Zero: Jurnal Sains, Matematika, dan Terapan
E-ISSN: 2580-5754
P-ISSN: 2580-569X
Vol. 8, No. 1, June 2024
DOI: 10.30829/zero.v8i1.20713
pp. 26-33

Analysis Of Factors Affecting Breast Cancer At Haji General Hospital Medan Using The Principal Component Analysis Method

[1]Khairiyah Nurfalija Daulay
[1]Department of Mathematics, State University of Medan, Medan, Indonesia

Article Info

Article history:

Received 04 24, 2024
Revised 06 25, 2024
Accepted 06 30, 2024

Keywords:

Breast Cancer, Analysis, PCA, Detect.

ABSTRACT

Based on data from the International Agency for Research on Cancer in Globocan (Globocan cancer Statistics) in 2020, breast cancer was ranked first with 68,858 cases (16.6%) of the total 396,914 new cases of cancer in Indonesia. This study aims to determine the factors that have the greatest influence on breast cancer patients at Haji General Hospital Medan. This research was conducted using factor analysis with the Principal Component Analysis (PCA) method. It is hoped that this will provide additional information for the pathology team at Haji Hospital Medan in detecting breast cancer. The population that was also sampled in this study was ≤ 153 data on breast cancer patients at Haji General Hospital Medan in 2022. The results of factor analysis using the PCA method showed that breast cancer at Haji General Hospital Medan 2022 was influenced by three main factors amounting to 57.135011. The main factor that most influences breast cancer is gender with the component having the highest eigenvalue of 1.59666 with a total percentage of variance (cumulative percent of variance) of 22.63581.

Corresponding author:

Khairiyah Nurfalija Daulay
Department of Mathematics,
State University of Medan, Medan, Indonesia
Email: khairiyahdaulay@gmail.com

1. INTRODUCTION

Cancer is defined as the growth of abnormal cells that spread uncontrollably and have the potential to harm other body components [1]. One of the most commoncancers is breast cancer. Breast cancer is malignant cancer of the breast that originates from gland cells, gland ducts, and breast supporting tissue, but does not include breast skin [2]. The most common cancer disease in Indonesia according to data from the International Agency for Research on Cancer in Globocan (Globocan cancer Statistics) in 2020 is breast cancer. Of the 396,914 cases with 22,430 (9.6%) deaths due to this disease, there were around 65,858 new cases (16.6%) of breast cancer detected nationally, of which North Sumatra is in the top 10 highest contributors of all provinces in Indonesia. The increase in the incidence of breast cancer always increases every year in North Sumatra, especially the city of Medan. In 2021, breast cancer sufferers in the city of Medan will reach 824 cases. In 2022 there will beapproximately 100 cases of breast cancer at the Haji Provincial General Hospital (RSUP) Medan.

What is needed when analyzing the data above is a method that is able to analyze severalvariables and can also measure the relationship between variables. A statistical method that iscapable of analyzing several variables is multivariate statistics. In multivariate statistics, each method has its own benefits. Factor analysis is a multivariate statistical technique that is commonly used [3]. Principal component analysis is a multivariate statistic that can be used to describe how a set of uncorrelated data (parameters) can vary into several independent parameters (principal components) [4].

Factor analysis using principal components is a type of data analysis that can be used to extract assessment indicators to create new factors that are not correlated with each other and strengthen an assessment category. The degree of data variance in all indicators is calculated in the main components [5].

Previous research conducted by [6] entitled "Breast Cancer Detection with Feature Selection based on Principal Component Analysis and Random Forest" aims to achieve a high level of accuracy in breast cancer detection. This research uses a classification model, namely Principal Component Analysis based on Random Forest, to predict the problem under study, where the results of the model evaluation will be seen for its accuracy value. The findings of this study show that when Random Forest and logitboost feature selection are combined, the principal component analysis-based feature selection method significantly improves its classification performance.

Based on this background and considering previous research, in this research the author is interested in studying using the PCA (Principal Component Analysis) method to analyze several factors that influence breast cancer at Haji Medan General Hospital by considering factors that have the potential to cause breast cancer. So the the author will conduct research with the title "Analysis of Factors Affecting Breast Cancer at Haji General Hospital Medan Using the Principal Component Analysis Method".

2. RESEARCH METHODS

2.1. Factor Analysis

Factor analysis is a part of the multivariate statistical analysis technique that attempts to identify relationshipsbetween several variables that previously did not depend on each other to create one variable or group (Simarmata et al., 2015).

The following initial model of factor analysis is as follows:

$$X_i = B_i F_1 + B_{i2} F_2 + \cdots + B_{ij} F_j + \cdots + B_{im} F_m + V_i \mu_i$$

Where

X_i = standardized i^{th} variable (mean = 0, standard deviation = 0)

B_i = partial regression coefficient carried out on the j^{th} common factor F_i = common factor J^{th}

V_i = factor coefficients that are frozen on the i^{th} unique factor

μ_i = unique factor of the i^{th} variable

m = many common factors

The methods contained in the factor analysis model are principal components, unweighted least squares, generalized least squares, maximum likelihood, principal axis factoring, alpha factoring, and image factoring. Among these methods, only two are often used in parameter estimation, namely PCA because it can overcome multicollinearity problems [7] and Maximum Likelihood can provide the best estimation results [8].

According to [8] because correlation is the basic idea of factor analysis, the correlational assumptions that will be applied are:

1. The correlation between independent variables must be strong enough, for example greater than 0.5.
2. The magnitude of the partial correlation, namely the correlation between two variables with the assumption that the other variables remain constant and must be small.
3. Measure Sampling Adequacy (MSA) or Barlett's Test of Sphericity is used to test the full correlation matrix.

The steps for carrying out factor analysis essentially consist of (Kusno, 2019: 63):

a. To find out how closely the variables are related to each other, a data matrix is compiled in the form of a correlation matrix between the original variables.
b. Barlett's Test of Sphericity, KMO, and MSA were used to test the relationship between several variables.
c. Extracting factors or factoring, to reduce data from several indicators (variables), produces smaller factors that can explain the relationship between the indicators studied.
d. Rotating factors is carried out if feature extraction (factoring) still does not obtain clear main factor components.
e. Interpreting factor rotation results (done by looking at factor loadings).

2.2. Principal Component Analysis (PCA)

There are two methods for extracting factors, namely principal component analysis and common factor analysis [9]. This method is quite effective, overcoming the problem of multicollinearity and eliminating correlation between independent variables until they are not correlated. The advantage of this method is that it can eliminate correlation without reducing or

eliminating the original variables. The detailed aim of PCA is to eliminate and can also be said to simplify factors that are less influential or less related without losing the reasons and objectives of the original data [10].

The steps in Principal Component Analysis (PCA) are as follows [11]: Calculates the variance covariance matrix from observational data

Variance is used $\left(V_{ar}(x)\right)$ to find the spread of data in a collection to determine the deviation of data in a sample data set. Covariance Matrix $\left(C_{ov}(x,y)\right)$ is a matrix in which the covariance values in each cell are obtained from the sample, provided that x and y are random variables [12].

$(\bar{x})$ and $(\bar{y})$ are respectively the sample average (mean) of the variables x and y. After calculating using the formula above, we get the n x n formulated as elow: matrix. The covariance matrix can also below:

$$\left(\mathrm{V_{ar}(x)}\right) = \sigma^2 = \frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}}\left(\mathrm{Z_{ij}} - \mu_{\mathrm{j}}\right)^2$$

$$\mathrm{C_{ov}(x,y)} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}}\left(\mathrm{X_{ij}} - \mu_{\mathrm{xj}}\right)\left(\mathrm{Y_{ij}} - \mu_{\mathrm{xj}}\right)$$

where
$N =$ number of observations
$X_i, Y_i =$ value of the observation μ_x, μ_y = mean vector
Calculate eigenvectors (V) and eigenvalues $(\boldsymbol{\lambda})$
This step is used after getting the covariance matrix value. The eigenvalues of $(\boldsymbol{\lambda})$ that have been computed are then transformed (orthogonal varimax rotation, which means minimizing the number of variables that have high loadings on a factor) using the following formula [11]:

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

Then, calculate the vector by solving the following equation:

$$[A - \boldsymbol{\lambda I}][\boldsymbol{X}] = \mathbf{0}$$
$$Ax = \boldsymbol{\lambda x}$$

Where
$A =$ matrix n x n
$I =$ Identity matrix
$\lambda =$ eigenvalue $(\boldsymbol{\lambda})$

Determining a new variable or what can be said to be the Principal Component (PC) by multiplying the original variable and the eigenvector matrix.

Last, from each eigenvalue of the new variable (PC), determine the proportion value of the Principal Component (PC) in (%) using the formula:

$$PC(\%) = \frac{NilaiEigen(\boldsymbol{\lambda})}{VarCov} \times 100\%$$

3. RESULTS AND ANALYSIS

In factor analysis using PCA, the first step was to collect data obtained by collecting information from the medical records of inpatient and outpatient patients at Haji Medan Hospital in 2022. The data collected was 153 data on breast cancer patients with several causal factors. The step after collecting data is checking data to see empty data (missing values). The results obtained were 143 data.

3.1 Correlation Matrix Between Variables

Next, form a correlation matrix with the aim of seeing the closeness of the relationship between variables.

a. Barlett's Test and KMO (Kaiser Meyer Olkin)

Bartlett's test of sphericity is a test used to test the correlation between variables inthe sample [13]. The requirement for the Barlett test is a sig value $< \alpha = (0{,}05)$. From the results of the Barlett test that has been carried out, it can be seen that the significance value is 0.0000, which means that there is a correlation betweenvariables and the process can continue.

The KMO test is carried out to see the adequacy of sampling in a study. From the results of the KMO test that has been carried out, it shows that this analysis is suitable foranalysis because the KMO value is > 0.5 and meets the criteria $0.5 < \text{KMO} \leq 0.6 =$ datato be analyzed (sufficient). Based on calculations using Python, the KMO value $= 0.522771$, meaning that the sampling adequacy is acceptable because it is greater than 0.5 . The following results in table 1 show that factor analysis is worthy of analysis.

Table 1. Barlett's test and KMO

	Nilai
KMO	0,522771
Approx Chi-Square	57,3281
Df	21,0000
p- *Value*	0,0000

b. MSA (Measure of Sampling Adequacy)

The next stage is testing the anti-image correlation matrix. The degree of correlation can be seen through the MSA value obtained from the results of table 3. Theseresults show that the MSA value is > 0,5, which means that seven variables have a strong correlation and can be analyzed further (Manullang et al., 2023).

Table 2. MSA values for 7 variables

Variable	**MSA Value**
Age	0,549204
Gender	0,509952
Genetics (Family History)	0,526959
Smoking Habits	0,506005
Marital Status	0,509817
Highest Level of Education	0,624533
Employment	0,54696

3.2 Factoring or Factor Extraction

a. Communalities

The next stage is the extraction process using the PCA method. The next stage after getting the MSA test value for each variable is to look for the communalities value. This value shows whether the variable being studied is able to explain the factors or not.A variable is considered capable if the extraction value is > 0.50. The higher the communalities value, the stronger the relationship between variables related to the resulting factor.

The results in table 3 show that the variable that has the highest communalities value is gender at 0,579526 = 57,9526%. This means that the last education variable can explain 57,9526% of the variance in the factors formed. The greater the value of communalities, the closer the relationship between variables and factors is formed. So, it can be concluded from table 3 that only the variables of gender and smoking habits canbe used to explain factors related to breast cancer.

Table 3. Communalities

Variable	Initial	Extraction
Age	1,0	0,254896
Gender	1,0	0,579526
Genetics (Family History)	1,0	0,132110
Smoking Habits	1,0	0,502474
Marital Status	1,0	0,207037
Highest Level of Education	1,0	0,175955
Employment	1,0	0,248604

b. Total Variance Explained

Next is the calculation of the total variance explained value to find out theeigenvalues and explained variance values for each attribute. To find out more specific extraction results using the PCA method, it can be seen from the total variance test which shows the eigenvalues. Therefore, only eigenvalues that are more than one or equal to oneare components that form factors. So from the 7 variables extracted in this study, three main components were formed, where component 1 had a value of 1,595666, component 2 was 1,293432, and component 3 was 1,138518 shown in the following table.

Table 4. Total Variance Explained

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	1,595666	22,635821	22,635821
2	1,293432	18,348382	40,984203
3	1,138518	16,150808	57,135011

c. Scree Plot

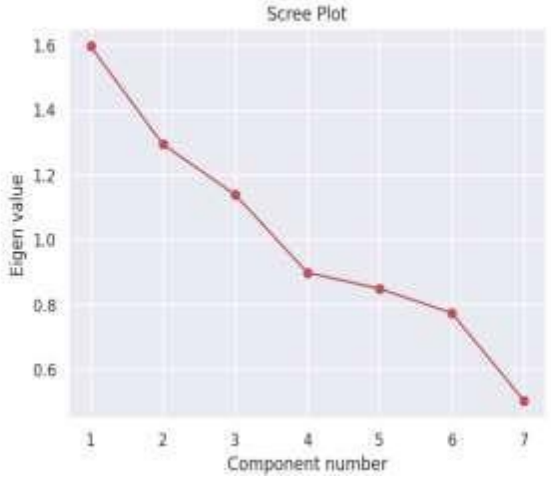

Figure 4.1: Scree Plot

Figure 4.1 shows the results of component points that have eigenvalues > 1, namely the three highest factors with values above 1. The image above shows the graphicform of the eigenvalues of each factor formed.

3.3 Factor Rotation

Before interpreting the factor results, the first step that must be taken is factor rotation to look for correlations between factors and variables. Factor rotation was also carried out to reviewthe placement of variables that were still not appropriate. Only correlation represented by loadingfactors (factors that correlate with each other) with a value > 0.30 is considered quite strongly correlated [14]. Factor rotation is also carried out to clarify the position of the variables without looking at the loading values. The following is a table of matrixcomponents before and after rotation. The results of factor rotation can be seen in table 5 and 6.

From table 5 of the component matrix below, it is still not very easy to find the appropriate place for the variables, for example in the Smoking Habits (X_4) and Marital Status (X_5) variables, the difference in loading values between variables is very small.

Table 5. Matrix Components Before Rotation

Variable	**1**	**2**	**3**
Age	0,292868	-0,499918	0,186189
Gender	0,643587	0,245831	-0,008245
Genetics (Family History)	0,001159	-0,171942	0,654749
Smoking Habits	0,584507	0,317105	-0,187595
Marital Status	0,225738	0,131222	0,654769
Highest Level of Education	-0,269073	0,422691	0,231263
Employment	0,187126	-0,603028	-0,138465

The rotational component matrix below shows a clearer and more appropriate distributionof variables. It can be seen from table 6 that the previously small loading factors are increasingly being reduced, and large loading factors are increasingly being enlarged.

Table 6. Matrix Components After Rotation

Variable	**1**	**2**	**3**
Age	-0,054685	0,635004	0,299782
Variable	**1**	**2**	**3**
Gender	0,843773	-0,090923	-0,117416
Genetics (Family History)	0,175548	0,077832	0,696968
Smoking Habits	-0,837988	0,014018	-0,088675
Marital Status	-0,226020	-0,111860	0,723848
Highest Level of Education	0,124753	-0,606150	0,147246
Employment	0,069274	0,732409	-0,054732

The result is that the 7 variables have been reduced, leaving only 3 factors, namely:

I. Factor 1 has only 1 forming variable consisting of: Gender.

II. Factor 2 has 3 forming variables consisting of: Age, Smoking Habits, and Occupation.

III. Factor 3 has 3 forming variables consisting of: Genetics (Family History), Marital Status,and Last Education.

The following table below explains which factor a variable will go into (factor group),namely:

Table 7. Rotation Result Factor Group

Variabel	1	2	3
Age		2	
Gender	1		
Genetics (Family History)			3
Smoking Habits		2	
Marital Status			3
Highest Level of Education			3
Variabel	**1**	**2**	**3**
Employment		2	

3.4 Interpretation of results

The loading value identifies the relationship between the factors formed and the variables. The higher the loading value, the closer the variable is to the factor. In the results above, all variables form a factor based on their largest loading value, so that the factors are interpreted in the table below as follows:

It can be seen in table 8, the main factor (component) formed is the gender factor with an eigenvalue of 1,595666. So, the factor that most influences breast cancer at Haji General Hospital Medan in 2022 is gender.

Tabel 8. Table of Variable Interpretation Results

No	Variabel	Eigen Values	Loading Faktor	% Variance	% Kumulatif
1	(X_2)	1,595666	0,635004	22,635821	22,635821
2	(X_4)	1,293432	0,843773	18,348382	40,984203
3	(X_1)		0,696968		
4	(X_7)		0,837988		
5	(X_3)		0,723848		
6	(X_5)	1,138518	0,147246	16,150808	57,135011
7	(X_6)		0,732409		

4. CONCLUSION

Several factors that influence breast cancer at RSUP Haji Medan consist of 7 variables which are classified into 3 factors, namely: the first factor with an eigenvalue of 1.595666%, the second factor with an eigenvalue of 1.293432%, and the third factor with an eigenvalue of 1.293432%. 1.138518%. The factor that most influences breast cancer at RSUP Haji Medan in 2022 is gender.

REFERENCE

[1] Zafrial, R. M., & Amalia, R. (2018). Artikel Tinjauan: Anti Kanker Dari Tanaman Herbal. Farmaka, 16(1), 15–23.

[2] Direktorat Pengendalian Penyakit Tidak Menular. (2014). Buku Saku Pencegahan Kanker Leher Rahim dan Kanker Payudara. Departemen Kesehatan RI.

[3] Wiratmanto. (2014). Analisis faktor dan Penerapannya dalam Mengidentifikasi FaktorFaktor yang Mempengaruhi Kepuasan Konsumen Terhadap Penjualan Media Pembelajaran. Universitas Negeri Yogyakarta.

[4] Nyoman Radiarta, I., Akhmad Mustafa, dan, Penelitian dan Pengembangan Perikanan Budidaya Jl Ragunan, P., Minggu, P., Selatan, J., & Penelitian dan Pengembangan Budidaya Air Payau, B. (2012). Sitakka No. 129, Maros 90512. In Februari.

[5] Wangge, M. (2021). Penerapan Metode Principal Component Analysis (PCA) Terhadap Faktor- faktor yang Mempengaruhi Lamanya Penyelesaian Skripsi Mahasiswa Program Studi Pendidikan Matematika FKIP UNDANA. 05(02), 974–988.

[6] Fauzi, A., Supriyadi, R., & Maulidah, N. (2020). Deteksi Penyakit Kanker Payudara dengan Seleksi Fitur berbasis Principal Component Analysis dan Random Forest. In Jurnal (Vol. 2, Issue 1). http://ejournal.bsi.ac.id/ejurnal/index.php/infortech96

[7] Kusuma, F. M., & Wibowo, A. (2017). kusuma. Biometrika Dan Kependudukan, 6(2), 89–97. Manullang, S., Aryani, D., & Rusydah, H. (2023). Analisis Principal Component Analysis (PCA) dalam Penentuan Faktor Kepuasan Pengunjung terhadap Layanan Perpustakaan Digilib. Edumatic: Jurnal Pendidikan Informatika, 7(1), 123–130. https://doi.org/10.29408/edumatic.v7i1.14839

[8] Santoso, S. (2012). Aplikasi SPSS pada Statistik Multivariat (Revisi). Elex Media Komputindo. Simarmata, I., Arma, A. J. A., & Arnita. (2015). arma. Kebijakan, Promosi Dan Biostatistika, 1(2), 1–10.

[9] Santoso, S. (2010). Mastering SPSS 18 (1st ed.). Elex Media Komputindo.

[10] Nasution, M. Z., Nababan, A. A., Syaliman, K. U., Novelan, M. S., Jannah, M., Dan Teknologi, S., Lunak, R. P., Informatika, T., Pancabudi, U. P., Jendral, J., & Subroto, G. (2019). Penerapan Principal Component Analysis (PCA) Dalam Penentuan Faktor Dominan Yang Mempengaruhi Pengidap Kanker Serviks (Studi Kasus: Cervical Cancer Dataset). Jurnal Mantik Penusa, 3(1), 204–210.

[11] Johnson, R. A., & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis (6th ed). Pearson Prentice Hall.

[12] Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). Springer Science & Business Media.

[13] Muliady Faisal, Syalam Ali Wira Dinata, & Dewi Ratna Sari. (2023). Analisis Komponen Utama Pada Dinas Ketenagakerjaan Bagian Penempatan Dan Perluasan Kerja Mencari Pekerjaan Menurut Golongan Pekerjaan. Journal of Innovation Research and Knowledge, 2(12), 4561–4568. https://doi.org/10.53625/jirk.v2i12.5627

[14] Fadilah, F., & Mahyuny, S. R. (2019). Analisis Faktor Yang Mempengaruhi Locus Of Control Mahasiswa Pendidikan Matematika FKIP Universitas Samudra. Jurnal IPA & Pembelajaran IPA, 2(2), 100–105. https://doi.org/10.24815/jipi.v2i1.10731