H_0: There is no significant correlation between the variables

H_1: There is a significant correlation between the variables

Statistics test:

$$x^2_{obs} = -\left[(N-1) - \frac{(2p+5)}{6}\right] \ln |R|$$

Where:

N = Amount observation

p = Consist of variables

$|R|$ = Determinant of correlation matrices

Following the decision criterion of not rejecting H_0 at the significance level α, the subsequent step is to perform the Bartlett test of sphericity, utilized to assess the correlation among the variables in the sample.

2.3 Kaise Meyer Oklin (KMO) Test

This test analyzes the appropriateness of observational data for analysis via principal component analysis. The Kaiser-Meyer-Olkin (KMO) statistic quantifies sample adequacy, utilizing the following formula:

$$KMO = \frac{\sum_i \sum_j r_{ij}^2}{\sum_i \sum_j r_{ij}^2 + \sum_i \sum_j \alpha_{ij}^2} \tag{1}$$

which:

r_{ij} = Coefficcient correlation between variable i and j

a_{ij} = Partial coefficcient correlation between variable i and j

The accomplishment of a minor partial correlation coefficient relative to the correlation coefficient will yield a KMO value nearing 1. A low KMO score suggests that the application of factor analysis should be reevaluated, as the interrelations among variables cannot be elucidated by other variables. The criteria for decision-making are as follows.

Table 1: Decision Criteria

Value of KMO	Interpretation
KMO > 0.9	The data is very good.
$0.8 <$ KMO ≤ 0.9	Good data
$0.7 <$ KMO ≤ 0.8	The data is quite good.
$0.6 <$ KMO ≤ 0.7	Insufficient Data
$0.8 <$ KMO ≤ 0.6	Bad data
KMO < 0.5	Data is not eligible.

2.4 Evaluation of Component Factors by Eigenvalues

The eigenvalue signifies the degree of influence a variable exerts on the development of qualities represented by λ. Factor extraction is a prevalent technique employed to identify eigenvalues that are greater than or equal to 1 or 0, as well as to analyze the scatter plot. Factors with an eigenvalue exceeding 1 are preserved, whilst those with an eigenvalue below 1 are excluded from the model. An eigenvalue signifies the extent to which a factor contributes to the variance of all original variables. Only factors exhibiting a variance over 1 are used into the model. Factors exhibiting a variance below 1 are suboptimal, as the original variables have been normalized, resulting in a mean of 0 and a variance of 1.

2.5 Determination of Principal Component Analysis

Three methods are employed to ascertain the quantity of main components for subsequent investigation. The initial approach involves examining the explained variance, which ought to exceed 80%. The alternative approach involves examining the eigenvalues that exceed one. The third way involves examining the scree plot, particularly the elbow point indicated on it. This study determines the number of principle components generated in principle Component Analysis (PCA) by examining eigenvalues exceeding one.

3. RESULT AND ANALYSIS

This research covers all provinces in Indonesia, which number 34. Data obtained in 2024, which describes health characteristics in Indonesia. The secondary data used comes from the 2024 Indonesian Health Profile publication. This publication is issued by the Ministry of the Republic of Indonesia and includes data and indicators from each relevant province.

The following are the indicators found in the publication "Indonesia Health Profile." This study did not include all available indicators, but only selected ones:

Table 2: Indonesia Health Profile Indicators

No	Indicator Name	Indicator Explanation
1	Life expectancy	The average number of years a person who reaches age X in a given year is expected to live, based on the mortality conditions prevailing in the community.
2	Health center ratio	Comparison between the number of community health centers and the population in a province.
3	Hospital ratio	Comparison between the number of hospitals and the population in a province.
4	Percentage of households with clean and healthy lifestyle behaviors	Households that implement 10 indicators of clean and healthy living.
5	Percentage of households with Adequate Sanitation	Households using communal toilets, swan-neck toilets, and septic tanks.
6	Percentage of low-birth-weight babies	Babies weighing less than 2.5 kg at birth.
7	Percentage of infants exclusively breastfed	Exclusive breastfeeding without other foods or drinks until the baby is 6 months old, and then continuing until the baby is 2 years old, even after the baby starts eating.
8	Diarrhea morbidity rate	Bowel movements more than three times a day with loose stools.

Based on the results of the Kaiser–Meyer–Olkin (KMO) Measure of Sampling Adequacy test, an Overall MSA value of 0.65 was obtained. This value falls within the range of $0.6 < \text{KMO} \leq 0.7$, indicating that the data is categorized as sufficient (insufficient but acceptable) for factor analysis or Principal Component Analysis (PCA). Therefore, in general, the data meets the minimum eligibility criteria for further analysis using PCA.

Next, the MSA values for each variable show varying levels of feasibility. Variables X2 (0.74), X5 (0.71), X6 (0.69), X7 (0.63), and X8 (0.88) have MSA values above 0.6, indicating that they are feasible and make a good contribution to factor formation. Variable X8 even shows a very good MSA value, indicating a strong correlation with other variables in the data structure.

However, variables X3 (0.37) and X4 (0.34) have MSA values below the minimum threshold of 0.5. This indicates that both variables contribute less to factor formation and have relatively high partial correlations. This condition needs to be addressed, as variables with low MSA values can potentially reduce the quality of the factor analysis results.

Furthermore, the results of Bartlett's Test of Sphericity show a test statistic value of $X^2 = 127.5629$ with degrees of freedom (df) = 28 and a p-value of 1.158031×10^{-14}. The p-value, which is much smaller than the significance level of 0.05, leads to the rejection of H_0, thus concluding that there is a significant correlation between the variables in the data. In other words, the correlation matrix is not an identity matrix, so the data meets the assumption for factor analysis or PCA. Overall, the results of the KMO and Bartlett's Test indicate that the data are sufficiently suitable for analysis using Principal Component Analysis, although there are several variables with low MSA values that need to be considered in the advanced analysis stage.

```
eigen() decomposition
$values
[1] 3.3158429 1.447092 1.0700606 0.8132127 0.6402824 0.4166662 0.1038008

$vectors:
            [,1]      [,2]       [,3]       [,4]       [,5]        [,7]
[1,] -0.5125583 -0.07707531 0.004964793 0.10618311 0.06662279 -0.82722268
[2,]  0.3773977 -0.17821059 0.316931573 0.27935746 0.23343895  0.63666137
[3,] -0.2081658 -0.59298809 0.424915789 0.25540126 0.29436378 -0.12524279
[4,] -0.1480435 -0.29140871 0.672087564 0.62266437 0.22269538 -0.11712430
[5,] -0.4555447 -0.29675941 0.670284747 0.14669388 0.27666338 -0.09411002
[6,] -0.4581222 -0.45976923 0.360416799 0.55700891 0.18891390 -0.67007004
[7,] -0.4278222 -0.28407629 0.333309918 0.14669384 0.02103340 -0.51663719
[8,] -0.2962181 -0.36473825 0.124388393 0.09314834 0.05294548  0.09513617
            [,0]
X *** eigen() vectors
[1,]   0.73401550
[2,]  -0.10931735
[3,]  -0.21772225
[4,]   0.19942981
[5,]   0.01315199
[6,]  -0.22047004
[8,]  -0.04555882
```

Figure 1: Matrix Varians and Covarian

Eigenvalues are used to describe how much variance can be captured by the principal components. Eigenvectors, on the other hand, are the coefficients used to form the principal components from the variables, acting as loadings. Eigenvalues are typically used to create points on a Scree Plot, which serves as a reference for plotting the k-th principal component against its eigenvalue.

Regarding the method for determining the number of principal components, by looking at the proportion of variance of the principal components obtained from the eigenvalues of the principal components corresponding to the total of all eigenvalues, we obtain principal components that cover at least 80% of the cumulative variance in the data or at least 80% of the data's variability.

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC7	PC8
Standard deviation	1.8209	1.2029	1.0344	0.9018	0.80130	0.4396	0.32218
Proportion of Variance	0.4145	0.1809	0.1338	0.1017	0.08026	0.5206	0.02294
Cumulative Proportion	0.4145	0.5954	0.7291	0.8308	0.91103	0.96308	1.00000
Cumulative Proportion	0.4145	0.5954	0.7291	0.8308	0.91103	0.96308	1.00000

Figure 2: Determining the Value of the Principal Component Function

Based on the information obtained, the number of main components extracted was 4. This is in accordance with the minimum coverage of 80 percent of the data variance. These four components are able to capture 83.08 percent of the total data variability.

Additionally, the method for determining the number of principal components using a Scree Plot was also obtained. A Scree Plot represents the principal components with their eigenvalue variances. The number of principal components is determined from the extreme point where the curve line begins to flatten or is insignificant in adding to the total data variance. The scree plot is shown in the following graph:

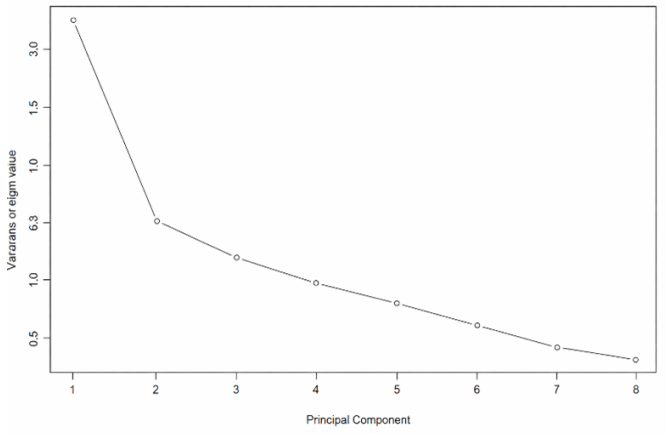

Figure 3: Creating a Screeplot

Based on the graph above, the main components extracted are 4. This is indicated by the extreme point where the curve line begins to flatten, shown at the 4th component. Therefore, the number of components selected based on the cumulative variance proportion method with the Scree Plot method is 4 main components.

3.1 Calculating Principal Component Coefficients

The equation of the principal components formed is used to interpret each principal component. It can also be used to name the principal components. As a result, the principal component equations are obtained based on the values of the largest variable coefficients because they are no longer influenced by the variance of the respective variable. Here is the equation formed after standardization. Main Component 1 (MC_1)

$$MC_1 = -00770X_1 - 0.1782X_2 + 0.5928X_3 - 0.2941X_4 + 0.2967X_5 + 0.4569X_6 - 0.2816X_7 + 0.38473X_8$$

This component describes the hospital ratio size, the incidence of diarrhea, and the number of low birth weight (LBW) babies. These three indicators are the largest indicators in component MC_1. Based on this information, MC_1 is named Population Health Services.

Main Component 2 (MC_2)

$$MC_2 = 0.0049X_1 - 0.3815X_2 + 0.4249X_3 + 0.6708X_4 + 0.0702X_5 - 0.3686X_6 - 0.2598X_7 - 0.1243X_8$$

This component is a measure of the ratio of hospitals, households with clean and healthy living behaviors, and low birth weight babies. Based on these indicators, it is the largest indicator within the component. Therefore, this information, MC_2, is referred to as Healthy Living Behaviors and Infant Health Conditions.

Main Component 3 (MC_3)

$$MC_6 = -0.2877X_1 + 0.6366X_2 + 0.1225X_3 - 0.1417X_4 - 0.023X_5 - 0.4482X_6 - 0.5166X_7 + 0.0961X_8$$

This component is a measure of the ratio of health centers, the ratio of hospitals, and the incidence of diarrhea. It is the largest indicator in this component. Based on that information, it is referred to as MC_3 and is named as a Health Facility.

$$MC_8 = 0.7349X_1 - 0.1029X_2 - 0.2771X_3 - 0.1204X_4 + 0.0318X_5 + 0.02017X_6 - 0.5950X_7 - 0.0495X_8$$

This component is a measure of life expectancy, households with adequate sanitation, and infants with low birth weight. Based on that information, it is the largest indicator in that component. Thus, $MC_4$is obtained, which is referred to as Community Group Health.

Based on principal component analysis, it is known that the Indonesian health data used in this study, consisting of 8 main components, can be reduced to 4 indicators while still representing the diversity of the initial data. The four indicators include Population Health Services, Healthy Lifestyle and Infant Health Conditions, Health Facilities, and Community Group Health.

4. CONCLUSION

This research shows that applying Principal Component Analysis (PCA) is effective in reducing the complexity of Indonesia's 2024 multidimensional health data, which has correlations between variables. Thru the dimension reduction process, PCA is able to simplify the data structure while retaining most of the important information, making the data easier to analyze and interpret without losing the substantive meaning of the health indicators being analyzed.

Additionally, the results of principal component analysis-based visualization show that PCA can present patterns, trends, and structures in health data more clearly and informatively. Visualization in low-dimensional space allows for the identification of dominant characteristics and differences in health conditions between regions or indicators. Thus, PCA serves not only as a statistical technique but also as an analytical tool supporting data-driven decision-making in the management and formulation of national health policies.

REFERENCES

[1] N. H. Alfajr and S. Defiyanti, "Heart disease prediction using random forest and pca," *Jurnal Informatika dan Teknik Elektro Terapan*, 2025.

[2] Anonymous, "Data reduction using pca: theoretical underpinnings in public health," *JCMEDU*, 2024.

[3] U. Hasanah *et al.*, "Pca and clustering on health and socio-economic indicators in east java," *Techno.Com*, 2024.

[4] A. Jamil *et al.*, "Empirical evaluation of dimensionality reduction for clinical narratives using pca, t-sne and umap," *Scientific Reports*, 2025. [Online]. Available: https://doi.org/10.1038/s41598-025-30537-w

[5] Z. Jayidan *et al.*, "Improving heart disease prediction using pca for feature extraction," *Jurnal Teknik Informatika*, 2024.

[6] A. A. Nasser, M. N. Aljober, A. S. A. Alghawli, and A. A. K. Essayed, "Revealing principal components, patterns, and structural gaps in health security among high-income countries using pca," *F1000Research*, vol. 14, p. 769, 2025. [Online]. Available: https://doi.org/10.12688/f1000research.168082.2

[7] G. S. M. Khamis, "Using fuzzy c-means clustering and pca in public health," *Heliyon*, 2025.

[8] B. Koichubekov *et al.*, "Functional pca for forecasting healthcare workforce needs," *International Journal of Environmental Research and Public Health*, 2025.

[9] F. A. Shalih *et al.*, "Principal component analysis: methods and application review," *JEM, E-Journal UPI*, 2024.

[10] S. Wibowo *et al.*, "Cardiovascular health patterns using pca and k-medoids clustering," *Jurnal Ilmu Komputer dan Sistem Informasi*, 2025.

[11] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A*, vol. 374, no. 2065, 2016. [Online]. Available: https://doi.org/10.1098/rsta.2015.0202

[12] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1–3, pp. 37–52, 1987. [Online]. Available: https://doi.org/10.1016/0169-7439(87)80084-9

[13] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010. [Online]. Available: https://doi.org/10.1002/wics.101

[14] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 11, pp. 559–572, 1901. [Online]. Available: https://doi.org/10.1080/14786440109462720

[15] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 1933. [Online]. Available: https://doi.org/10.1037/h0071325

[16] M. Ringner, "What is principal component analysis?" *Nature Biotechnology*, vol. 26, pp. 303–304, 2008. [Online]. Available: https://doi.org/10.1038/nbt0308-303

[17] J. Shlens, "A tutorial on principal component analysis," *arXiv preprint*, 2014. [Online]. Available: https://doi.org/10.48550/arXiv.1404.1100