International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 10 Oct. 2022, pp: 1189-1196 www.ijaem.net ISSN: 2395-5252

Breast Cancer Prediction using Principal Component Analysis with Logistic Regression

Mohammad Kaosain Akbar
Concordia Institute for Information Systems Engineering (CIISE)
Concordia University, Montreal, Quebec, Canada.

Date of Submission: 15-10-2022 Date of Acceptance: 31-10-2022

ABSTRACT: Breast Cancer is considered as the major andmost common form of cancer among women. Breast cancer issecond source of death causes by cancer, with first being thelungs cancer. To tackle this cancer, rigorous efforts areconstantly being given by scientists across the globe. Besides, thefield of Machine Learning and Data Mining has made asignificant progress over the years for extracting and gatheringvaluable information, even from the most complex data sources.Based on the information extracted from data, the MachineLearning model is also capable of performing certain degree ofprediction, classification, and clustering. In this paper, we haveexplored the relation between diagnosis of breast cancer withmultiple number of attributes of a dataset. We have used asupervised learning classification algorithm called LogisticRegression for predicting the existence of breast cancer, basedon five different attributes of X-ray images dataset. The datasetwas obtained from the University of Wisconsin Hospitals,Madison from Dr. William H. Wolberg. Prior to applying theMachine Learning algorithm, we performed a dimensionalityreduction technique over the data, known as PrincipalComponent Analysis (PCA). Then, we analysed theperformance of the Machine Learning algorithm based on theaccuracy, precision, recall, f1-score, and support. Additionally,we also analysed the above scores generated by the modelwithout the use of the PCA analysis.

KEYWORDS: Logistic Regression, Machine Learning, PrincipalComponent Analysis, Breast Cancer, Prediction.

I. INTRODUCTION

Cancer is a horrible disease which occurs when there is abnormal growth of cells within our body. These cells can spread across the body, destroying normal body cells and can form a mass commonly known as tumour [1]. Breast Cancer is developed from abnormal cells growth in breast and signs mostly include change in breast shape and skin and lump in the breast. Although this kind of cancer is invasive cancer in women, breast cancer can also occur in men [2]. Based on the situation of cancer, Doctors arrange treatment plans,but any misdiagnoses can undoubtedly lead patient to loss significant amount of curing time which in turn may end up in dire consequences. Therefore, it is very important to diagnose and predict breast cancer at early stage

Our study chooses Logistic Regression in order to work predict diagnosis of breast cancer because this algorithm is generally used to predict the probability of a target value. Often the target value is dichotomous, which means that the target value will have two outcomes. In our case that would be either the targeted patient is diagnosed with cancer, or the patient is not diagnosed with cancer. We used Pythonprogramming language as the essential tool for developing theMachine Learning model and predicting patients beingdiagnosed with breast cancer or not based on the datasetgenerated from X-ray images. The Breast Cancer datasetwhich we used to train and test out model was obtained fromreputed data repository platform called “Kaggle”.

We first performed PCA analysis on the breast cancerdataset. PCA analysis reduced the dimension of the data andthen the data was used to train the Machine Learning modelwhich then we calculated scores based on the predictionsbased by the model. Moreover, we also trained the MachineLearning model without reducing the data using PCAanalysis. Then we compared the two set of scores, generatedby two Logistic Regression

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 10 Oct. 2022, pp: 1189-1196 www.ijaem.net ISSN: 2395-5252

Models. results show dynamic behaviour under various operating and environmental conditions anddemonstrate advantages of adaptive control over the non-adaptive type.

II. DATASET DESCRIPTION

The Breast Cancer dataset was obtained from a well know online data repository named "Kaggle". This dataset is a collection of data from different patients who performed Xrays to check whether they were infected with Breast Cancer. The dataset has six attributes; five attributes represent different values obtained from the X-ray images such mean of radius, mean of the texture of the X-ray, etcetera. The final attribute indicates whether the patient is diagnosed with Breast Cancer or not. The provided information by the breast cancer datasets are as follows:

- mean_radius : Mean radius of the lump
- mean_texture: Mean texture of the X-ray image.
- mean_perimeter: Mean perimeter of the lump
- mean_area: Mean area of the lump
- mean_smoothness: Mean of smoothness of the image
- diagnosis: Whether patient is diagnosed with Cancer ornot.

0 represents patients not being infected with breastcancer.
1 represents patients who were diagnosed with breastcancer.

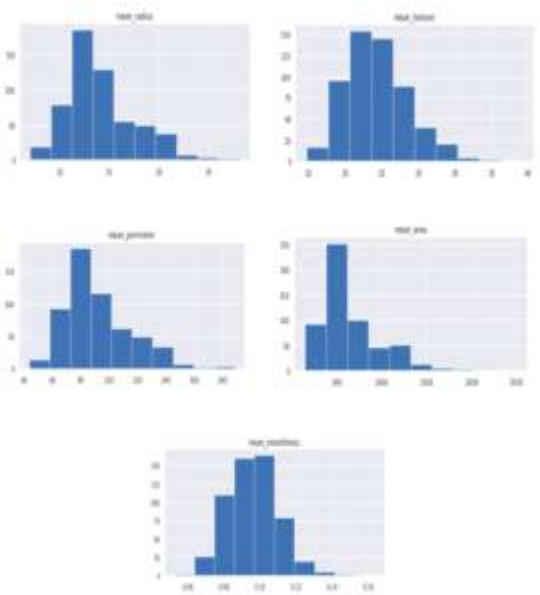

Figure 1: Histogram Representation of the attributes which is used for training the Machine Learning model.

Apart from the diagnosis attribute, we will use all the reamingattributes to train our machine learning models based onLogistic Regression. The data was gathered from theUniversity of Wisconsin Hospitals. The data was a publicdataset, and no additional charges were made whiledownloading the data.Figure 1 represents the histogram representation of theattributes. The figure shows the different values of theattributes and their frequency of occurrence within the breast

III. PRINCIPAL COMPONENT ANALYSIS

A. Theoretical Description

Principal Component Analysis is a statisticaltechniquewhich is used which is greatly used to reduce the dimensionsof the data. In addition to reduce dimensions, PCA alsoemphasize on variation and highlights strong patterns in adataset so data can be easily explored and visualized.Although, PCA reduces the dimension of the data, PCA triesto retain as much information possible from the original data.PCA generates Principal Components, and the first Principalcomponents represents the larger variance of the data whichmeans that this component accounts for most of thevariability in the data. Likewise, thesecond Principalcomponent represents the second most significant varianceand so on.In the follows, we describe the steps ofperforming thePCA analysis:

- First, we try to center the original data by computing thecentered data matrix Y = HX by subtracting off-columnmeans.
- Next, for the second step, using the centered data, wecompute the covariance matrix S, represented by p × p.

$$s = \frac{1}{n-1} Y'Y.$$

- Then, for the covariance matrix S, we compute theeigenvectors and eigenvalues using eigendecomposition

$$s = A\Lambda A' = \sum_{i=1}^{p} \lambda_j a_j' a_j.$$

Here:

- A is a p x p orthogonal matrix (i.e A'A = I) whosecolumns aj = (aj1; aj2; :::; ajp) are the eigenvectorsof S.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 10 Oct. 2022, pp: 1189-1196 www.ijaem.net ISSN: 2395-5252

- Λ = diag($\lambda 1$; $\lambda 2$; :::; λp) is a p x p diagonal matrixwhose elements are the eigenvalues of S arranged indecreasing order.

For the last step, we compute the transformed data matrixwhich is Z = Y A and this matrix is of size nxp.

$$Z = (z'_1, z'_2, ..., z'_i, ..., z'_p) = \begin{bmatrix} z_{11} & z_{12} & .. & .. & z_{1p} \\ z_{21} & z_{22} & .. & .. & z_{2p} \\ . & . & . & . & . \\ . & . & . & . & . \\ z_{n1} & z_{n2} & .. & .. & z_{np} \end{bmatrix}$$

It transformed data matrix contains the coordinatesof the original data in the new coordinate system definedby the Principal Components. The rows of Z matrixcorrespond to observations Zi = A' (xi − $\bar{x}$), while itscolumns correspond to PC scores [3].

B. Application of PCA in reduction of Breast Cancer datadimensions

The data was obtained from Kaggle and was created for"AI for Social Good: Women Coders'Bootcamp". The datado not have time parameters. It only consists of X-ray imagefeatures and output of patient being diagnosed with Cancer ornot

#	Column	Non-Null Count	Dtype
0	mean_radius	569 non-null	float64
1	mean_texture	569 non-null	float64
2	mean_perimeter	569 non-null	float64
3	mean_area	569 non-null	float64
4	mean_smoothness	569 non-null	float64

Figure 2: Datatype of the individual attributes along with the non-nullcount.

From figure 2, we can see that all attributes of BreastCancer dataset have datatype of float64 and the dataset do not have any null values.

1) Box Plot
Initially, we normalized the data so that values of numericcolumns of the dataset gets into same scale.

$$\frac{value - \mu}{\sigma}$$

Then, we constructed the side-by-side box plots of the fiveattributes of the dataset.

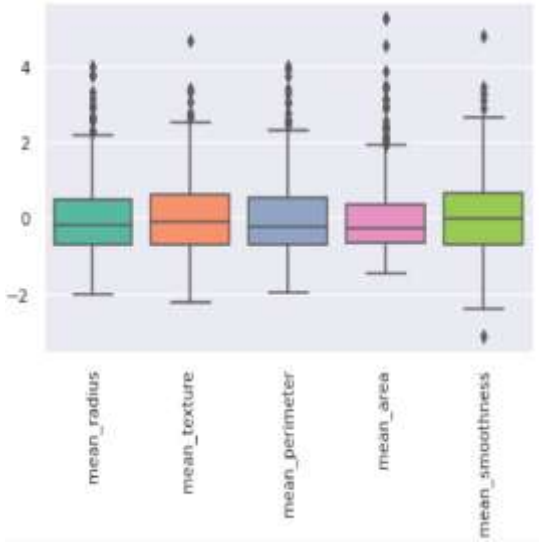

Figure 3: Box plot of the dataset

Figure 3 shows the box plot of the dataset. From the boxplot, we can see that all the attributes have certain numbers of outliers. These outliers occurred because health and medicaldata often contains abnormal patient condition, instrumentation errors or even recording errors [4].

2) Covariance Matrix
After the data is centred, we compute the CovarianceMatrix. Covariance Matrix helps us to measure of howvariables change with respect to each other. It is positivewhen variables tend to show similar behaviour and negativeotherwise.
From figure 4, we can see the covariance matrix of thedataset. It can be observed that the values at diagonals of thematrix are 1. since along the diagonal, variables are same withrespect to one another. It is also observed that all the valuesof the covariance matrix are positive, lying between 0 and 1.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 10 Oct. 2022, pp: 1189-1196 www.ijaem.net ISSN: 2395-5252

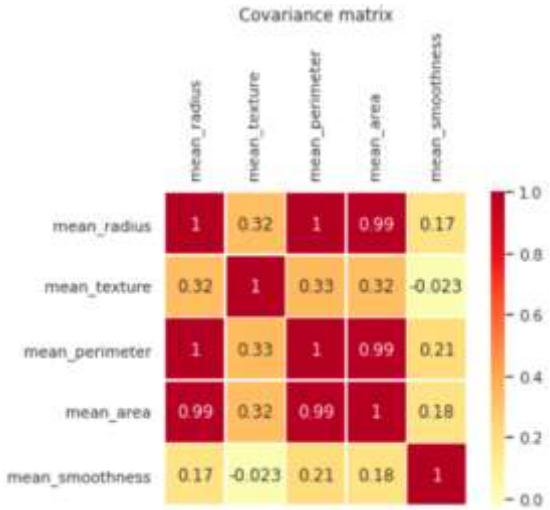

Figure 4: Covariance Matrix of the dataset

3) Pair-plot

Pair-plots usually plots a pairwise relationships in adataset [5]. Pair-plot represents the distribution of singlevariables as well as correlation between two variables. Fromfigure 5, we can see that certain attributes shows linearrelationships with other attributes of the dataset which meansthey are attributes are highlycorrelated. Additionally, fromthe diagonals of the Pair-plot we can see the histograms anddetermine the skewness.

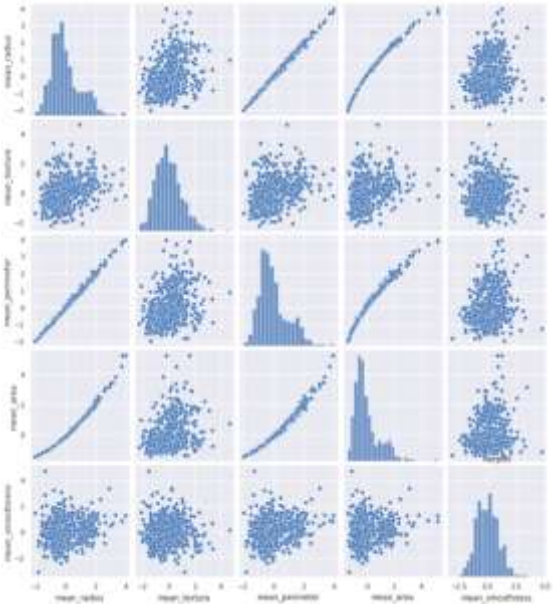

Figure 5: Pair-plot of the breast cancer dataset

4) Scree plot and Pareto Chart

Next, we applied Principal Component analysis on thedataset. In our Python programming language, we used thebuilt-in framework library called the scikit-learn framework.Then we obtain the Principal Components.

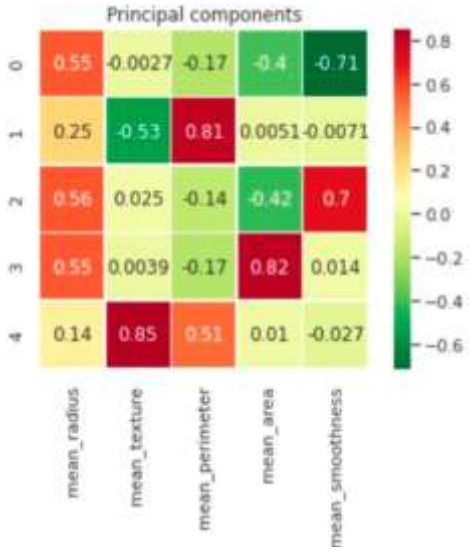

Figure 6: Principal Components generated after performing principal component analysis on the dataset.

In order to decide how many Principal Components shouldwe select; we first need to plot the Scree plot and Pareto Chart.The Scree plots provides us with explained variance ofindividual Principal Components that is which is thepercentage of variance counted for by the j^{th} PrincipalComponent.

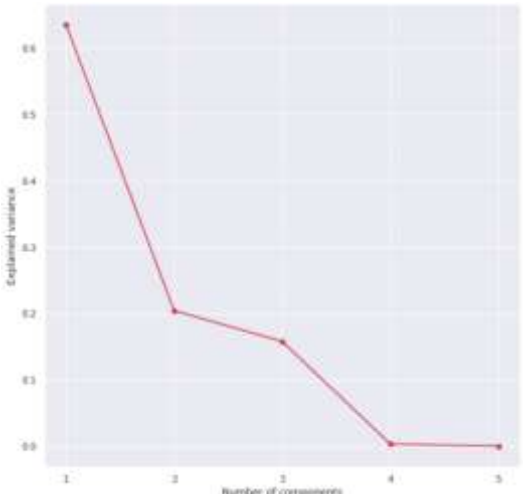

Figure 7: Scree Plot

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 10 Oct. 2022, pp: 1189-1196 www.ijaem.net ISSN: 2395-5252

Component. Explained variance is given by:

$$l_j = \frac{\lambda_j}{\sum_{j=1}^{p}} * 100\%, \forall j \in [\![1, p]\!]$$

From our scree plot, we can see that we can select the valueof "r" as 2, based on the elbowgenerated from the beginningof horizontal point 2.

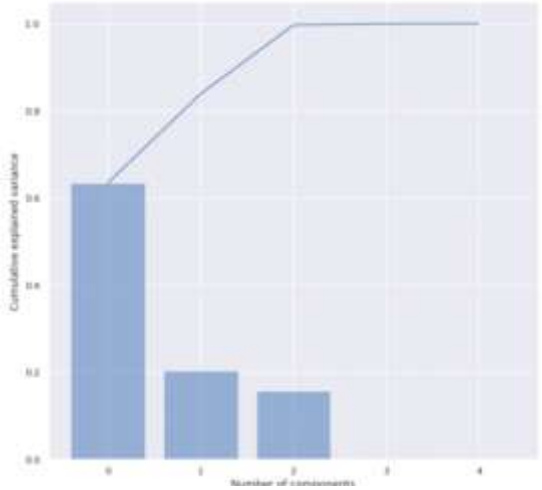

Figure 8: Pareto Chart

Generally, the rule is to be considered retain severalcomponents that represent at least 80% of the data variance.From our Pareto chart, we can see that almost 82% of the totalvariance can be explained by two of the initial PrincipalComponents that was generated right after the PrincipalComponent Analysis.

5) Scatter Plot

In Figure 9, we can see the scatter plot from which we canfind out which attributes have similar contribution as that ofthe computed PrincipalComponents.Considering the parameters mean_smoothness which is situated on the top left corner ofthe plot, we can deduce that the attribute of mean smoothnesshas high contribution on Principal Component 2 (A2).

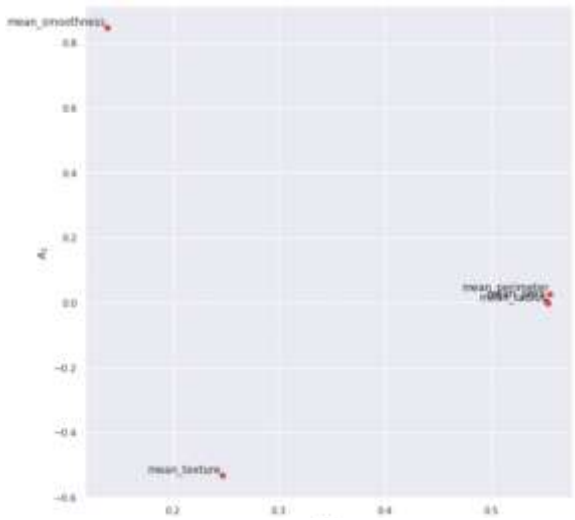

Figure 9: Scatter Plot

6) Biplot

From biplot, for each observation, we can see principalcomponent coefficient and principal component scorestogether. Points on the biplot represent observations whereelse the lines on the biplot are the attributes of the dataset.From the length and direction of the vectors, we can deducehow each attribute contribute towards the individual PCs.Considering mean_texture, this attribute contributesmore to PC2 than mean_smoothness attribute, becausemean_texture is closer to the PC2 than mean_smoothness.

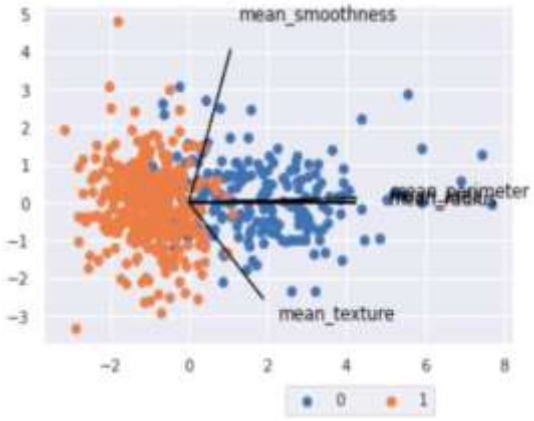

Figure 10: Biplot

IV. MACHINE LEARNING ALGORITHM: LOGESTIC REGRESSION

Dealing with Cancer detection data is a very sensitive issue as,we need to properly identify the cancer patient efficiently.Since our dataset deals with classification problem, we findthat for solving our problem, Logistic Regression wouldefinitely be an idle solution.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 10 Oct. 2022, pp: 1189-1196 www.ijaem.net ISSN: 2395-5252

Logistic regression is basically a supervised classificationalgorithm. In a classification problem, the target variable (oroutput), y, can take only discrete values for given set offeatures (or inputs), X.

Contrary to popular belief, logistic regression is aregression model. The model builds a regression model topredict the probability that a given data entry belongs to thecategory numbered as "1". Just like Linear regression assumesthat the data follows a linear function, Logistic regressionmodels the data using the sigmoid function [7].

$$g(z) = \frac{1}{1+e^{-z}}$$

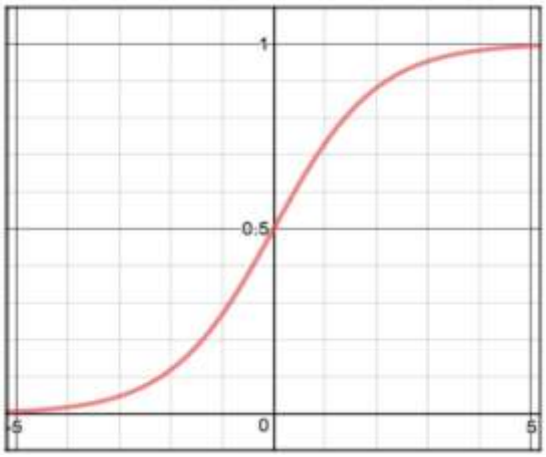

Figure 11: Graph of a sigmoid Function.

Logistic regression becomes a classificationtechniqueonly when a decision threshold is brought into the picture. Thesetting of the threshold value is a very important aspect ofLogistic regression and is dependent on the classificationproblem itself.

V. EXPERIMENT WITH MACHINE LEARNING ALGORITHM

A. Applying Logistic Regression on the original BreastCancer Datasets

Initially, the original dataset is provided to the LogisticRegression. We use then use sklearn framework to performLogistic Regression on the original dataset. We used train andtest split operation to divide the data as training and testingset. After training the model with training set, we test ourmodel using the testing set. Then we print the AccuracyScore, Classification report and Confusion matrix, once weare done testing our model.

	precision	recall	f1-score	support
0	0.91	0.86	0.89	36
1	0.94	0.96	0.95	78
accuracy			0.93	114
macro avg	0.92	0.91	0.92	114
weighted avg	0.93	0.93	0.93	114

Accuracy of ML Algorithm on the Full Data: 0.9298245614035088

Chart 1: Classification Report and accuracy score of the Machine Learningmodel on original dataset.

From the classification report, we can observe that our modelgenerated a satisfactory output. The model had 91% precisionin identifying patients who does note have breast cancer and94% precision in identifying patients who is suffering frombreast cancer. Both outcomes have an average f1-score andrecall score of more than 90%. The accuracy of the trainingmodel is 93%.

Figure 12: Confusion Matrix for the ML algorithm on the Original Dataset.

From the confusion matrix generated by the model trainedand tested using the original data, we can see that truenegative score is 31 and true positive score us 75. Both of thescores are significantly higher than false positive and falsenegative scores, indicating that model was almost efficientlygenerating the expected output.

B.Applying Logistic Regression on the data generatedafter Principal Component Analysis.

Next, we used the reduced data generated by the PCA to trainour Machine Learning model based on Logistic Regression.Again, we used sklearn framework [6] to perform logisticoperation but this time on the PCA generated dataset. Weagain used the similar train and test split operation to dividethe data as training and testing set. After training the modelwith training set, we test our model using the testing set.Finally, we printed the Accuracy Score, Classification reportand Confusion matrix, once we are done

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 10 Oct. 2022, pp: 1189-1196 www.ijaem.net ISSN: 2395-5252

testing our modelwith data generated after Principal Component Analysis.

	precision	recall	f1-score	support
0	0.91	0.95	0.93	41
1	0.97	0.95	0.96	73
accuracy			0.95	114
macro avg	0.94	0.95	0.94	114
weighted avg	0.95	0.95	0.95	114

Accuracy of PCA Data: 0.9473684210526315

Chart 2: Classification Report and accuracy score of the Machine Learningmodel on PCA generated dataset.

From the classification report, we can observe that our modelhad 91% precision in identifying patients who does not havebreast cancer and 97% precision in identifying patients whois suffering from breast cancer. Both outcomes have a recallscore 95% and f1-score of more than 90%. The accuracy ofthe training model is 95%.

From the confusion matrix generated by the model trainedand tested using the PCA generated data, we can see that truenegative score is 39 and true positive score us 69. Again, bothscores are significantly higher than falsepositive andfalse negative scores, indicating that model was almostefficiently generating the expected output.

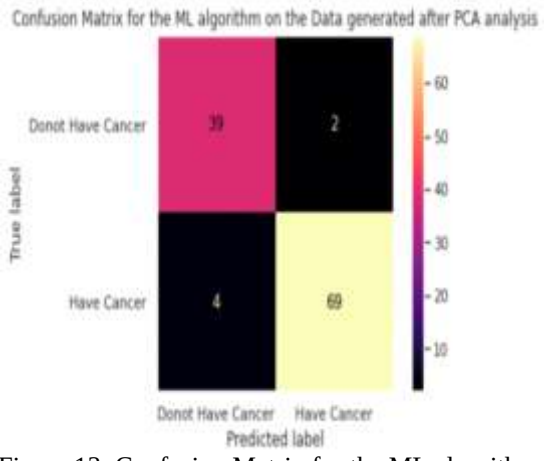

Figure 13: Confusion Matrix for the ML algorithm on the Dataset obtainedafter PCA.

VI. FINDINGS

Based on the classification report, accuracy scores andconfusion matrix, it can be said for that the classificationreport generated by the model which was trained by thedataset generated after PCA analysis shows better accuracyand precision score in comparison to the report generated bythe Machine Learning model trained by only the originaldataset. We can also look at the confusion matrices andconclude that the total number of false negative and falsepositive generated by the model trained with PCA generateddata is lower in contrast to that of the model trained by theoriginal dataset. Considering all these findings we can saythat applying PCA on our breast cancer datasets not onlyreduces the dimensions of our data but also create a machinelearning model from which we are also getting betteraccuracy score, higher precision, recall and f1-scores as wellas getting a better confusion matrix outcome. Although weget similar result from the model trained by the originaldataset, yet model trained by PCA datasets showed betteroutput for breast cancer detection data using LogisticRegression Algorithm.

VII. CONCLUSION

In this paper, we have demonstrated how to perform PCAanalysis on a dataset. Initially we performed PrincipalComponent Analysis on the breast Cancer Datasets, and weshowed that the attributes are correlated but act differently onthe principal components based on biplot. Then we appliedthe original data and data generated after

PCA analysistotrain two seperate ML model using Logistic Regression. Wecompared their classification reports, accuracy scores andconfusion matrix. We found that Machine Learning modeltrained with data generated by PCA had better accuracyprecision, f1 and recall score in comparison to model whichwas trained using original breast cancer datasets. In the futurewe hope to repeat these experiments with breast cancer datasetswith more attributes so that we can obtain the exact pictureabout the models trained by PCA generated datasets andoriginal datasets.

REFERENCES

[1]. Cancer Treatment Centers for America https://www.cancercenter.com/
[2]. MayoClinichttps://www.mayoclinic.org/diseases-conditions/breastcancer/symptoms-causes/syc-20352470
[3]. A. Ben Hamza, Chapter 5 : Statistical Process and Quality Control,INSE 6220 course content, Winter 2017..K. Elissa, "Title of paper ifknown," unpublished.
[4]. V. Deneshkumar, K. Senthamaraikannan, M. Manikandan,Identification of Outliers in Medical Diagnostic System Using DataMining Techniques, International

Journal of Statistics andApplications, Vol. 4 No. 6, 2014, pp. 241-248. doi:10.5923/j.statistics.20140406.01.

[5]. PythonTutorial https://pythonbasics.org/seabornpairplot

[6]. Sklearn: https://scikit-learn.org/

[7]. Sigmoid Function: https://en.wikipedia.org/wiki/Sigmoid_function