Mathematical Journal of Modelling and Forecasting
Vol. 2, No. 2, December 2024, pp. 30-36

Application of Principal Component Analysis in Identifying Factors Affecting the Human Development Index

Muhammad Faisal[1,*] , Fadhilah Fitri[1], and Zilrahmi[1]

[1] Departement of Statistic, Faculty of Mathematics and Natural Sciences, Universitas Negeri Padang, Indonesia
[*]mhmdfaisal864@gmail.com

Abstract. This study examines the Human Development Index (HDI) in West Sumatra Province in 2023. The HDI is an essential indicator for measuring the success of efforts to improve the quality of human life. This research aims to identify the key factors that influence the HDI. The HDI is constructed from three fundamental dimensions that indicate human quality of life: health, education, and economy. The factors within each dimension tend to be strongly correlated, as they mutually influence one another, potentially leading to multicollinearity issues. Therefore, an analysis is conducted to reduce the number of original variables into new orthogonal variables while preserving the total variance of the original variables using Principal Component Analysis (PCA). Based on this background, the study applies PCA to address multicollinearity and to identify new, more representative variables. The study findings indicate that the factors influencing the HDI are the education and economic and health welfare indexes.

Keywords: Human Development Index, Principal Component Analysis, Eigen Value, Scree Plot

1 Introduction

The Human Development Index (HDI) is an important indicator for measuring success in efforts to improve human quality of life. The HDI is built from three basic dimensions used as measures of human quality of life. The dimensions include health, education, and economy [1]. In this study, to measure the health dimension, the variable that can be used is the number of health facilities [2]. Meanwhile, the education dimension can be assessed through the gross enrollment ratio and net enrollment ratio [3]. Apart from these two dimensions, there is the economic dimension, where the economic dimension can be measured with variables such as the percentage of the poor population, the number of people aged 15 and above who are employed, and the labor force participation rate [4].

The factors in each basic dimension of the Human Development Index (HDI) tend to have a strong relationship with each other, because they influence each other, which can cause multicollinearity problems [4]. Multicollinearity is a condition where there is a correlation between independent variables [5]. Therefore, it is carried out to reduce several original variables into new orthogonal variables while maintaining the total variance of the original variables by using Principal Component Analysis (PCA) [6]. PCA is a statistical method aimed at transforming a number of correlated original variables into a smaller set of new independent variables (not correlated with each other) [7]. Several previous studies on Principal Component Analysis have been conducted, such as Dwi Retno's research, which reduced the variables affecting tapioca production from 4 variables to 2 variables [8]. Sudianto, Dita, and Hanifah conducted research using Principal Component Analysis to reduce the variables affecting digital library services from 12 variables to 2 variables [7]. The next study was also by Wangge, who reduced the variables affecting the duration of thesis completion by Mathematics Education students at FKIP UNDANA from 13 variables to 10 variables [9]. Based on that background, this research uses Principal Component Analysis to address multicollinearity and identify new, more representative variables.

2 Research Methods

2.1. Types and Sources of Data

In this study, the data used is secondary data for 2023 sourced from *www.sumbar.bps.go.id*. The data contains 19 Regency/City data in West Sumatra Province. The following table below are the variables used in the study.

Table 1. Research Variables

Variable	Description
X_1	Health Facilities
X_2	Labor Force Participation Rate
X_3	Percentage of Poor Population
X_4	Population aged 15 years and above who are employed
X_5	Pure Participation Rate (PPR) with Senior High School
X_6	Gross Participation Rate (GPR) with Senior High School

The selection of these variables refers to previous studies [2], [3], [4]which show that the number of health facilities, PPR, GPR, percentage of poor population, number of population aged 15 years and above who are employed, and labor force participation rate are valid representations to measure the health, education, and economic dimensions of HDI.

2.2. Research Procedures

Principal Component Analysis (PCA) was first introduced by Karl Pearson in 1901 for analysis purposes in the field of biology. Over time, the method was adapted by Karhunen in 1947 and further developed by Loeve in 1963. Therefore, in telecommunications, PCA is often referred to as the Karhunen-Loeve transformation [10]. PCA is a multivariate statistical analysis that can free data from multicollinearity, which is done by reducing (simplifying) a number of independent variables by transformation and regrouping and then giving a new identity to the main components formed by reviewing the dominant characteristics of the variables that compose it [11]. PCA is one of the methods used to reduce data dimensions at the pre-processing stage [12]. The problem that often arises in the process of reducing factors or variables is how to minimize the number of variables but still retain important information or characters contained in the data [13]. After simplification, PCA will find that the orthogonal basis becomes a new basis [14].

Research procedures in PCA:

1. Collect data from www.sumbar.bps.go.id
2. Descriptive data.
3. Data standardization if data units are different.
4. Determining the covariance matrix and correlation matrix (different variable units).
   a) Covariance matrix

$$S_{ik} = \frac{\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{n-1} \qquad i = 1, 2, \ldots, p \qquad k = 1, 2, \ldots, p \tag{1}$$

$$\boldsymbol{S} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{bmatrix}$$

   b) Correlation matrix

$$r = \frac{S_{ik}}{\sqrt{S_{ii}}\sqrt{S_{kk}}} = \frac{\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)^2}\sqrt{\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2}} \tag{2}$$

$$\boldsymbol{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

5. Determine eigenvalues and eigenvectors.
If A is an n×n matrix then the nonzero vector x in R^n which is the eigenvector of A if A$\underline{x}$ is a scalar multiple of $\underline{x}$, i.e. :

$$A\underline{x} = \lambda\underline{x} \quad (3)$$

A scalar λ is called an eigenvalue of A and $\underline{x}$ is said to be the eigen vector corresponding to λ.

6. Determine the number of principal components with 3 methods:
   a) Using eigen value > 1 .
   b) Using the proportion of cumulative variance.
   According to Johnson, the minimum percentage of diversity that can be explained is 80%. The value of the proportion of cumulative variance is obtained by means of:

$$prop_k = \frac{\sum_{k=1}^{p} \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \qquad k = 1, 2, \ldots, p \quad (4)$$

   c) Using scree plot.
   The screeplot is a graph between the eigenvalue λ_k and k. In this method, the number of principal components chosen is k, which is a left-skewed curve but right-skewed at point k. The idea behind this method is to select the number of principal components so that the difference between successive eigenvalues is no longer too large. Determining the right number of components, look for elbows (bends) in the scree plot [15].

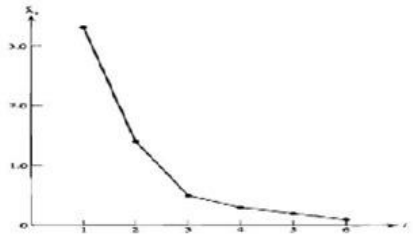

Fig 1. Scree Plot

7. Interpretation of research results.
8. Conclusion and suggestions.

3 Results and Discussion

3.1 Descriptive Analysis

The first step in a study is descriptive analysis. The purpose of descriptive analysis is to understand the meaning of data through its characteristics and description. Descriptive data can be seen from the minimum, maximum, average (mean) and standard deviation values of each independent variable in the study [11]. The following below is a descriptive analysis table of the data variables used in the study.

Table 2. Descriptive Analysis

Variable	Min.	Mean	Max.	Sd
X_1	45.00	175.20	387.00	101.12
X_2	65.00	70.69	80.07	4.12
X_3	2.27	5.90	13.72	2.39
X_4	27035	149733	426765	10495.5
X_5	56.74	69.21	84.41	7.83
X_6	77.79	91.97	114.29	8.59

Based on the table above, the minimum value shows that the lowest percentage of research data is in the variable, and vice versa for the maximum value. Of all the data variables, there is variable X_4 which has the highest data diversity because the standard deviation value is the largest. While the mean value of each variable is greater than the standard deviation. Then the data can be used as a good data representation.

3.2 Determining the Covariance Matrix and Correlation Matrix

In the process of determining principal component analysis, the covariance matrix must first be determined. The covariance matrix is used to measure the magnitude of the relationship between two variables. The matrix below is a covariance matrix which uses standardized data.

$$S = \begin{bmatrix} 1.000 & -0.125 & 0.148 & 0.911 & -0.227 & -0.168 \\ -0.125 & 1.000 & 0.487 & -0.300 & -0.464 & -0.421 \\ 0.148 & 0.487 & 1.000 & 0.070 & -0.629 & -0.711 \\ 0.911 & -0.300 & 0.070 & 1.000 & -0.117 & -0.097 \\ -0.227 & -0.464 & -0.629 & -0.117 & 1.000 & 0.864 \\ -0.168 & -0.421 & -0.711 & 0.911 & -0.227 & -0.168 \end{bmatrix}$$

Variables are said to be linearly related to each other if the covariance value is 0. The table above shows that the variables contain positive covariance and negative covariance, so it can be concluded that there is no strong relationship between these variables. In this study, the data used is not the same unit, so it is necessary to find the correlation matrix value. The following is a correlation matrix.

$$R = \begin{bmatrix} 1.000 & -0.125 & 0.148 & 0.911 & -0.227 & -0.168 \\ -0.125 & 1.000 & 0.487 & -0.300 & -0.464 & -0.421 \\ 0.148 & 0.487 & 1.000 & 0.070 & -0.629 & -0.711 \\ 0.911 & -0.300 & 0.070 & 1.000 & -0.117 & -0.097 \\ -0.227 & -0.464 & -0.629 & -0.117 & 1.000 & 0.864 \\ -0.168 & -0.421 & -0.711 & 0.911 & -0.227 & -0.168 \end{bmatrix}$$

3.3 Determining Eigen Values and Eigen Vectors

The next step is to find the eigenvalue decomposition to make it easier to determine the number of main components. Table 3 below is the eigen value decomposition obtained from the correlation matrix.

Table 3. Eigen Value

Component	1	2	3	4	5	6
Eigen Value	2.87	2.00	0.56	0.39	0.12	0.07

Table 3 above shows that there are 6 variables or components that are analyzed in this study. In components 1 and 2 the eigen value is > 1, while components 3 - 6 eigen value < 1. The next step is to find the eigen vector value which will be the main component coefficient. The following is a eigen vector.

$$V_1 = \begin{bmatrix} -0.185 \\ -0.354 \\ -0.498 \\ -0.113 \\ 0.536 \\ -0.540 \end{bmatrix} \quad V_2 = \begin{bmatrix} 0.640 \\ -0.349 \\ -0.073 \\ 0.678 \\ 0.022 \\ -0.044 \end{bmatrix} \quad V_3 = \begin{bmatrix} -0.293 \\ -0.820 \\ -0.001 \\ -0.111 \\ -0.268 \\ -0.397 \end{bmatrix} \quad V_4 = \begin{bmatrix} 0.078 \\ 0.205 \\ -0.840 \\ -0.034 \\ -0.472 \\ -0.151 \end{bmatrix} \quad V_5 = \begin{bmatrix} 0.093 \\ -0.150 \\ 0.205 \\ -0.169 \\ -0.628 \\ 0.710 \end{bmatrix} \quad V_6 = \begin{bmatrix} -0.674 \\ 0.126 \\ -0.001 \\ 0.696 \\ -0.152 \\ 0.147 \end{bmatrix}$$

Based on the eigen vector above, the 1st component absorbs most of the variables, the 2nd component absorbs most of the remaining variance after absorbing the 1st component, and so on.

3.4 Determining the Number of Principal Components

Using Eigen Value > 1

The first step in determining the number of main components is the eigenvalue whose value is more than one. Based on table 3, the eigen value that is more than one is 2 main components.

Using Cumulative Proportion

The main component taken is the main component that covers at least 80% of the cumulative variance in the data or can be said to be able to capture at least 80% of the diversity of the data [15].

Table 4. Cumulative Proportion(%)

Component	1	2	3	4	5	6
Cumulative Proportion	0.4776	0.8104	0.9038	0.9686	0.9886	1.0000

Based on the information from table 4, the number of principal components that can be taken is 2 principal components. This is because 2 principal components have captured at least 80% of the data variance. One principal component has been able to capture 81.04% of the total data diversity.

Using Scree Plot

The Scree Plot is a plot between the principal component and the eigen value (variance). The number of principal components taken is at the extreme point where the curve line starts to slope.

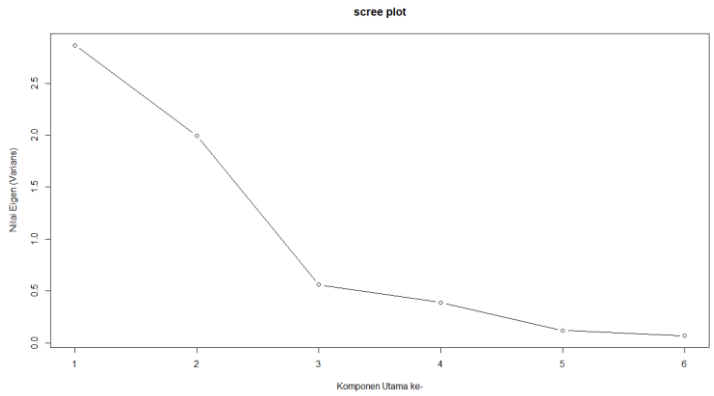

Fig 2. Scree Plot PCA

Based on the Scree Plot above, the number of principal components selected is 2. This is because the extreme point where the curve line begins to slope is shown in the 2nd component.

3.5 Number of Major Components

In determining the number of main components, the results obtained on the eigenvalue > 1 are 2 main components, the cumulative proportion is 2 main components, and based on the scree plot is 2 main components. So it can be concluded that from these three determinations, the number of main components is 2 main components.

3.6 Principal Component Formed

Step on HDI data consists of 6 variables; after analysis, 2 main components are formed. Below is a table of the main component equation, which is a result of simplifying the dimensions of the variables formed into several main components.

Table 5. Principal Component Formed

Component	Component Label	Variable	Variable Label
$\mathbf{PC_1} = -0.185X_1 - 0.354X_2 - 0.498X_3 - 0.113X_4 + 0.536X_5 + 0.540X_6$	Education Welfare Index	X_3	Percentage of Poor Population
		X_5	Pure Participation Rate (PPR) with Pure Participation Rate (PPR) with Senior High School
		X_6	Gross Participation Rate (GPR) with Senior High School
$\mathbf{PC_2} = 0.640X_1 - 0.349X_2 - 0.073X_3 + 0.678X_4 + 0.022X_5 + 0.044X_6$	Health and Economic Welfare Index	X_1	Health Facilities
		X_2	Labor Force Participation Rate
		X_4	Population aged 15 years and above who are employed

The results obtained show that there are two main component factors formed, namely the Education Welfare Index which represents data X_3, X_5, and X_6;. In contrast, the Economic and Health Welfare Index represents data X_1, X_2, and X_4. Based on the 2 principal component equations formed, it can be concluded that:

Principal Component 1 (PC_1)
PC_1 is correlated with the variables Percentage of Poor Population (X_3), Pure Participation Rate (PPR) with Senior High School (X_5) and Gross Participation Rate (GPR) with Senior High School (X_6). When PC_1 increases, this indicates that the number of these three variables increases. The factor formed in PC_1 is the Education Welfare Index factor.

Principal Component 2 (PC_2)
PC_2 is correlated with the variables Health Facilities (X_1), Labor Force Participation Rate (X_2) and Population Aged 15 Years and Above Who are Employed (X_4). When PC_2 increases, the sum of these three variables increases. The factors formed in PC_2 are the Health and Economic Welfare Index.

4 Conclusions and Suggestions

The discussion results show that factors affect the Human Development Index in West Sumatra Province, namely the Education Welfare Index and the Economic and Health Welfare Index. The new factors or indicators formed can be used for further regression analysis and have overcome the problem of multicollinearity because they are independent of each other.

Suggestions for further research include developing other methods for finding factors in a problem. In addition, the number of variables considered capable of influencing the Human Development Index (HDI) can be increased so that other new variables that affect HDI are obtained. The results of these studies can be compared with previous studies, including this study.

References

[1] T. Faizia, A. Prahutama, and H. Yasin, "PEMODELAN INDEKS PEMBANGUNAN MANUSIA DI JAWA TENGAH DENGAN REGRESI KOMPONEN UTAMA ROBUST," vol. 8, no. 2, pp. 253–271, 2019, [Online]. Available: http://ejournal3.undip.ac.id/index.php/gaussian

[2] Dewi Khoirunnisa, "Analisis Pengaruh Angka Partisipasi Murni, Fasilitas Kesehatan dan Pertumbuhan Ekonomi Terhadap Indeks Pembangunan Manusia di Pulau Sulawesi Tahun 2011 - 2020," UIN syarif Hidayatullah, Jakarta, 2022.

[3] Rasdi Ekosiswoyo, Kardoyo, and Tri Joko Raharjo, "Strategi Akselerasi Pencapaian IPM Bidang Pendidikan untuk Mendukung Keberhasilan Pembangunan Jangka Menengah Kota Semarang," *Riptek*, vol. 1, no. 2, pp. 23–33, 2008.

[4] Dwi Maumere Putra, "Pemodelan Indeks Pembangunan Manusia (IPM) Provinsi Jawa Timur dengan Menggunakan Metode Regresi Logistik Ridge," Institut Teknologi Sepuluh November, Surabaya, 2015. Accessed: Oct. 24, 2024. [Online]. Available: https://repository.its.ac.id/59984/1/1311100108-Undergraduate%20Thesis.pdf

[5] Kusnandar and Dadan, *Metode Statistik dan Aplikasinya dengan Minitab dan Excel*. Yogyakarta: Madyan Press, 2003.

[6] Sigit Nugroho, *Statistika Multivariat Terapan*. Bengkulu: Unib Press.h.1, 2008.

[7] S. Manullang, D. Aryani, and H. Rusydah, "Analisis Principal Component Analysis (PCA) dalam Penentuan Faktor Kepuasan Pengunjung terhadap Layanan Perpustakaan Digilib," *Edumatic: Jurnal Pendidikan Informatika*, vol. 7, no. 1, pp. 123–130, Jun. 2023, doi: 10.29408/edumatic.v7i1.14839.

[8] D. R. P. Sari, "METODE PRINCIPAL COMPONENT ANALYSIS (PCA) SEBAGAI PENANGANAN ASUMSI MULTIKOLINEARITAS," *PARAMETER: Jurnal Matematika, Statistika dan Terapannya*, vol. 2, no. 02, pp. 115–124, Nov. 2023, doi: 10.30598/parameterv2i02pp115-124.

[9] M. Wangge, "Penerapan Metode Principal Component Analysis (PCA) Terhadap Faktor-faktor yang Mempengaruhi Lamanya Penyelesaian Skripsi Mahasiswa Program Studi Pendidikan Matematika FKIP UNDANA," *Jurnal Cendekia : Jurnal Pendidikan Matematika*, vol. 5, no. 2, pp. 974–988, Apr. 2021, doi: 10.31004/cendekia.v5i2.465.

[10] M. S. Noya van Delsen, A. Z. Wattimena, and S. Saputri, "PENGGUNAAN METODE ANALISIS KOMPONEN UTAMA UNTUK MEREDUKSI FAKTOR-FAKTOR INFLASI DI KOTA AMBON," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 11, no. 2, pp. 109–118, Dec. 2017, doi: 10.30598/barekengvol11iss2pp109-118.

[11] N. I. Tahir, Wahidah Alwi, and Khalilah Nurfadilah, "APLIKASI METODE ANALISIS KOMPONEN UTAMA (AKU) DALAM MENGIDENTIFIKASI FAKTOR YANG MEMENGARUHI KEMISKINAN DI KABUPATEN/KOTA PROVINSI SULAWESI SELATAN," *Journal of Mathematics: Theory and Applications*, pp. 38–44, Dec. 2021, doi: 10.31605/jomta.v3i2.1222.

[12] D. Sartika and I. Saluza, "Penerapan Metode Principal Component Analysis (PCA) Pada Klasifikasi Status Kredit Nasabah Bank Sumsel Babel Cabang KM 12 Palembang Menggunakan Metode Decision Tree," *Generic*, vol. 14, no. 2, pp. 45–49, Jul. 2022, doi: 10.18495/generic.v14i2.130.

[13] D. R. P. Sari, "METODE PRINCIPAL COMPONENT ANALYSIS (PCA) SEBAGAI PENANGANAN ASUMSI MULTIKOLINEARITAS," *PARAMETER: Jurnal Matematika, Statistika dan Terapannya*, vol. 2, no. 02, pp. 115–124, Nov. 2023, doi: 10.30598/parameterv2i02pp115-124.

[14] Mahmoudi, "Principal Component Analysis to study the relations between the spreas rates of COVID-19 in high risks countries," *Alexandria Engineering Journal*, vol. 60, no. 1, 2021.

[15] Richard A. Johnson and Dean W. Wichern, *Applied Multivariate Statistical Analysis*. New Jersey: University of Wisconsin, 1982.