Principal component analysis and Machine learning classification algorithms for the diagnosis of breast cancer

Concordia Institute for Information System Engineering

Concordia university

Narina Hari Krishna Ayyappa

Student ID: 40186038

***Abstract*— Breast cancer is the most common cancer in women that is significantly increasing every year and it became the second most common cancer that is affecting the people all over the world. Therefore, to diagnose this cancer, analysing and gaining information from the previous data will be very helpful. There are various methods or techniques to analyse or predict the results from the previous data. Principal component analysis is a multivariate technique that is used to reduce the dimensionality of data for analysis. This technique is used in the current project to analyse data for the diagnosis of breast cancer. Different machine learning classification algorithms are also applied on the data to predict the need for diagnosis and decide the best algorithm based on the confusion matrix and accuracy scores.**

***Keywords*—Principal component analysis, classification algorithm, confusion matrix, accuracy score**

I. INTRODUCTION

There are different types of cancers that people are being affected with every year in the today's world. Among them, breast cancer has become one of the most common cancer especially in women and became the reason for the second most cancer deaths after lung cancer. On average, one in every eight women is suffering with breast cancer and about two-thirds of them having this cancer are 55 or older. The rest of the cases are aged between 35 and 54. The above statistics about the breast cancer gives a best example to explain the importance of data. Data is a collection of information that comes from observations, measurements, counts and responses. The statistics about breast cancer as mentioned above or to analyse or gain information in any other field, data plays a vital role. Manufacturing companies and various organisations use data to check the quality of a product or process and to determine the ways to improve it. It plays a major role to draw conclusions or take importance decisions. It is required to predict the problems in advance and to mitigate them. Data has become the key element in every organisation to improve the quality, predict the problems, make decisions and increase the revenue or profits. There are many techniques and tools that are being used to analyse the data. One of the most important technique or method is Principal Component Analysis (PCA). Principal Component Analysis is a multivariate technique that reduces the dimensionality of dataset by producing a new set of variables that is smaller than the original set of variables but contains most of the information of the original dataset. Hence it is also called data reduction technique. The new variables which are obtained from this technique are called the principal components and are not correlated. In this project the PCA technique is applied on the breast cancer data set to produce the new variables that contribute to more than 80% of the information of the original data. Breast cancer is curable and can be treated if spotted it early. Breast cancer can be predicted early by analysing the previous data containing attributes responsible for the cancer. This can be achieved by applying the binary classification of data using classification machine learning algorithms. The different classification algorithms are Logistic regression, Naive bayes, K-Nearest neighbours, SVM etc. In the current project, we have applied various Classification algorithms on the breast cancer dataset to predict the diagnosis of cancer and to find the algorithm that best fits the data.

II. DATASET DESCRIPTION

The dataset used in this project is taken from the Kaggle website. This dataset consists of the information related to breast cancer which was obtained from the Wisconsin Hospitals. Breast cancer is the most common type of cancer in women and the second highest in death rates. Therefore, predicting the cancer at the early stages is very important to diagnose it. Diagnosis of breast cancer is done when an abnormal bump is identified and to check further if it is cancerous and spread to other parts of body. The need for diagnosis can be analysed or predicted based on the features of the current dataset. The data consists of the below measurements based on which the data is classified into two types i.e., if diagnosis is required or not required.

Features:

1) mean_radius
2) mean_texture
3) mean_perimeter
4) mean_area
5) mean_smoothness

Class Variables:

1) Diagnosis(0/1)

The boxplots in the fig.1 shows the distribution of the data of different attributes that contribute to the analysis/prediction of the diagnosis for breast cancer. The box plots were constructed after normalising the data of all the attributes and could see the outliers for each attribute which indicate the abnormality and need for diagnosis.

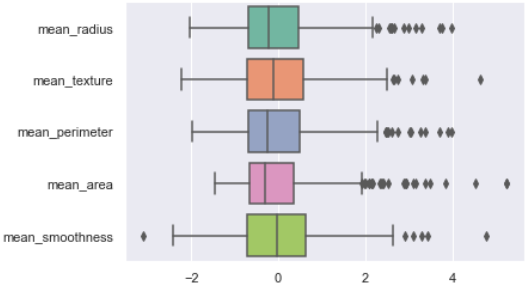


Fig. 1 Box Plot

III. PRINCIPAL COMPONENT ANALYSIS

A. *Theory*

Principal Component Analysis (PCA) is an exploratory multivariate technique which helps in simplifying the complex data sets. Hence this technique is called a dimensionality reduction technique or data reduction technique which helps in identifying the new set of variables smaller than the original set of variables and retains most of the information of the sample data. These new variables are called the principal components which are uncorrelated and are ordered by the fraction of total information each variable retains. Therefore, the idea of PCA is to minimize the number of variables of a dataset while preserving most of the information of the original dataset.

PCA is applied on the n-dimensional data which is organised as a matrix data where rows are the observations and columns are the variables. The PCA algorithm consists of four main steps to convert an original data matrix X into a transformed data matrix Z.

Step1: Centralization of the data matrix X

The data in the matrix is centred by subtracting each value with the mean of the respective column transforming the data matrix into centralized data matrix X.

Step2: Covariance matrix computation

Covariance matrix is a *p* x *p* symmetric matrix that explains the covariances associated with all possible pairs of the original variables. Where p is the number of variables.

The covariance matrix S of the centred data matrix is computed as follows

$$S = \frac{1}{n-1} X'X$$

Step3: Compute the eigen values and eigen vectors of the covariance matrix

An eigen vector is a nonzero vector that changes by a scalar factor when liner transformation is applied and the corresponding eigen value is the factor by which the eigen vector is scaled.

The eigen vectors and eigen values of the covariance matrix S is computed using the eigen-decomposition

$$S = A \wedge A' = \sum_{j=1}^{p} \lambda_j a_j a_j'$$

Where A = ($a_1, a_2, a_3, \ldots, a_p$) is a *p* x *p* orthogonal matrix whose columns are the eigen vectors of S.

Λ = diag($\lambda_1, \lambda_2, \lambda_3, \ldots\ldots, \lambda_p$) is a *p* x *p* diagonal matrix whose elements are the eigen values of S which are arranged in the decreasing order .

Step4: Compute the transformed data matrix Z

The transformed data matrix Z is computed as follows

$$Z = XA$$

Where X = centralized data matrix

A = *p* x *p* orthogonal matrix (eigen vectors matrix)

The size of matrix Z is *n* x *p*

$$Z = (z_1', z_2', \ldots, z_i', \ldots, z_p') = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1j} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2j} & \cdots & z_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ z_{i1} & z_{i2} & \cdots & z_{ij} & \cdots & z_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nj} & \cdots & z_{np} \end{pmatrix}$$

The rows of the transformed matrix Z represent the observations while the columns represent the principal component scores.

B. *Application of PCA on breast cancer dataset*

In the current project, Principal Component analysis is applied on the breast cancer dataset to analyse data and transform the original set of variables into new set of variables which are uncorrelated. The dataset is normalised before the application of PCA to get all the variables into the same scale. The PCA algorithm is applied on the scaled data and it consists of four main steps to transform the data matrix into a transformed data matrix Z. For PCA to work properly, the data is centralized by subtracting each value with their respective means of the dimensions.

Once the data is centralized, covariance matrix S is computed from the centred breast cancer data matrix X. The covariance matrix in fig. 2 shows the correlations between the variables that constitute to the breast cancer diagnosis. From the fig.2 we could see that the highest positive correlation exists between mean radius, mean perimeter and mean area. The positive correlation between these variables can also be observed from pair plot fig. 3 where the scatter plots between these variables has a linear shape. There is a negative correlation between the mean smoothness and mean texture and these variables are weakly correlated with all other variables. This can be clearly seen from pair plot as well where the scatter plots for these variables mean texture and mean smoothness are nonlinear.

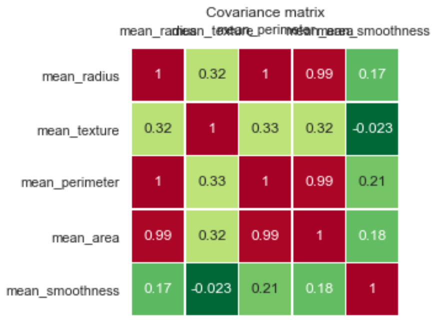


Fig. 2 Covariance matrix

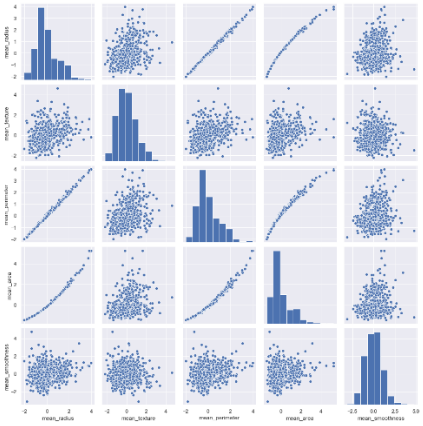


Fig. 3 Pair Plot

Now the Eigen values and Eigen vectors of the covariance matrix S from breast cancer data set are computed. The Eigen vectors are given by the following 5 x 5 matrix.

$$A = \begin{bmatrix} 0.553713 & -0.00269 & -0.17452 & -0.39872 & -0.7099 \\ 0.246878 & -0.53356 & 0.808878 & 0.005146 & -0.00715 \\ 0.55562 & 0.024738 & -0.14439 & -0.41808 & 0.70359 \\ 0.55179 & 0.003859 & -0.17094 & 0.816148 & 0.014001 \\ 0.138804 & 0.845385 & 0.514971 & 0.010485 & -0.02743 \end{bmatrix}$$

The Eigen values λ are given by

$$\lambda = \begin{bmatrix} 3.17234 \\ 1.02109 \\ 0.78858 \\ 0.016592 \\ 0.001395 \end{bmatrix}$$

The transformed data matrix Z is computed by using the Eigen vectors matrix A and centred data matrix X as follows.

$$Z = YA$$

Which means PCA transformed the five original centred variables of breast cancer dataset (X_1, X_2, X_3, X_4, X_5) into five new variables (Z_1, Z_2, Z_3, Z_4, Z_5) such that the Z variables are uncorrelated and called the principal components.

Since we know that the purpose of PCA is to reduce the dimensionality of data, the amount of variance accounted for by each principal component needs to be calculated. This can be achieved through eigen values where the percentage of variance accounted by the j^{th} principal component is given by

$$l_j = \frac{\lambda_j}{total\ sum\ of\ \lambda} \times 100$$

Where λ_j is eigen value of j^{th} PC

From the above calculations, the explained variance of each principal component is PC_1 =63.44%, PC_2=20.42%, PC_3=15.77%, PC_4=0.3%, PC_5=0.02%. We could see that the first two principal components combined account for 83.86% of the total variation in the breast cancer dataset. The explained variance of the principal components can also be observed from scree and pareto plots (fig. and fig.). Therefore, based on the explained variance by the first two principal components (PC_1 and PC_2) and from the scree and pareto plots, the dimensionality of the data can be reduced to two new variables to represent the breast cancer data set as more than 80% of the variance is contributed from the first two principal components.

Using the columns of the eigenvector matrix, the principal components PC_1 and PC_2 can be calculated as follows.

$$Z_1 = 0.55X_1 + 0.25X_2 + 0.56X_3 + 0.55X_4 + 0.14X_5 \quad (1)$$

$$Z_2 = -0.53X_2 + 0.02X_3 + 0.85X_5 \quad (2)$$

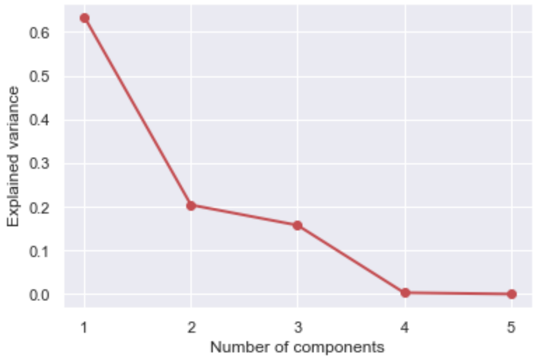


Fig. 4 Scree plot

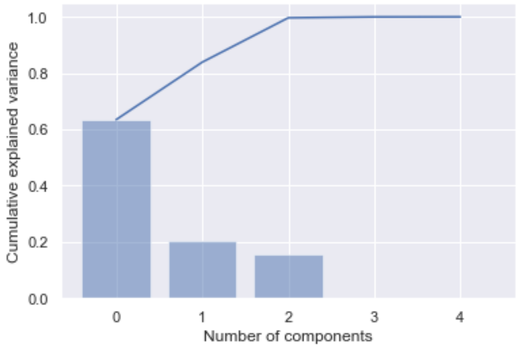


Fig. 5 Pareto plot

From the coefficients of the above equation (1), we could see that all the original variables of the breast cancer data set contributed to the PC_1 and among which, mean_radius, mean_perimeter and mean_area are equally accounted for PC_1. The equation (2) explains that PC_2 appears to be a contrast between mean_texture and mean_smoothness which can be evidenced from the positive coefficient of mean_smoothness and negative coefficient of mean_texture.

The scatter plot of PC2 coefficients vs. PC1 coefficients can be seen under fig. 6. This plot helps in understanding which variables have similar involvement within the PCs. From this scatter plot, we could see that the variables mean_radius, mean_perimeter and mean_area are contributed to PC_1 and variables mean_texture and mean_smoothness are contributed to PC_2 which is consistent with the values of the coefficients of PC_1 and PC_2 .

The 2D biplot of PC_2 vs. PC_1 is shown in the Fig. 7. Biplot is a graphical tool that provides information on both the variables and observations of a data matrix by displaying them graphically and this is the reason it is called biplot. The observations are displayed as points and the variables are displayed as vectors where the length and direction of the vector indicates how each original variable accounts to the two principal components. The axes in the biplot represent the principal components where the horizontal axes

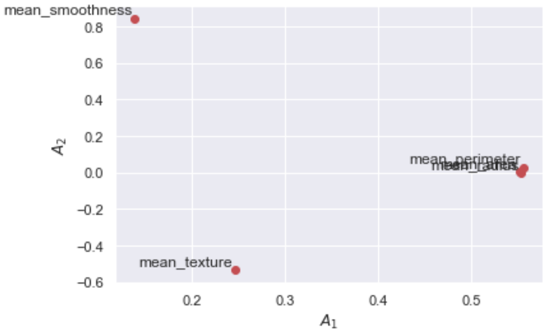


Fig. 6 scatter plot of PC_1 coefficients vs. PC_2 coefficients.

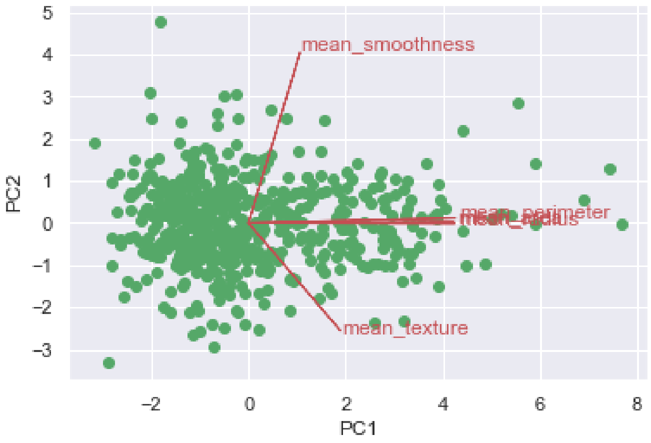


Fig. 7 Biplot

represent the PC_1 and the vertical axes represent the PC_2. From the 2D biplot Fig.7, we can see that the first principal component has positive coefficients for all the original variables of the breast cancer data matrix. This corresponds to all the vectors directed into the right half of the plot. In Biplot, a narrow angle between vector and principal components indicates that the variable plays a major contribution in the Principal component. Therefore, in the fig. 7, we could see that the angle between the vectors of variables mean_radius, mean_perimeter and mean_area and PC_1 is very narrow as these three variables play the highest contribution in PC_1.

The second principal component has two positive coefficients for variables mean_smoothness and mean_perimeter and one negative coefficient for variable mean texture. This corresponds to vectors with positive coefficients are directed into the top half and vector with negative coefficient is directed into bottom half of the plot. From the biplot, we could see a narrow angle between mean_smoothness and PC_2 as this variable provides the highest contribution in PC_2.

Each of the observations of the transformed matrix is represented by a point in the above biplot and their locations represent the score of each observation for the two principal components. For example, the points near the left edge in the Biplot indicate that those observations have the least or lowest scores for the first principal component. We can see from the biplot that the variable vectors are represented by rays extending out from the origin of the plot. Rays that tend to point in the same direction indicates that these variables are positively correlated. For example, the rays for mean_radius, mean_perimeter and mean_area point in the same direction which indicates the positive correlation of these variables with one another and the rays for mean_texture and mean_smoothness are pointed in the opposite direction representing a negative correlation with one another. Therefore, from the above principal component equations and biplot, we could see how the original variables are contributed to the first and second principal components.

IV. Classification Algorithms

Classification in Machine learning is a process of categorizing a particular set of data into classes and it can be performed on both structured and unstructured data. It predicts the class of a given data points where the classes are often referred to as target, labels or categories. The main goal of classification algorithms is to predict the class or category to which the new data falls into.

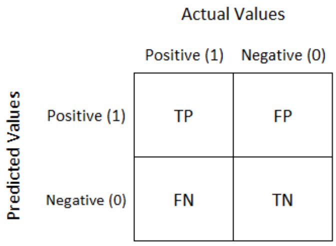
For the current project, detection of diagnosis for breast cancer can be identified as a classification problem and this is a binary classification as there can be only two classes i.e., diagnosis required or not required. In this case, the classifier needs the training data to understand how the variables of the dataset are related to the class. Once the classifier is trained, it can be used to predict whether diagnosis is required or not based on the input data of a particular patient.

There are many classification algorithms in machine learning. In the current project, we have used the logistic regression and Naive Bayes for classifying the breast cancer data set to predict if diagnosis is required or not.

Logistic regression and naïve bayes algorithm are applied on the Breast cancer original dataset, transformed dataset and the first two principal components to analyse and compare the results. The main goal is to assess the impact of original variables and first two principal components on the dependent variable i.e., to predict the binary variable if diagnosis is required or not.

Cross validation is an essential tool in data science to utilize the data better. Instead of using the traditional way of splitting the data into training and test sets, where training test is used to train the model and test set is used to predict and validate the data (For example, splitting the data into simple 80%(train) and 20%(test)), we have used the 5-fold cross validation where the data is split into k parts. For example, let us take the value of K=3 such that the observations of the data set were split into three parts and three different models were built where each model is trained on two parts and tested on the third (i.e., first model is trained on part 1 and part 2 and tested on part 3, second model is trained on part 1 and part 3 and tested on part 2 and so on). Therefore, logistic regression algorithm and naïve bayes are evaluated based on the 5-fold cross validation technique to estimate the skill of each model on our dataset.

The Performance of these classification algorithms when applied on the original dataset(X), transformed dataset(Z) and first two principal components (PC_1 and PC_2) is compared through accuracy scores and the accuracy scores are calculated based on the confusion matrix.



Predicted Values	Actual Values: Positive (1)	Actual Values: Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Fig. 8 Confusion matrix

Based on the breast cancer dataset:

True Positives (TP): When the actual value represents diagnosis is required and the predicted also indicates diagnosis is required.

False Positives (FP): When the actual value represents diagnosis is not required but the predicted indicates diagnosis is required.

True Negatives (TN): When the actual value represents diagnosis is not required and the predicted also indicates diagnosis is not required.

False Negatives (FN): when the actual value represents diagnosis is required but the predicted indicates diagnosis is not required.

Therefore, the accuracy is the ratio of number of observations that are predicted correctly to total number of observations.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

A. *Logistic Regression*

Logistic Regression of classification algorithm in machine learning take one or more independent variables to train and predict the outcomes. This is a binary classification algorithm, and the outcome will have only two possible outcomes. Logistic regression helps to find the best fitting relationship between the dependent variables and the independent variables and quantitatively explains the features or variables that lead to classification. It helps in understanding how a set of independent variables affect the result of the dependent variable. This classification algorithm is mostly used in identifying the diseases, weather predictions, voting applications etc.

The logistic regression equation is:

y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x))

This algorithm is applied on the breast cancer dataset by using python and the below table represents the accuracy scores when 5-fold cross validation technique is used on the original dataset(X), transformed dataset(Z) and first two principal components (PC_1 and PC_2).

K-fold	Accuracy (Logistic Regression)		
	Original dataset(X)	Transformed dataset(Z)	PC_1 and PC_2
K = 1	94.74%	90.35%	89.47%
K = 2	93.86%	95.61%	95.61%
K = 3	92.98%	91.23%	93.86%
K = 4	90.35%	93.86%	89.47%
K = 5	92.04%	91.15%	87.61%
Average	92.79%	92.44%	91.20%

Table. 1

From the table. 1, we could see that the average accuracy is almost same for original and transformed dataset since the new variables are not reduced in transformed dataset and hence 100% variance is retained from the original data. But we could see that the accuracy when first two principal components are used is comparatively less as from the PCA we observed that that the first two principal components combined account for only 83.86% of the total variation in the breast cancer dataset. Therefore, we could identify that logistic regression accuracy depends on the variance accounted by the variables.

B. *Naïve Bayes*

Naïve Bayes is a classification algorithm which is based on Bayes theorem and gives an assumption of independence among the predictors. In simple words, a naïve bayes algorithm assumes that the presence of a particular attribute or feature in a class is not related to the presence of other features. Although the features depend on each other, all of these attributes or features contribute to the probability of an occurrence independently. This algorithm or model is mainly useful for comparatively large datasets. Naïve bayes model is easy to make but it still performs better than most of the other classification algorithms in machine learning. The major use of this model is prediction of diseases, spam filters, sentiment analysis etc.

The Naïve bayes algorithm is based on the below Bayes theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

This model is applied on the breast cancer dataset by using python and the below table represents the accuracy scores when 5-fold cross validation technique is used on the original dataset(X), transformed dataset(Z) and first two principal components (PC_1 and PC_2).

K-fold	Accuracy (Naïve Bayes)		
	Original dataset(X)	Transformed dataset(Z)	PC_1 and PC_2
K = 1	88.60%	87.72%	92.11%
K = 2	88.60%	85.09%	95.61%
K = 3	94.74%	90.35%	88.60%
K = 4	93.86%	92.98%	87.72%
K = 5	88.50%	92.92%	92.04%
Average	90.86%	89.81%	91.22%

Table. 2

From the table. 2, we could see that the average accuracy of naïve bayes is comparatively less than the logistic regression accuracy for original and transformed datasets, which indicates that the logistic regression classification algorithm best fits these two datasets to predict the diagnosis for breast cancer. The accuracy scores when applied on first two principal components is same for both the classification algorithms in the table.1 and table. 2 which indicates that both logistic regression and naïve bayes best predicts the output when applied on first two principal components.

V. Conclusion

In this project, principal component analysis is applied on the breast cancer dataset to analyse the data and transform it into a new set of uncorrelated variables where it is observed that 83.86% of the total variance was accounted by the first two principal components. Therefore, the dimensionality of the data is reduced to two variables and analysed how the original variables contributed to these principal components. Once the relation between the original and new variables analysed, we have used these datasets(original(X) and transformed(Z)) and the first two principal components for the classification of breast cancer data i.e., to predict the need for diagnosis of breast cancer by using the classification algorithms Logistic regression and Naïve bayes. The accuracy scores are calculated for each model applied on these three datasets where it is observed that logistic regression best fits the original and transformed datasets and for the first two principal components, both the algorithms best fit the data as both had the same average accuracy scores. Therefore, based on the PCA and classification algorithms we have analysed the breast cancer dataset and identified the features with high variability and the algorithms that best predict the need for diagnosis of breast cancer.

REFERENCES

[1] A. Ben Hamza, "Advanced Statistical Approaches to Quality", unpublished.

[2] Principal Component Analysis for Breast Cancer Data with R and Python by Rukshan Manorathna Jan 09,2021

[3] https://www.webmd.com/breast-cancer/understanding-breast-cancer-basics

[4] https://www.c-q-l.org/resources/guides/12-reasons-why-data-is-important/#:~:text=Good%20data%20allows%20organizations%20to,benchmarks%20and%20set%20performance%20goals.

[5] https://www.kaggle.com/merishnasuwal/breast-cancer-prediction-dataset - dataset

[6] https://medium.com/analytics-vidhya/understanding-principle-component-analysis-pca-step-by-step-e7a4bb4031d9

[7] https://builtin.co m/data-science/step-step-explanation-principal-component-analysis

[8] https://www.edureka.co/blog/classification-in-machine-learning/#:~:text=MNIST%20Digit%20Classification-,What%20is%20Classification%20In%20Machine%20Learning,as%20target%2C%20label%20or%20categories.