Breast Cancer Prediction

Pavan Gandhi (202118008)
MSc Data Science
DAIICT
Gandhinagar, Gujarat, India
pavangandhi3020@gmail.com

Pranay Kothari (2021180010)
MSc Data Science
DAIICT
Gandhinagar, Gujarat, India
pranaykothari2000@gmail.com

Nidhi Sadhwani (202118043)
MSc Data Science
DAIICT
Gandhinagar, Gujarat, India
nidhi.sadhwani03@gmail.com

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.

1 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:-

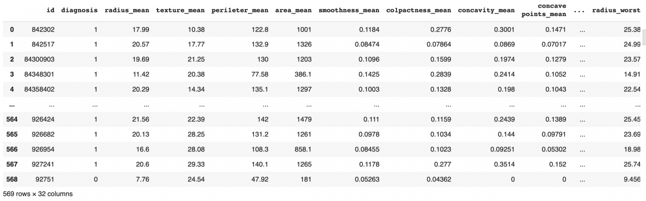
1. Age: women with age (>50) are more likely to get breast cancer.
2. The individual's history of cancer: - If a woman has cancer in one breast, then there are high chances that the woman might have cancer in the other breast too.
3. Family History of breast cancer: - a woman is breast cancer-prone if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

2 Dataset Description

The dataset used in this paper for applying the random forest algorithm is the University of California, Irvine Breast Cancer Wisconsin (Diagnostic) Dataset. This dataset is recorded for 569 medical records and belongs to multidimensional data. The dataset records the confirmed diagnosis of breast cancer cases from 10 dimensions. Include radius (mean of distances from center to points on the perimeter), texture (stan- dard deviation of gray-scale values) perimeter area, smoothness (local variation in radius lengths), compactness (perimeterZ / area - 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, fractal dimension ("coastline approximation" - 1), a total of 10 attributes, each of which is divided into average values, standards The difference and the worst value are 3 attributes, so there are 30 attributes for recording the detection result. The diagnosis results are classified by malignant = 1, benign = 0.

Dataset Characteristics	Multivariate	**Number of Instances**	569
Attribute Characteristics	Real	**Number of Attributes**	32



	id	diagnosis	radius_mean	texture_mean	perileter_mean	area_mean	smoothness_mean	colpactness_mean	concavity_mean	concave points_mean	...	radius_worst
0	842302	1	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	...	25.38
1	842517	1	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	...	24.99
2	84300903	1	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	...	23.57
3	84348301	1	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	...	14.91
4	84358402	1	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	...	22.54
...	...	...	...	...	...	...	...	...	...	...	...	...
564	926424	1	21.56	22.39	142	1479	0.111	0.1159	0.2439	0.1389	...	25.45
565	926682	1	20.13	28.25	131.2	1261	0.0978	0.1034	0.144	0.09791	...	23.69
566	926954	1	16.6	28.08	108.3	858.1	0.08455	0.1023	0.09251	0.05302	...	18.98
567	927241	1	20.6	29.33	140.1	1265	0.1178	0.277	0.3514	0.152	...	25.74
568	92751	0	7.76	24.54	47.92	181	0.05263	0.04362	0	0	...	9.456

569 rows × 32 columns

Figure 1: Data Description

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.

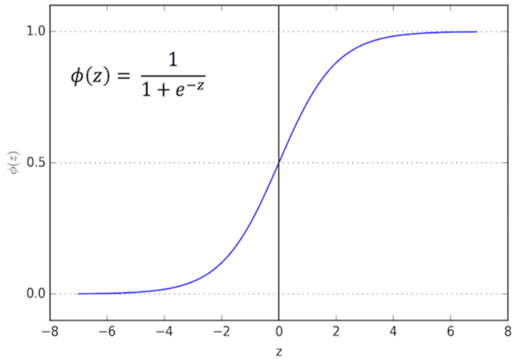
3.3 How does PCA help in our model

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.

4 Classification Models

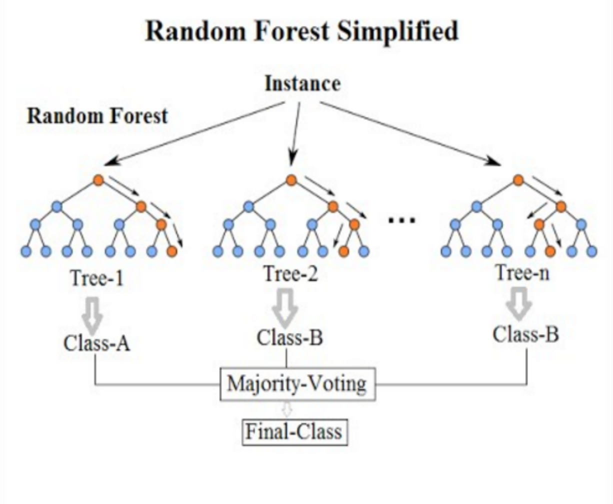
4.1 Logistic Regression

Logistic Regression is one of the most commonly used classification algorithms. Logistic Regression is used when the dependent variable is categorial. Sigmoid function is applied on the predicted values to get the probability. As probability ranges from 0 to 1, our result from this regression will range between 0 to 1. According to the result, we classify to either of the class.



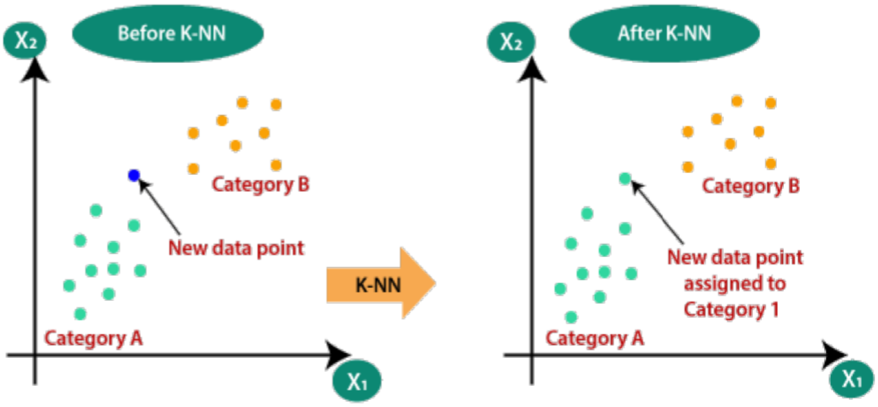
4.2 Random Forest Classifier

Random Forest Classifier is a kind of ensemble classifier. It uses decision tree algorithm in randomized way. The whole dataset is converted to boot strap dataset. Based on the total number of features, we randomly make multiple decision trees and output is generated. If majority of the output is 0 then 0 is the output or vice versa.



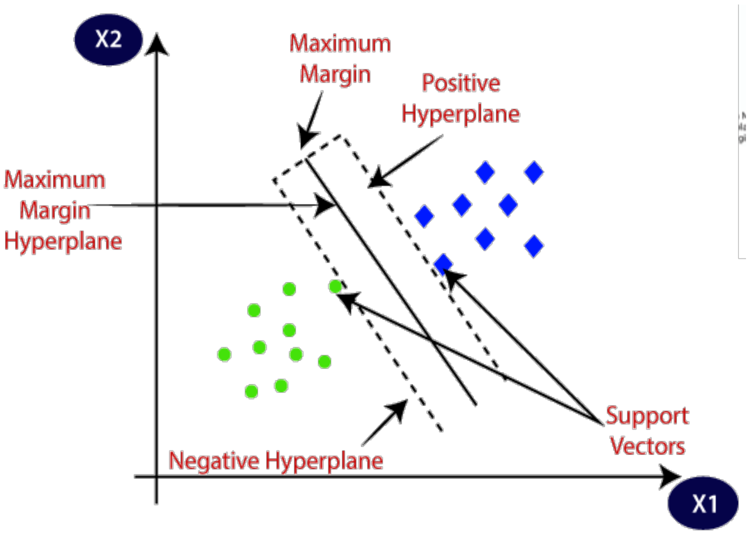
4.3 KNN Classifier

KNN represents K nearest neighbor Classification. Based on Euclidean Distance, K number of neighbor predicted values are taken to consideration. In our project we have taken 5 nearest neighbors into consideration. Based on the majority of the output of those 5 values, we declare the result.



4.4 SVM(Support vector machine)

Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.



5 Evaluation Strategy

The evaluation of classifiers is done using several evaluation matrices depends on the confusion matrix. Among of those criteria are Accuracy, precision, recall and f-score. They are calculated according to the following equations:

Accuracy is to find whether the given model is best fitted or not for finding true predictions, patterns and relationships between variables.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision is the quality of positive predictions made by the model. Exact positive refers to the number of true positves divided by the total number of predictions.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall tells us about the percentage of classes that we are focused on which were reproduced by the model.

$$\text{Recall} = \frac{TP}{TP+FN}$$

F1 score is represented by the weighted average of precision and recall, where 1 indicates the best and 0 indicates worst score.

$$\text{F-Score} = \frac{2*Precision*Recall}{Precision+Recall}$$

Where, TP represents the number of true positive
TN represents the number of true negatives
FP represents the number of false positives
FN represents the number of false negatives classes.

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.

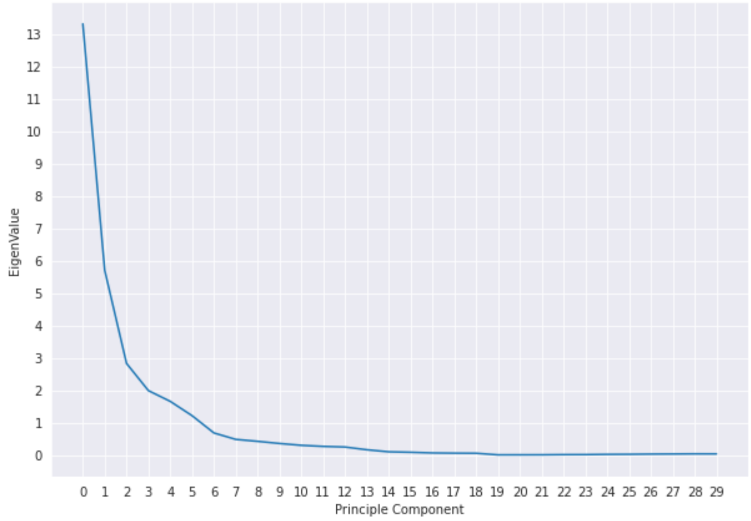


Figure 2: Scree Plot(PCA vs Eigenvalue)

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.

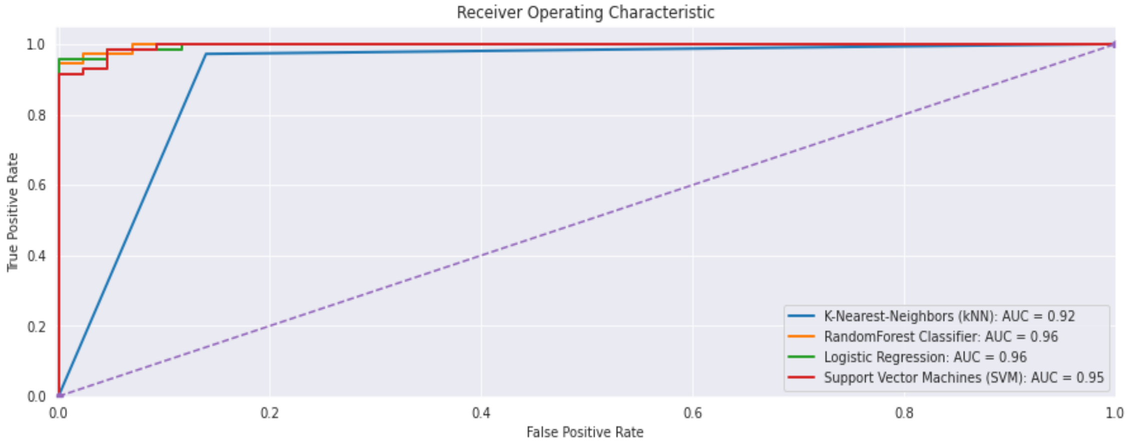


Figure 3: ROC Curve(Without PCA)

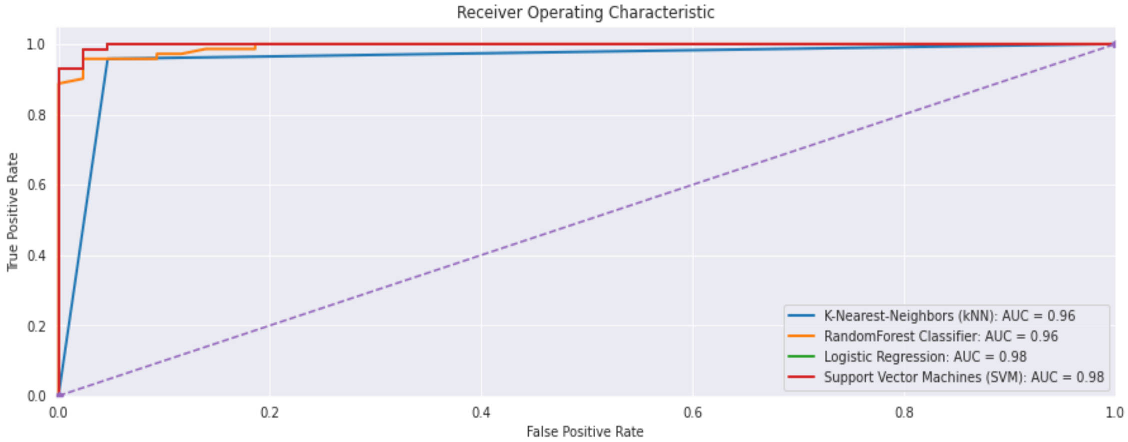


Figure 4: ROC Curve(PCA)

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.