INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 4 - V7I4-1539)
Available online at: https://www.ijariit.com

Breast Cancer Prediction using Machine Learning model

Sanath Kumar A.
sanathkumaratikukke@gmail.com
N. M. A. M. Institute of Technology, Nitte, Karkala, Karnataka

ABSTRACT

Day by day the number of cancer cases around the world are increasing and it has become the common health issue. One of the most commonly seen cancer is Breast cancer. It has become one among the leading causes of death. in women Breast cancer can be mainly seen in women, but it can be seen in men too. There is a belief that men cannot get breast cancer, but it is not true. Cancer can be treated only if it is detected in the initial stages, if not then it will spread to various parts of the body and its effect is irreversible and it will cause a threat to life. But most of the time it is difficult to detect the cancer and also sometimes people ignore the symptoms. In this matter artificial intelligence can be of a great help. Here we have collected a set and then we have built a prediction model to detect stroke based on the different algorithms that are available on machine learning.

Keywords: *Breast Cancer, Machine Learning Model*

1. INTRODUCTION

Breast Cancer is the most common type of cancer that is found in women around the world where most of the cases are diagnosed in late stages. It is a type of tumor that occurs in the tissues of the breast and this tumor is caused by the changes that occur in the DNA cells. Both men and women are affected by breast cancer. It is seen most in women and rarely seen in men. Today there are various means to detect and treat these cancer cells. Doctors say that this breast cancer occurs when some of the breast cells start to grow abnormally. It is also found that these cells have the capacity to divide rapidly than the healthy cells and finally it accumulates as lump. It may begin with the glandular tissue called lobules or in the other tissues or cells. Some of the common symptoms of breast cancer are fatigue, lump or area of thickening that can be felt under the skin, changes in the colour of the skin, difficulty swallowing, unexplained joint or muscle pains, unexplained bruising or bleeding etc.

Different forms of breast cancer exist, which arises when cancerous cells and tissues spread throughout the body. DCIS is a type of breast cancer that is non-invasive or pre-invasive. This indicates that the cells that line the ducts have transformed into cancer cells, but they have not moved beyond the duct walls into the surrounding breast tissue. Invasive Breast cancer is the second type of breast cancer and it is the most common type. This type of cancer is defined as the as the breast cancer that has spread to the surrounding tissue. Invasive breast cancer is the most common type of breast cancer, but there are several forms of invasive breast cancer. Invasive ductal carcinoma(IDC) and invasive lobular carcinoma are the two types most commonly encountered. Invasive ductal carcinoma occurs when a cancer that started in the milk ducts of the breast has migrated beyond the milk duct's lining and into the surrounding breast tissue. Invasive lobular carcinoma(ILC) occurs when a cancer that started in the milk-producing lobules of the breast has spread through the lobule's lining and into the surrounding breast tissue. Third type of breast cancer is Inflammatory breast cancer(IBC) which affects the blood vessels in the epidermis and/or the lymphatic vessels of the breast. It is an uncommon and aggressive form of invasive breast cancer. Triple-negative breast cancer(TNBC) is the fourth type of breast cancer which is distinct from other forms of invasive breast cancer in that it grows and spreads more quickly, has fewer treatment options, and has a poor prognosis.

Breast cancer most usually spreads to adjacent lymph nodes, but it can also spread to other parts of the body, including the bones, lungs, liver, and brain. Metastatic breast cancer, often known as stage IV breast cancer, is the most advanced form of the disease.

Machine learning can help us with this regard. If can find a way to detect the cancer on time/in the initial stages then we will be able to save a lot of lives by telling them about the various treatment that are available in the medical field.

2. METHODOLOGY

The breast cancer data is obtained from kaggle. We have a sample breast cancer dataset comprising of 569 rows and 32 columns. This analysis aims to observe which features are most helpful in predicting Malignant or Benign cancer to see general trends that may aid us in model selection and hyper parameter selection. To achieve this we have used machine learning classification methods to fit a function that can predict the discrete class of new input. The flow of data is shown in fig.1.

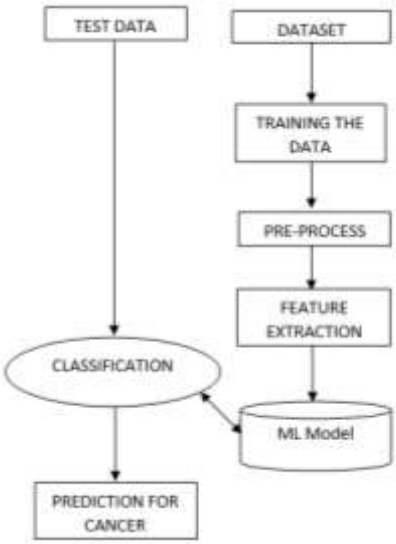

Fig. 1: Flow Chart

1. ID number 2) Diagnosis (M = malignant, B = benign) 3–32) Ten real-valued features are computed for each cell nucleus:
1. radius (mean of distances from center to points on the perimeter)
2. texture (standard deviation of gray-scale values)
3. perimeter
4. area
5. smoothness (local variation in radius lengths)
6. compactness (perimeter² / area — 1.0)
7. concavity (severity of concave portions of the contour)
8. concave points (number of concave portions of the contour)
9. symmetry
10. fractal dimension (“coastline approximation” — 1)

We have divided the dataset into four phases.The first phase is the data exploration. In this we will import the necessary libraries. We observe that the dataset contains 569 rows and 32 columns. ‘Diagnosis’ is the column in which we are going to predict cancer is Malignant (M) or Benign (B). ‘1’ means the cancer is Malignant and ‘0’ means Benign.

Diagnosis
0 357
1 212
dtype: int64

Table 1. Attribute Description

Attributes	Null
id	0
diagnosis	0
radius_mean	0
texture_mean	0
perimeter_mean	0
area_mean	0
smoothness_mean	0
compactness_mean	0
concavity_mean	0
concave points_mean	0
symmetry_mean	0
fractal_dimension_mean	0
radius_se	0
texture_se	0
perimeter_se	0
area_se	0
smoothness_se	0
compactness_se	0
concavity_se	0
concave points_se	0
symmetry_se	0
fractal_dimension_se	0
radius_worst	0
texture_worst	0
perimeter_worst	0
area_worst	0
smoothness_worst	0
compactness_worst	0
concavity_worst	0
concave points_worst	0
symmetry_worst	0
fractal_dimension_worst	0
Unnamed: 32 569	
dtype: int64	

The second phase is categorical data that the variables contain label values rather than numeric values. The numbers of possible values is often limited to fixed set. The data we use is usually split into training data and test data. The training dataset contains unknown outputs and the model learns on this data.

The third phase is feature scaling. In this most of the times your dataset will contain features highly varying in magnitudes, units and range.

Density Graph:

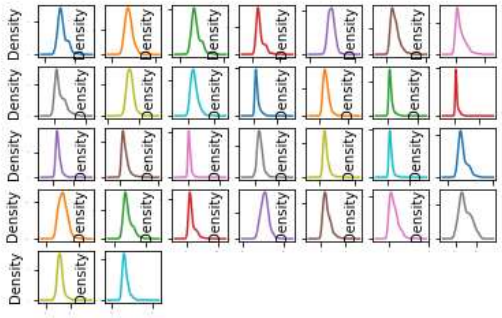

Fig 1: Density Graph

Colorbar Graph:

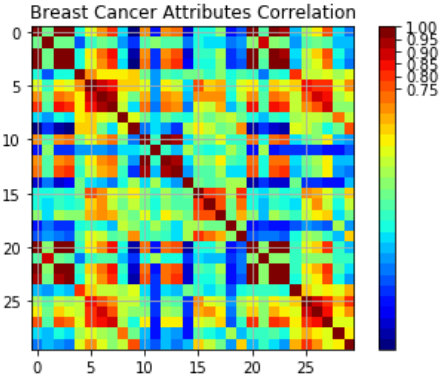

Fig 2: Colorbar Graph

Baseline algorithm checking:

From the dataset, we will analysis and build a model to predict if a given set of symptoms lead to breast cancer. This is a binary classification problem, and a few algorithms are appropriate for use.

CART: 0.918744 (0.041929) (run time: 0.422605)
SVM: 0.619614 (0.082882) (run time: 0.488190)
NB: 0.940773 (0.033921) (run time: 0.040002)

KNN: 0.927729 (0.055250) (run time: 0.190134)

Boxplot Graph:

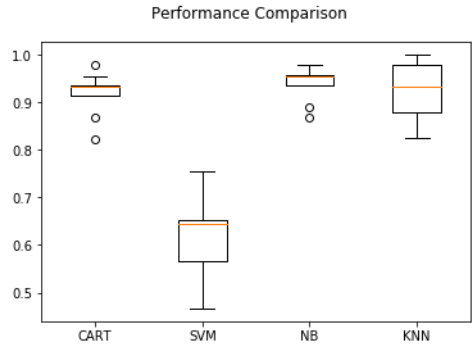

Fig 3: Boxplot Graph

ScaledCART: 0.912126 (0.038253) (run time: 0.195994)
ScaledSVM: 0.964879 (0.038621) (run time: 0.124014)
ScaledNB: 0.931932 (0.038625) (run time: 0.079986)
ScaledKNN: 0.958357 (0.038595) (run time: 0.103998)

Evaluation of algorithm on Standardised Data:
The performance of the few machine learning algorithm could be improved if a standardized dataset is being used. The improvement is likely for all the models. I will use pipelines that standardize the data and build the model for each fold in the cross-validation test harness.
ScaledCART: 0.912126 (0.038253) (run time: 0.195994)
ScaledSVM: 0.964879 (0.038621) (run time: 0.124014)
ScaledNB: 0.931932 (0.038625) (run time: 0.079986)
ScaledKNN: 0.958357 (0.038595) (run time: 0.103998)

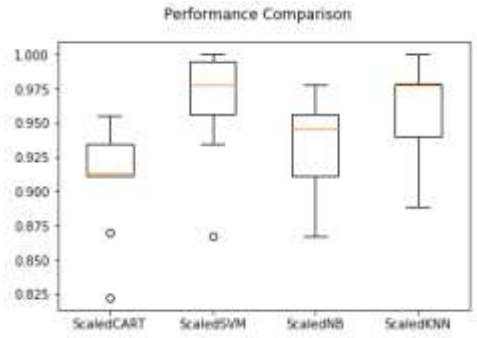

Fig 4: Boxplot Graph

The last phase is model selection. This is the most exciting phase in applying machine learning to any dataset. It is also known as algorithm selection for predicting the best results. The algorithm used is SVM.

SVM:
It is a classification method. In this algorithm, we plot each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinate.

Algorithm Tuning - Tuning SVM:
Best: 0.969231 using {'C': 2.0, 'kernel': 'rbf'}
0.964835 (0.026196) with: {'C': 0.1, 'kernel': 'linear'}
0.826374 (0.058723) with: {'C': 0.1, 'kernel': 'poly'}
0.940659 (0.038201) with: {'C': 0.1, 'kernel': 'rbf'}
0.949451 (0.032769) with: {'C': 0.1, 'kernel': 'sigmoid'}
0.962637 (0.029474) with: {'C': 0.3, 'kernel': 'linear'}
0.868132 (0.051148) with: {'C': 0.3, 'kernel': 'poly'}
0.958242 (0.031970) with: {'C': 0.3, 'kernel': 'rbf'}
0.958242 (0.033368) with: {'C': 0.3, 'kernel': 'sigmoid'}
0.956044 (0.030933) with: {'C': 0.5, 'kernel': 'linear'}
0.881319 (0.050677) with: {'C': 0.5, 'kernel': 'poly'}
0.964835 (0.029906) with: {'C': 0.5, 'kernel': 'rbf'}
0.953846 (0.026785) with: {'C': 0.5, 'kernel': 'sigmoid'}
0.953846 (0.031587) with: {'C': 0.7, 'kernel': 'linear'}
0.885714 (0.038199) with: {'C': 0.7, 'kernel': 'poly'}
0.967033 (0.037271) with: {'C': 0.7, 'kernel': 'rbf'}
0.953846 (0.028513) with: {'C': 0.7, 'kernel': 'sigmoid'}
0.951648 (0.028834) with: {'C': 0.9, 'kernel': 'linear'}
0.887912 (0.038950) with: {'C': 0.9, 'kernel': 'poly'}
0.967033 (0.037271) with: {'C': 0.9, 'kernel': 'rbf'}
0.949451 (0.034009) with: {'C': 0.9, 'kernel': 'sigmoid'}
0.953846 (0.026546) with: {'C': 1.0, 'kernel': 'linear'}
0.890110 (0.038311) with: {'C': 1.0, 'kernel': 'poly'}
0.967033 (0.033027) with: {'C': 1.0, 'kernel': 'rbf'}
0.947253 (0.032755) with: {'C': 1.0, 'kernel': 'sigmoid'}
0.956044 (0.025765) with: {'C': 1.3, 'kernel': 'linear'}
0.894505 (0.039427) with: {'C': 1.3, 'kernel': 'poly'}
0.967033 (0.028188) with: {'C': 1.3, 'kernel': 'rbf'}
0.942857 (0.031144) with: {'C': 1.3, 'kernel': 'sigmoid'}
0.958242 (0.024765) with: {'C': 1.5, 'kernel': 'linear'}
0.896703 (0.039791) with: {'C': 1.5, 'kernel': 'poly'}
0.967033 (0.028188) with: {'C': 1.5, 'kernel': 'rbf'}
0.940659 (0.035237) with: {'C': 1.5, 'kernel': 'sigmoid'}
0.956044 (0.021766) with: {'C': 1.7, 'kernel': 'linear'}

0.903297 (0.033409) with: {'C': 1.7, 'kernel': 'poly'}
0.967033 (0.024479) with: {'C': 1.7, 'kernel': 'rbf'}
0.945055 (0.035539) with: {'C': 1.7, 'kernel': 'sigmoid'}
0.956044 (0.021766) with: {'C': 2.0, 'kernel': 'linear'}
0.909890 (0.033680) with: {'C': 2.0, 'kernel': 'poly'}
0.969231 (0.022370) with: {'C': 2.0, 'kernel': 'rbf'}
0.931868 (0.028237) with: {'C': 2.0, 'kernel': 'sigmoid'}
Application of SVC on dataset:
Run Time: 0.007997

Accuracy Score:
Accuracy score 0.991228

	precision	recall	f1-score	support
0	1.00	0.99	0.99	75
1	0.97	1.00	0.99	39
avg / total	0.99	0.99	0.99	114

[[74 1]
[0 39]]

3. CONCLUSION

The conclusion we can see that the accuracy we achieved of 99.1% on the held-out test dataset. From the confusion matrix, there is only 1 case of miss-classification. The performance of this algorithm is expected to be high given the symptoms for breast cancer should exhibit certain clear patterns.

The model shows that supervised learning on this model of SVM gives the best accuracy of 99.1% which is amazing.

4. REFERENCES

[1] Ahmad Taher Azar, Shaimaa Ahmed El-Said, “Probabilistic neural network for breast cancer classification,” Neural Computing and Applications, Springer, vol. 23, pp.1737-1751, 2013.
[2] Emina Alicˇkovic´, Abdulhamit Subasi, “Breast cancer diagnosis using GA feature selection and Rotation Forest”

Neural Computing and Applications, Springer, Volume 28, issue 4, pp 753–763, April 2017.

[3] Fadzil Ahmad, Nor Ashidi Mat Isa, Zakaria Hussain, Siti Noraini Sulaiman, "A genetic algorithm-based multi-objective optimization of an artificial neural network classifier for breast cancer diagnosis", Neural Computing and Applications, Springer, Volume 23, Issue 5, pp 1427–1435 ,October 2013,.

[4] M. K. Hasan, M. M. Islam and M. M. A. Hashem, "Mathematical model development to detect breast cancer using multigene genetic programming," 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), Dhaka, pp. 574-579, 2016.

[5] H. AttyaLafta, N. KdhimAyoob and A. A. Hussein, "Breast cancer diagnosis using genetic algorithm for training feed forward back propagation," 2017 Annual Conference on New Trends in Information & Communications Technology Applications (NTICT), Baghdad, pp. 144- 149, 2017.

[6] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), Ras Al Khaimah, pp. 1-4, 2016.

