Available online at https://journal.rescollacomm.com/index.php/ijqrm/index

International Journal of Quantitative Research and Modeling

e-ISSN 2721-477X
p-ISSN 2722-5046

Vol. 5, No. 3, pp. 273-279, 2024

Determination of Dominant Factors Affecting Lung Cancer Patients Using Principal Component Analysis (PCA)

Moh Alfi Amal[1*], Nurnisaa binti Abdullah Suhaimi[2], Arla Aglia Yasmin[3]

[1,2,3]*Master of Mathematics Program, Faculty of Mathematics and Natural Sciences, Padjadjaran University, Sumedang 45363, West Java, Indonesia*

[*]*Corresponding author email: moh23006@mail.unpad.ac.id*

Abstract

The diagnosis of lung cancer is one of the most pressing health issues as the disease is often only detected at an advanced stage, leading to a poor prognosis for patients. Therefore, in an effort to detect, prevent, and manage the disease more effectively, this study utilized screening variables. Screening is an important endeavor in the early detection of diseases or abnormalities that are not yet clinically apparent using various tests, examinations, or procedures. The use of screening variables is very important in the early detection process because it can help in this study to understand the risk factors and causes of disease. The purpose of this study is to determine the dominant factors affecting people with lung cancer using Principal Component Analysis (PCA). Based on the results of the study, it was found that there are 20 dominant screening variables that have a considerable correlation to the formation of early detection of lung cancer with a total proportion of covariance variance of 100% when, the total variance obtained from the 20 screening variables is 100%. The final PCA results show that the factor loading values are used to determine which variables are most influential by comparing the magnitude of the correlation. As a result, the main factor causing lung cancer was Fatigue which had a factor loading of 7.87%, followed by the variables Age and Alcohol use with a factor loading of 6.02%. Other variables also showed certain factor loadings that indicated the cause of lung cancer. These findings are very important in efforts to improve early detection and management of lung cancer through more effective and targeted screening.

Keywords: Lung Cancer, Principal Component Analysis, Covariance Variance, Eigen Value, Eigen Vector

1. Introduction

Cancer is a malignant disease caused by the uncontrolled growth of cells in the body. According to the World Health Organization (WHO), cancer is one of the leading causes of death worldwide with nearly 10 million deaths by 2020, which means nearly one in six deaths is caused by cancer. The most common cancers are breast, lung, colon and rectum, and prostate cancer. Lung cancer is a malignancy of the lungs that occurs due to genetic changes in airway epithelial cells, resulting in uncontrolled cell proliferation. This malignancy can originate from the lung organ itself (primary) or from other parts of the body that spread to the lungs (metastasis). In Indonesia, lung cancer is the leading cause of death among men and third among women, as well as being the leading cause of death among all cancers. Therefore, screening and early detection are very important.

Symptoms of lung cancer, although often undetectable in the early stages, can include a chronic cough that does not go away, coughing up blood, shortness of breath, chest pain, excessive fatigue and pain that spreads throughout the body. As these symptoms are often ignored or dismissed as other respiratory problems, screening and early detection are crucial in identifying lung cancer early, increasing the chances of a cure and extending the life expectancy of sufferers.

It is important to know if cancer is present, especially if you have symptoms or risk factors. The difference between screening and detection is that screening is done on people without symptoms but at risk, while early detection is done on people with symptoms through follow-up examinations. Therefore, screening is crucial in recognizing the crucial first step in identifying lung cancer at an early stage, even before symptoms manifest. Through screening, individuals who are at high risk or exposed to certain risk factors can be identified, allowing for further investigations that may include medical tests or imaging. This provides an opportunity to detect cancer at an earlier stage, where treatment is often more effective and chances of cure are higher.

Therefore, early research to understand the dominant factors affecting lung cancer, as well as to develop effective screening methods, is an important step in efforts to prevent and control this disease. This study aims to simplify and eliminate less dominant or relevant screening factors or indicators using Principal Component Analysis (PCA), without compromising the essence and purpose of the original data.

2. Literature Review

2.1. Principal Component Analysis

Principal Component Analysis (PCA) is a multivariate technique that analyzes data tables in which observations are described by several correlated quantitative dependent variables. The goal is to extract important information from the table, to represent it as a new set of orthogonal variables called principal components, and to display the similarity pattern of the observations and variables as points on a map (Abdi and Williams, 2010).

PCA is concerned with explaining the variance-covariance structure of a set of variables through several linear combinations of these variables (Johnson and Wichern, 2007). By using PCA, n variables will be selected into k new variables called principal components, with the number of k being less than n. Using only k principal components will produce the same value as using n variables. The variable resulting from the selection is called principal component.

2.2. Variance-Covariance

Variance and covariance are two important concepts in statistics that are used to describe the characteristics of data distribution. Variance is a measure of the dispersion of a dataset from its mean. Variance measures how far the values in a dataset are spread around the average (mean) of the dataset.

Suppose x and y are random variables, then

$$Var(x) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \quad (1)$$

where x_i is each value in the dataset, μ is the average (mean) of the dataset, and n is the number of observations.

Covariance is a measure of how two variables change together. Covariance indicates the direction of the linear relationship between two variables (Jolliffe, 2002).

$$Cov(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y) \quad (2)$$

where x_i and y_i are the dataset values of different factors, μ_x and μ_y are the average (mean) of each factor, and n is the number of observations.

2.3. Eigenvalues and Eigenvectors

Definition 1. Suppose a matrix $A_{n\times n}$ and x is a nonzero vector in R^n and λ is a real scalar such that:

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

where λ is eigenvalue and **x** is eigenvector.

The calculated eigenvalues are then transformed using the eigenvector equation as follows:

$$Det(A - \lambda I) = 0 \quad (4)$$

where A is $n \times n$ matrix, I is identity matrix, and this eigenvector equation is the characteristic equation of λ.

3. Materials and Methods

3.1. Research Data

In this research, the data used is a dataset containing factors that can affect the outbreak of lung cancer. This lung cancer data involves 150 samples with 20 factors.

4. Results and Discussion

This study uses secondary data obtained from the Kaggle website. The data obtained consists of 150 samples and 20 factors. The factors are labeled as follows;

Table 1. Factors that can influence the development of lung cancer

No.	Factor	Lable
1	Age	X_1
2	Gender	X_2
3	Air Pollution	X_3
4	Alcohol use	X_4
5	Dust Allergy	X_5
6	Occupational Hazards	X_6
7	Genetic Risk	X_7
8	chronic Lung Disease	X_8
9	Balanced Diet	X_9
10	Obesity	X_{10}
11	Smoking	X_{11}
12	Passive Smoker	X_{12}
13	Chest Pain	X_{13}
14	Coughing of Blood	X_{14}
15	Fatigue	X_{15}
16	Weight Loss	X_{16}
17	Shortness of Breath	X_{17}
18	Swallowing Difficulty	X_{18}
19	Clubbing of Finger Nails	X_{19}
20	Dry Cough	X_{20}

4.1. Preprocessing data result

Table 2. Actual data

No.	X_1	X_2	X_3	X_4	X_5	...	X_{20}
1	33	1	2	4	5	...	3
2	17	1	3	1	5	...	7
3	35	1	4	5	6	...	7
4	37	1	7	7	7	...	7
5	46	1	6	8	7	...	2
⋮	⋮	⋮	⋮	⋮	⋮		⋮
150	32	1	2	3	6	...	5

Table 3. Mean and standard deviation from actual data

Factor	X_1	X_2	X_3	X_4	X_5	...	X_{20}
Mean	36.68	1.47	3.89	4.73	5.38	...	4.11
SD	12.74	0.50	1.97	2.58	1.83	...	2.08

Furthermore, the data is normalized by standardizing the data so that the data interval becomes more proportional. Data normalization uses the Z-Score method,

$$Z = \frac{(x-\mu)}{\sigma} \qquad (5)$$

where Z is the standardized value, x is the research data, μ is the average of each factor, and σ is the standard deviation per factor. The results of data normalization can be seen in Table 4.

Table 4. Normalized data

No.	X_1	X_2	X_3	X_4	X_5	...	X_{20}
1	-0.28892088	-0.94485058	-0.95939273	-0.28193962	-0.20716221	...	-0.53590535
2	-1.54509855	-0.94485058	-0.45267122	-1.44591053	-0.20716221	...	1.38950310
3	-0.13189866	-0.94485058	0.05405030	0.10605068	0.33800151	...	1.38950310
4	0.02512355	-0.94485058	1.57421483	0.8820313	0.88316525	...	1.38950310
5	0.73172350	-0.94485058	1.06749332	1.27002159	0.88316525	...	-1.01725747
⋮	⋮	⋮	⋮	⋮	⋮		⋮
150	-0.36743197	-0.94485058	-0.95939273	-0.66992992	0.33800151	...	0.42679888

Table 5. Mean and Variance from normalized data

Factor	X_1	X_2	X_3	X_4	X_5	...	X_{20}
Mean	1.2×10^{-17}	-1.4×10^{-16}	-3.1×10^{-16}	-1.8×10^{-16}	1.7×10^{-16}	...	4.7×10^{-17}
Var	1	1	1	1	1	...	1

4.2. Variance Covariance calculation results

Table 6. Variance Covariance calculation results

No	X_1	X_2	X_3	X_4	X_5	...	X_{20}
X_1	1	-0.091803	0.186602	0.259819	0.106929	...	-0.051883
X_2	-0.091803	1	-0.369480	-0.314960	-0.343123	...	-0.155070
X_3	0.186602	-0.369480	1	0.729182	0.636071	...	0.225599
X_4	0.259819	-0.314960	0.729182	1	0.787274	...	0.167516
X_5	0.106929	-0.343123	0.636071	0.787274	1	...	0.261606
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
X_{20}	-0.051883	-0.155070	0.225599	0.167516	0.261606	...	1

4.3. Eigenvalue calculation results

To determine the value of the Principal Component (PC) proportion, the following equation is used,

$$PC(\%) = \frac{Eigenvalue}{Variance-Covariance} \times 100\% \qquad (6)$$

where the variance-covariance is the sum of the diagonal values of the covariance matrix. The selected principal component has the maximum amount of contribution based on the proportion of variance of each selected principal component.

Table 7. Eigenvalue calculation results

Eigenvalue	**PC**	**PV**	**PK**
8.947083234	44.74%	0.447354	0.44735416
2.430628521	12.15%	0.121531	0.56888559
1.57454445	7.87%	0.078727	0.64761281
1.203666333	6.02%	0.060183	0.70779613
1.139334279	5.70%	0.056967	0.76476284
0.933940238	4.67%	0.046697	0.81145985
0.776519422	3.88%	0.038826	0.85028582
0.553241749	2.77%	0.027662	0.87794791
0.504965968	2.52%	0.025248	0.90319621
0.386407603	1.93%	0.01932	0.92251659
0.346827429	1.73%	0.017341	0.93985796
0.29833557	1.49%	0.014917	0.95477474
0.224551083	1.12%	0.011228	0.96600229
0.027918502	0.14%	0.001396	0.96739822

0.040769994	0.20%	0.002038	0.96943672
0.171462034	0.86%	0.008573	0.97800982
0.076612909	0.38%	0.003831	0.98184047
0.10700197	0.54%	0.00535	0.98719056
0.133977101	0.67%	0.006699	0.99388942
0.122211611	0.61%	0.006111	1

4.4. Eigenvector calculation results

Table 8. Eigenvector calculation results

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
X_1	-0.06	-0.08	-0.13	0.69	-0.16	-0.23	0.55	0.01	0.29	-0.01
X_2	0.13	0.03	0.12	-0.05	-0.43	0.73	0.30	-0.18	-0.03	-0.08
X_3	-0.26	-0.01	-0.13	0.06	0.04	-0.16	0.02	-0.58	-0.48	-0.09
X_4	-0.30	-0.05	-0.21	0.06	0.01	0.09	0.07	-0.07	0.04	-0.05
X_5	-0.27	-0.13	-0.09	-0.07	0.05	0.03	-0.17	-0.27	0.42	-0.32
X_6	-0.30	-0.01	-0.18	0.01	0.09	0.11	-0.13	0.14	0.24	-0.22
X_7	-0.31	-0.03	-0.17	0.02	0.07	0.09	-0.06	0.02	0.01	0.28
X_8	-0.29	0.09	-0.16	0.06	0.12	0.03	0.01	0.29	-0.16	0.07
X_9	-0.28	0.11	0.09	-0.21	-0.30	0.02	0.07	0.11	0.14	0.05
X_{10}	-0.27	-0.04	0.24	0.13	-0.09	0.13	-0.16	-0.05	-0.02	-0.35
X_{11}	-0.23	0.31	0.08	-0.03	0.03	-0.01	0.34	0.12	-0.14	0.19
X_{12}	-0.25	0.23	0.26	-0.06	0.08	0.03	0.22	-0.16	-0.19	0.03
X_{13}	-0.29	0.19	-0.06	-0.10	-0.06	0.05	-0.08	0.17	-0.03	0.30
X_{14}	-0.27	0.08	0.18	0.10	-0.24	-0.02	-0.22	-0.10	0.08	-0.02
X_{15}	-0.14	-0.21	0.52	0.16	-0.15	-0.14	-0.08	0.46	-0.21	-0.24
X_{16}	-0.07	-0.47	0.26	0.28	0.20	0.15	-0.11	-0.18	-0.22	0.31
X_{17}	-0.13	-0.49	0.05	-0.16	-0.17	0.04	0.00	-0.01	0.25	0.50
X_{18}	-0.02	0.17	0.49	-0.07	0.58	0.06	0.20	-0.15	0.39	0.05
X_{19}	-0.09	-0.37	-0.23	-0.16	0.37	0.31	0.33	0.28	-0.19	-0.27
X_{20}	-0.08	-0.29	0.09	-0.52	-0.18	-0.45	0.38	-0.10	0.00	-0.14

	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20
X_1	0.10	0.08	-0.08	-0.04	-0.03	-0.12	-0.03	0.03	-0.07	0.00
X_2	0.22	-0.03	0.20	-0.04	0.03	0.11	-0.01	0.05	-0.12	0.03
X_3	-0.04	0.03	0.26	-0.06	-0.02	-0.22	0.23	0.03	-0.32	-0.19
X_4	-0.18	-0.17	0.21	0.37	0.00	0.11	0.41	-0.05	0.41	0.50
X_5	0.01	-0.42	0.01	0.16	0.16	-0.06	-0.46	0.23	-0.14	-0.04
X_6	0.15	-0.12	0.08	-0.53	-0.45	0.26	0.27	-0.09	-0.01	-0.19
X_7	0.13	0.08	-0.23	-0.29	0.70	0.25	0.13	-0.01	-0.18	0.14
X_8	0.41	0.12	0.34	0.40	0.00	0.12	-0.30	-0.37	-0.07	-0.20
X_9	0.05	-0.11	-0.15	0.08	0.20	-0.47	0.27	-0.08	0.32	-0.49
X_{10}	-0.10	0.43	-0.49	0.35	-0.11	0.24	0.12	0.00	-0.15	-0.09
X_{11}	-0.49	-0.18	0.06	0.02	-0.04	0.41	-0.14	0.32	0.00	-0.30
X_{12}	0.03	-0.27	-0.38	-0.18	-0.18	-0.12	-0.26	-0.47	0.04	0.31
X_{13}	0.28	0.12	-0.09	0.06	-0.32	-0.30	-0.02	0.56	-0.17	0.31
X_{14}	-0.22	0.47	0.35	-0.32	0.07	-0.08	-0.35	-0.02	0.34	0.12

X_{15}	-0.01	-0.30	0.23	-0.07	0.16	-0.06	0.12	0.06	-0.24	0.16
X_{16}	0.25	-0.14	-0.09	-0.01	-0.09	0.11	-0.08	0.20	0.42	-0.20
X_{17}	-0.33	0.03	0.07	0.09	-0.22	-0.04	0.01	-0.28	-0.35	0.00
X_{18}	0.09	0.19	0.24	0.05	0.07	-0.08	0.21	0.02	-0.09	0.00
X_{19}	-0.23	0.21	-0.06	-0.13	0.05	-0.31	-0.14	0.07	0.05	0.01
X_{20}	0.30	0.16	0.01	-0.03	-0.01	0.29	-0.03	0.11	0.13	0.07

4.5. Final PCA result

Table 9. Final PCA result

No.	PC	Factor	Lable	Loading
1	PC3	X_{15}	Fatigue	7.87%
2	PC4	X_1	Age	6.02%
3	PC5	X_{18}	Swallowing Difficulty	5.70%
4	PC5	X_{19}	Clubbing of Finger Nails	5.70%
5	PC6	X_2	Gender	4.67%
6	PC7	X_{20}	Dry Cough	3.88%
7	PC9	X_5	Dust Allergy	2.52%
8	PC10	X_{17}	Shortness of Breath	1.93%
9	PC11	X_8	Chronic Lung Disease	1.73%
10	PC12	X_{10}	Obesity	1.49%
11	PC12	X_{14}	Coughing of Blood	1.49%
12	PC13	X_3	Air Pollution	1.12%
13	PC16	X_{11}	Smoking	0.86%
14	PC19	X_9	Balanced Diet	0.67%
15	PC19	X_{16}	Weight Loss	0.67%
16	PC20	X_4	Alcohol use	0.61%
17	PC20	X_{12}	Passive Smoker	0.61%
18	PC18	X_{13}	Chest Pain	0.54%
19	PC17	X_6	Occupational Hazards	0.38%
20	PC15	X_7	Genetic Risk	0.20%
	Total			54.08%

5. Conclusion

Based on the results of research conducted using the lung cancer dataset, early detection of lung cancer disease passes 16 of the most dominant factors out of a total of 20 factors, because only 16 factors have a large enough correlation to the formation of early detection of lung cancer disease with a proportion of covariance variance of 100%, including the 3 highest most dominant factors, namely the fatigue factor with a variance proportion value of 7.87%, the age factor with a variance proportion of 6.02%, and the swallowing difficulty factor with a variance proportion of 5.70%. The total variance obtained from the 9 variables screening variables is 54.08%.

References

Abdi, H. and Williams, L.J. (2010). Principal Component Analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2, 433-459. http://dx.doi.org/10.1002/wics.101

Johnson, W. A., & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis (6th Edition). Pearson Prentice Hall.

Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). Springer New York. https://doi.org/https://doi.org/10.1007/b98835

Nasution, M.Z, et al. 2018. PCA Based Feature Reduction to Improve the Accuracy of Decision Tree C4.5 Classification. International Conference on Computing and Applied Informatics 2017. DOI: 10.1088/1742-6596/978/1/012058.

Nilsson R, Tong J. Opinion on reconsideration of lung cancer risk from domestic radon exposure. Radiat Med Prot. 2020;1(1):48–54.

World Health Organization. Cancer. WHO. 2022.