Breast Cancer Prediction Using Principle Component Analysis(PCA)

Nishant Koradia
202118009
M.Sc.(DS), DA-IICT
Ahmedabad, India
niskoradia@gmail.com

Dhairya Lakhani
202118012
M.Sc.(DS), DA-IICT
Vadodara, India
lakhanidhairya.07@gmail.com

Kandarp Parmar
202118027
M.Sc.(DS), DA-IICT
Ahmedabad, India
kandarpparmar16@gmail.com

Vidhi Shah
202118037
M.Sc.(DS), DA-IICT
Gandhinagar, India
vidhushah19@gmail.com

1 Abstract

In the last few years, the most common and arising cancer in women is breast cancer. There are many reasons for a woman to get affected by this and it has been noted that women with age more than 50 are most likely to gain this disease. In this situation, if the woman gets to know about her disease in an early stage, then there are chances that it can improve the prognosis and the chance of survival significantly. Also, if the diagnosis is accurate then it can help the patients to reduce their unnecessary costs. Thus, for this, we are implementing the machine learning algorithm on the dataset and predicting that the cancer cell is malignant or benign.

2 Introduction

According to a recent study conducted in 2018, over 9.5 million people died as a result of cancer that year. According to WHO cancer is the second most impacting cause of mortality in the whole world. The situation in India is no better, with over 1300 people dying every day from various types of cancer, according to current numbers. The number of cancer forms and causes has steadily increased over the last decade, which is bad news for the world's population. Among different types of cancer Breast Cancer and Lung Cancer were the most common cancers in the whole world comprising 12.5% and 12.2% of the total cases which were registered in 2020. Taking the most common cancer Breast cancer into consideration, let us see what are the factors and the symptoms related. The factors include:-

1. Age: women with age (>50) are more likely to get breast cancer.
2. The individual's history of cancer: - If a woman has cancer in one breast, then there are high chances that the woman might have cancer in the other breast too.
3. Family History of breast cancer: - a woman is breast cancer-prone if in her family her mother, sister, daughter, or any female relative is having breast cancer.

The other factors include Child bearing and Menstrual History also which have less but a significant impact on a woman to have breast cancer.

In our dataset, we have 32 features with us from which retrieval of useful information is not possible. So, we have to do dimensionality reduction of features from which we can visualize things in a better way. For dimensionality reduction, we use Principal Component Analysis (PCA). But first, we have to see what is PCA?

3 Principle Component Analysis

PCA is a linear dimensionality reduction approach (algorithm) that converts a collection of correlated variables (p) into a smaller k number of uncorrelated variables called principal components while preserving as much variance as feasible in the original data. PCA is an unsupervised machine learning approach that finds relevant variables that may be used for subsequent regression, grouping, and classification tasks apropos of Machine Learning.

3.1 Why PCA?

For the small datasets, PCA doesn't work because it would give all the columns as the principal components, but if the dataset is very large then we use PCA because it removes noise by reducing a large number of features to just a couple of principal components which are easy to visualize and calculate. Principal components are nothing but the orthogonal projections of the data from a higher-dimensional onto lower-dimensional space.

3.2 Steps for PCA

1. Standardize the dataset.
2. Calculate the covariance matrix for the features in the dataset.
3. Calculate the eigenvalues and eigenvectors for the covariance matrix.
4. Sort eigenvalues and their corresponding eigenvectors.
5. Pick k eigenvalues and form a matrix of eigenvectors.
6. Transform the original matrix.

4 Mathematical Formulation

Lets us understand the steps with the help of mathematical proof: Suppose we have a data with set of observations x_n where $n = 1\,2\,...\,N$ and x_n is the variable which is in D dimensional space. Our aim is to project these data onto a space that has M dimension where $M < D$. For instance consider that we are given the value of M to be 1 (i.e M=1) then ,the direction of space of M can be defined using the D dimensional vector u_1 which we take as the unit vector.So we get,

$$u_1^T u_1 = 1 \tag{1}$$

Each data point x_n can be projected onto a scaler value $u_1^T x_n$. Therefore the mean of the projected data becomes ,$u_1^T \bar{x}$ where is the sample set mean which is given by ,

$$\bar{x} = \frac{1}{N} \sum_{1}^{N} x_n \tag{2}$$

And the variance of the projected data is given by,

$$\frac{1}{N} \sum_{1}^{N} u_1^T x_n - u_1^T \bar{x} = u_1^T S u_1 \tag{3}$$

Where S is the covariance matrix of the data which is defined by,

$$S = \frac{1}{N} \sum_{1}^{N} (X_n - \bar{x})(X_n - \bar{x})^T \tag{4}$$

After this we have to maximize the projected variance $u_1^T S u_1$ with respect to u_1. The equation has to be a maximization constraint otherwise $u_1 \to \infty$. The best and the appropriate constraint can be derived from the normalization condition $u_1^T u_1 = 1$. To implement this condition we introduce a Lagrange multiplier that we shall denote by λ_1 and make an unconstrained maximization of

$$u_1^T S u_1 + \lambda_1 (1 - u_1^T u_1) \tag{5}$$

Thus when we do derivative of the above we do derivative with respect to u_1 and equate it to 0, thus we get a stationary point when

$$S u_1 = \lambda_1 u_1 \tag{6}$$

Thus from the above equation we get that u_1 must be an eigenvector of S. Premultiplying above equation with u_1^T and making use of equation $u_1^T u_1 = 1$ we see that the variance is given by,

$$u_1^T S u_1 = \lambda_1 \tag{7}$$

Thus, the variance will be maximum when we will set u_1 equal to the eigenvector having the largest eigenvalue λ_1. The eigen vector which we get by this way is called as the First Principal Component. In the similar way for the Second Principal Component we would take the second highest value of the eigenvalues and find the eigenvector corresponding to it and so on.

5 Solution

We have seen that we have 30 columns in our dataset, and working with this big dataset and extracting useful information looks difficult. Thus now we are reducing our dataset to 6 principal components using the steps which will make our calculation easy and understanding better. But, why did we selected 6 component vectors , why not more than or less than that?

The answer to this question is Kaiser's rule for selecting principal components which tells us that if the eigenvalue is greater than 1 than it should be considered and if it is less than 1 than we can drop that column and not consider as a principal component. By this method we get the principal components which gives us the maximum variance explained.

In our dataset we got 88.76% variability with the six principal components and only 11.24% of variance is loss.

6 Results and Analysis of Solution

6.1 Model Implementation(Without PCA)

First of all, we applied all four machine learning models on raw dataset without applying PCA to understand the accuracy of the models and running time. After applying it, we got the following results:-

Model	Accuracy
k-nearest neighborsClassifier	0.929825
RandomForestClassifier	0.964912
Logistic Regression	0.964912
SVM(linear)	0.956140

Table 1: Without PCA

6.2 Scree Plot

From scree plot we get the elbow curve from which we get to note that the principal component after 6 is having eigenvalue less than 1. Thereby we are selecting 6 principal components.

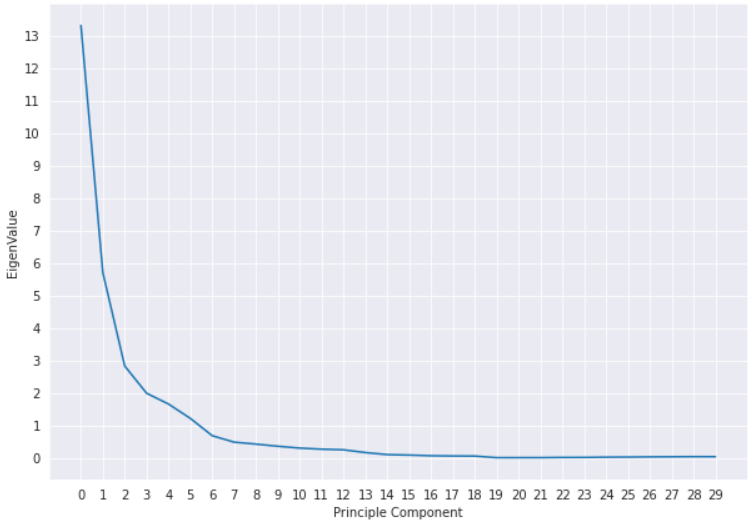

Figure 1: Scree Plot(PCA vs Eigenvalue)

6.3 Model Implementation(With PCA)

After applying PCA, we considered those components only as an attribute and applied again all four Machine Learning models which we applied for without PCA to understand the difference and how PCA is helping in improving accuracy. The results of that are as follows:-

Model	Accuracy
k-nearest neighborsClassifier	0.956140
RandomForestClassifier	0.956140
Logistic Regression	0.982456
SVM(linear)	0.982456

Table 2: With PCA

6.4 ROC Curves(Without and With PCA)

ROC curve,which is also known as "Receiver Operating Characteristics" Curve, is a metric used to measure the performance of a classifier model. The ROC curve shows us the rate of true positive(TPR) with respect to the rate of false positive(FPR), which means the sensitivity of the classifier model. It can be plotted with different thresholds settings. The TPR can be also considered as sensitivity, probability or recall of detection. The FPR

can be termed as probability of False Alarm and can be calculated by computing (1-Specificity). Thus, it can be also said that ROC Curve is recall or sensitivity as a function of fall-out.

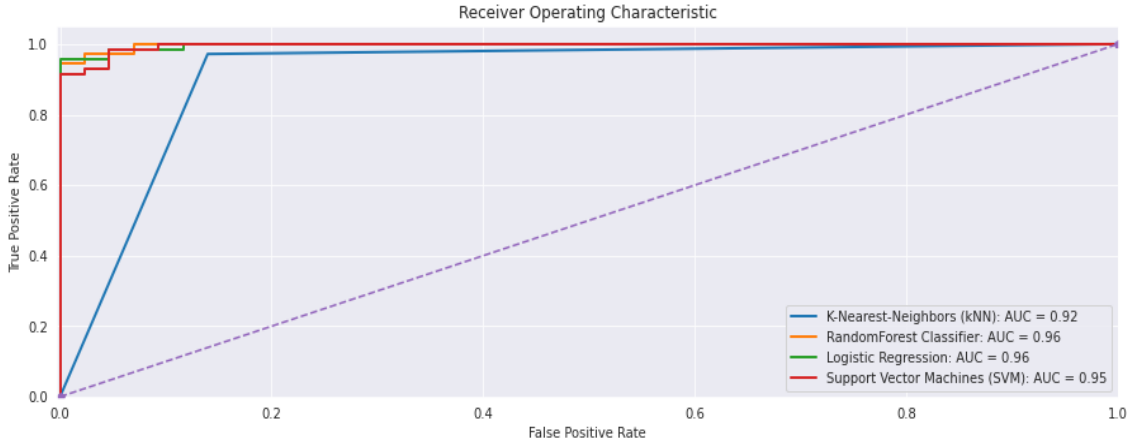

Figure 2: ROC Curve(Without PCA)

Figure 3: ROC Curve(PCA)

7 Conclusion

The primary purpose of this study is to create and execute a novel computation for predicting malignant and benign cancer due to which we are implementing models on our data and calculating accuracy of the particular model. In order to get more accuracy, we are implementing all the supervised learning models on our dataset before doing PCA and also after doing PCA. We saw that after performing component analysis we get more accuracy comparatively to the accuracy before PCA. And also it can be further implied that running time is also decreasing significantly after applying PCA as we are considering only six components instead of 30 features. Thus, we can say that model is very robust and works better after PCA and it can predict in better way which can further imply to work in a real world issue where we can predict an individual's cancer as malignant and benign which can save their lives.

References

[1] A. Mangal and V. Jain, "Prediction of Breast Cancer using Machine Learning Algorithms," 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2021, pp. 464-466, doi: 10.1109/I-SMAC52330.2021.9640813 .

[2] P. P. Sengar, M. J. Gaikwad and A. S. Nagdive, "Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020, pp. 796-801, doi: 10.1109/ICSSIT48917.2020.9214267.

[3] Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006.

[4] S. Ara, A. Das and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," 2021 International Conference on Artificial Intelligence (ICAI), 2021, pp. 97-101, doi: 10.1109/ICAI52203.2021.9445249.

[5] M. Shalini and S. Radhika, "Machine Learning techniques for Prediction from various Breast Cancer Datasets," 2020 Sixth International Conference on Bio Signals, Images, and Instrumentation (ICBSII), 2020, pp. 1-5, doi: 10.1109/ICBSII49132.2020.9167657.

[6] A. G´eron, Hands-On Machine Learning with Scikit-Learn & Tensor Flow, O'Reilly ISBN: 9781491962299.