Breast Cancer Prediction using Some Machine Learning Models by Dimensionality Reduction of Various Features

Article · February 2022

CITATIONS
0

READS
172

2 authors, including:

Dhanalakshmi Subramanian
SRM Institute of Science and Technology
12 PUBLICATIONS **5** CITATIONS

SEE PROFILE

Indian Journal of Natural Sciences www.tnsroindia.org.in ©IJONS

Vol.12 / Issue 70 / February / 2022 *International Bimonthly (Print)* *ISSN: 0976 – 0997*

RESEARCH ARTICLE

Breast Cancer Prediction using Some Machine Learning Models by Dimensionality Reduction of Various Features

S Dhanalakshmi[1]* and S. Thirunavukkarasu[2]

[1]Assitant Professor, Department of Mathematics, Faculty of Engineering and Technology, SRM IST, Ramapuram, Chennai - 600089, Tamil Nadu, India
[2]Assitant Professor, Department of Information Technology, Faculty of Engineering and Technology, BIHER, Chennai - 600073, Tamil Nadu, India

Received: 12 Nov 2021 Revised: 16 Dec 2021 Accepted: 11 Jan 2022

***Address for Correspondence**
S Dhanalakshmi
Assitant Professor,
Department of Mathematics,
Faculty of Engineering and Technology,
SRM IST, Ramapuram,
Chennai - 600089, Tamil Nadu, India
Email: dhanalas1@srmist.edu.in

ABSTRACT

Breast cancer is a disease that affects a large portion of the global population and is the world's second leading cause of death. At the same time, if detected early enough, it is one of the most treatable cancers. Early detection of breast cancer improves the prognosis and chances of survival by allowing patients to receive timely clinical treatment. Patients could avoid unneeded therapies if benign tumours were classified more precisely. Machine learning is a branch of artificial intelligence that use a variety of statistical, probabilistic, and optimization techniques to enable computers to "learn" from past examples and recognise difficult-to-find patterns in large, noisy, or complex data sets. In this work, we have used various supervised machine learning algorithms for predicting breast cancer problem. To improve the accuracy of those machine learning models from the existing one, we have used the PCA and LDA **to** reduce the dimensionality of the features, which helps to increase the efficiency of models with faster rate.

Keywords: Breast cancer, Machine learning, Dimensionality reduction, Supervised ML algorithms.

INTRODUCTION

The proper diagnosis of important information is a crucial issue in the field of medical research. The medical planning officer's knowledge and skill in the medical profession is frequently used to diagnose the ailment. As a result, there are instances of inaccuracies, unintended biases and the requirement for a lengthy period of time for a precise diagnosis of sickness. Breast cancer is one of the most harmful and heterogeneous diseases in today's world, killing a large number of people all over the globe. Early detection and prognosis of a cancer type has become a necessity in cancer research since it can help with patient clinical treatment. Many research teams, both biomedical and -bioinformatics, have been drawn to the need of classifying cancer patients into high and low risk groups.[1,2]As a result, these methods have been used to model the progression and therapy of malignant diseases. Furthermore, the ability of machine learning algorithms to find essential features in complicated datasets demonstrates their value. Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs), Decision Trees (DTs) are examples of these techniques. Even though it is clear that the application of machine learning algorithms can increase our understanding of cancer progression, adequate validation is required before these technologies can be used in clinical practice [3-5]. Lot of works have been carried out in diagnosing breast cancer with the help of machine learning techniques. Sun et al. presented a work on comparing feature selection approaches for unified breast cancer diagnosis in mammograms in 2005 [6]. Zheg et al. contributed his idea on a support vector machine (SVM) and a K-means algorithm for breast cancer diagnosis in 2014 [7]. Alireza Osarech and Bita Shadgar analysed the SVM classification algorithm to obtain 98.80 percent and 96.63 percent accuracy on two separate benchmark datasets for breast cancer[8].Two datasets from the UCI depository were used to test the classification techniques. One dataset was used for illness identification (WBCD), and the other was used to predict recurrence [9]. While working on breast cancer prediction, S.Kharya proposed Artificial Neural Networks (ANN). The study emphasised the benefits of applying machine learning methods such as SVM, Naive Bayes, Neural Networks, and Decisive Analysis.[10]

An Overview on Machine Learning

ML techniques often incorporate a learning process with the goal of learning to accomplish a task from "experience" (training data). In machine learning, data is made up of features, observations with labels. An individual example is usually described by a set of qualities, often known as features or variables, distinctive feature. The values can be nominal (enumeration), binary (i.e., 0 or 1), ordinal (e.g., A+ or B), or quantitative (integer, real). Machine Learning models and algorithms are evaluated using a variety of statistical and mathematical approaches. After completing the learning process our model performs the classification process based on training data. Also it will perform the prediction process associated with testing data.

Task of Learning

We can classify the ML task depends up on the dataset or features whether supervised or unsupervised learning. If the given dataset contain labelled data which called as Supervised Learning otherwise called as unsupervised learning. In supervised learning, we have to provide the well labelled data to teach or train the machine, which means each data is already combined with correct output data. The supervised learning which are classified into classification and Regression. Unsupervised learning is the reverse of supervised learning that means doesn't use any labelled data to train the machine. There isn't any one data already combined with correct output data. The machine doesn't follow others guidance for determining the required output.

Data Exploration and Cleaning

Remember that the quality of your inputs dictates the quality of your output. It makes reasonable to invest a large amount of time and attention to this step once you've finalised your company hypothesis. In my opinion, data research, cleaning, and preparation could take up to 70% of your total project time. The procedure for understanding, cleaning, and preparing your data for constructing a predictive model are outlined below:

Variable Identification
Univariate Analysis
Bi-variate Analysis
Missing values treatment
Outlier treatment
Variable transformation
Variable creation
In this paper,Jupiter Notebook tool has been used to implement the Breast Cancer datasetTo initiate, the following methods in Jupiter notebook are used to explore and clean data

Data Visualization
A key component of a data scientist's job is data visualisation. Exploratory Data Analysis (EDA) is frequently used in the early stages of a project to obtain insight into your data. Using visuals to make things clearer and easier to grasp is quite beneficial, especially when dealing with massive, multidimensional datasets. It's critical to be able to present your findings at the end of your project. Because visualisation makes it simpler to spot patterns, trends, and outliers, as well as providing a clear, better, and more dependable result, it is used in this work by constructing a count plot, a pair plot, and a scatter plot. Data visualisation is done with the help of the seaborn library in this work

Count Plot

```
ax=sns.countplot(df['diagnosis'],label="Count")
B,M=df['diagnosis'].value_counts()
print('Number of Benign', B)
print('Number of Malignant', M)
```

Number of Benign: 357
Number of Malignant: 212

The above count plot graph clearly shows that the data set contains a greater number of benign (B) stage cancer tumours, which can be cured. Also displays there is less number of malignant (M) stage of cancer tumours in the dataset which is having minimum possibility to be cured. Diagnosis column of a dataset have shown that 357 are benign patients and 212 are malignant patients.

Model Selection
Different categories of machine learning algorithm has been developed. Based on our dataset we have to choose the algorithm whether supervised or unsupervised. As our dataset has received labeled data here used supervised learning algorithms. Such classification algorithms are given as follows:

- Logistic Regression
- Decision Tree Classifier
- Naïve Bayes
- Random Forest Classifier

The following is how the data is set up for the model:

Split data set
In the given dataset, first we splitting the data into feature dataset, which is known as independent data set (X), and a target data set which is known as dependent data set (Y).Here we are applying the python conditional operator for getting the independent feature of our dataset. Here assigned all the rows and columns whatever given in the breast cancer dataset except diagnosis of class label into X variable. Then assigned only diagnosis columns of dependent value into Y.

```
x=df.iloc[:,df.columns!='diagnosis'].values
y=df['diagnosis'].values
```

Split up of Independent and Dependent Variable

The train-test split procedure is used to measure the performance of machine learning algorithms that make predictions on data that was not used to train the model. The train-test split technique which are mainly used to evaluate the performance of a machine learning algorithm. It can be used for any supervised learning technique and can be utilised for classification or regression tasks. Here we have used scikit Python machine learning library to implement the train-test split evaluation procedure via the train_test_split function.

```
from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test=train_test_split(x,y,test_size=0.25,random_state=42)
```

Splitting of 75% training set and 25% test set from DataFrame

Feature scaling

Feature scaling is a method of grouping the data's individual features into a range. During data pre-processing, it is utilised to handle dramatically shifting magnitudes, values, or units. A machine learning algorithm will presume larger values to be higher and smaller values to be lower if feature scaling is not done, regardless of the unit of measurement.

```
from sklearn.preprocessing import StandardScaler
sc=StandardScaler()
X_train=sc.fit_transform(X_train)
X_test=sc.transform(X_test)
```

Feature scaling

PCA and LDA

Principal Component Analysis is an unsupervised learning approach used in machine learning to reduce dimensionality. Feature scaling is a technique for grouping the data's independent features into a range. During data pre-processing, it's utilised to deal with magnitudes, values, and units that change significantly. A machine learning algorithm will presume that larger values are greater and smaller values are lower if feature scaling is not performed, regardless of the unit of measurement.

For example, a variable with a range of 0 to 100 will outperform a variable with a range of 0 to 1

The PCA algorithm is based on the following mathematical concepts:

- Variance and Covariance
- Eigenvalues and Eigen Factors

Linear discriminant analysis is supervised machine learning and it's a methodology for separating two or more classes of objects or events by finding a linear combination of features. Linear discriminant analysis, often known as LDA, separates numerous classes by computing the directions ("linear discriminants") that indicate the axis that improves separation. LDA is a technique for reducing dimensionality. As the name denotes dimensionality reduction techniques reduce the number of dimensions (i.e, variables or dimensions or features) in a dataset while withholding as much information as possible. The following machine learning codes of PCA and LDA has done dimensionality reduction for our breast cancer dataset. Here sklearn.decomposition and sklearn. discriminant analysis libraries has been used to import the PCA and LDA for dimensionality reduction from its features. By combining all the 31

features into a single feature without affecting the original value, we have used the above said analysis and its respective coding are shown below

```
from sklearn.decomposition import PCA
pca=PCA(n_components=1)
pca=PCA(0.9)
X_train=pca.fit_transform(X_train)
X_test=pca.transform(X_test)
explained_variance=pca.explained_variance_ratio_
```

```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
lda=LinearDiscriminantAnalysis()
X_train=lda.fit_transform(X_train,Y_train)
X_test=lda.transform(X_test)
explained_variance=lda.explained_variance_ratio_
explained_variance
```

```
array([1.])
```

Dimensionality Reduction by using PCA and LDA

Create a function for tracking different models

A function is built to hold all of the models that were used to classify the data in the dataset. The sklearn package in Python is used to import all the methods of classification algorithms. Sklearn library is used to import all the models to get the training accuracyin a function. Then fitting operation has been applied in all the models to get high accuracy. LogisticRegression is used as a first model .Secondly K Neighbours Classifier method is used to get the accuracy by taking three arguments such as neighbors=5,metric='minkowski' and p=2.Next SVM algorithm has beenused to analyse the output by SVC linear method of svm by considering two parameters such as kernel='linear', random state=0. Naïve Bayes Algorithm explored the accuracy using Gaussian NB model. To get accuracy state, Decision tree classifier tree class to use Decision Tree Algorithm and it gets two parametersas an argument such as criterion='entropy', random state=0. Finally RandomForestClassifier is imported through sklearn. Ensemble library. This model took one argument as estimators=100. It helps to improve decision tree accuracy by reducing over fitting. Then it is flexible to both classification and regression problems. We passed X_train and Y_train value to already defined function then received the accuracy by using the score method. Let us see the training accuracy associated with above mentioned models as given as follows

```
model=models(X_train,Y_train)
```

```
Logistic Regression Training Accuracy: 0.9694835680751174
KNearbors Training Accuracy: 0.9624413145539906
Support Vector Machine (Linear Classifier) Training Accuracy: 0.9694835680751174
Gaussion Naive Bayes Training Accuracy: 0.9694835680751174
Decision Tree Classifier Training Accuracy: 1.0
RandomForest Classifier Training Accuracy: 1.0
```

Accuracies of training set model

The correctness of testing data is utilised to test the model. Imported the confusion matrix method of the metric class to check the accuracy. The confusion matrix shown the accuracy of a classifier by comparing the actual and predicted classes. Confusion matrix determines and gives about right and error values by classification model. The number of

right and inaccurate predictions is broken by class and summarised with count values.It reveals the types of errors made by the classifier as well as the errors themselves.

```
from sklearn.metrics import confusion_matrix
for i in range(len(model)):
  cm=confusion_matrix(Y_test, model[i].predict(X_test))

  TP=cm[1][1]
  TN=cm[0][0]
  FP=cm[0][1]
  FN=cm[1][0]

  print(cm)
  print('Model[{}] Testing Accuracy = "{}!"'.format(i, (TP+TN)/(TP+TN+FN+FP)))
  print()
```

```
[[86  3]
 [ 1 53]]
Model[0] Testing Accuracy = "0.972027972027972!"

[[87  2]
 [ 2 52]]
Model[1] Testing Accuracy = "0.972027972027972!"

[[86  3]
 [ 1 53]]
Model[2] Testing Accuracy = "0.972027972027972!"

[[86  3]
 [ 1 53]]
Model[3] Testing Accuracy = "0.972027972027972!"

[[86  3]
 [ 2 52]]
Model[4] Testing Accuracy = "0.965034965034965!"

[[86  3]
 [ 2 52]]
Model[5] Testing Accuracy = "0.965034965034965!"
```

Confusion matrix of test set

Above displayed the confusion matrix accuracy of each model. First four models has produced high accuracy when compared to rest of the model. That means given the low error value from the existing one. The remaining two models also produce the better accuracy. The accuracy details are given in the following topics over confusion matrix with heat mapdiagram.

Cross Validation

The two common difficulties in machine learning are over fitting and under fitting both of which affect the performance of machine learning models.The main goal of each machine learning model is to produce the best result. After dividing the training data from the dataset which will produce stable and detailed output. Hence, we will get the better outcome of our model based on the variation of over fitting and under fitting. To overcome the respective problem, we have to apply the cross validation. The K Fold Cross-Validation has been applied to divide the input dataset into K groups of samples of equal sizes. The prediction function uses k-1 folds for each learning set, while the rest of the folds are used for the test set. This strategy is often used in CVs. Since it is simple to grasp and produces less biased results than methods. The 10 fold cross validation to be used to validate the result in the model

```
from sklearn.model_selection import cross_val_score
cross_validation0 = cross_val_score(model[0], X=X_train, y=Y_train, cv=10)
cross_validation1 = cross_val_score(model[1], X=X_train, y=Y_train, cv=10)
cross_validation2 = cross_val_score(model[2], X=X_train, y=Y_train, cv=10)
cross_validation3 = cross_val_score(model[3], X=X_train, y=Y_train, cv=10)
cross_validation4 = cross_val_score(model[4], X=X_train, y=Y_train, cv=10)
cross_validation5 = cross_val_score(model[5], X=X_train, y=Y_train, cv=10)

cross_validation0.mean(), cross_validation1.mean(), cross_validation2.mean(), cross_validation3.mean(), cross_validation4.
```

```
(0.9670542635658915,
 0.9600221483942415,
 0.9646733111849392,
 0.9693798449612403,
 0.9483388704318937,
 0.9483388704318937)
```

Cross validation result

RESULT AND DISCUSSION

The following table is for different machine learning models with its accuracy in training set, testing set and their corresponding cross validation. Above table indicates, cross validation accuracy is balanced with the training set accuracy and testing set accuracy of first four models whereas the cross validation accuracy is imbalanced with training set accuracy of 100% as well as testing set accuracy of 96.59% in both the Decision Tree Classifier and Random Forest Classifier models.

CONCLUSION

Breast cancer is the most common cause of mortality in women, and it is the only type of cancer that affects them worldwide. Several machine learning models have been developed for early identification and treatment of breast cancer, as well as to reduce the number of fatalities from the disease. Many breast cancer diagnosis approaches have been employed to improve diagnostic accuracy. This research looks at various supervised machine learning methods in order to find the most accurate model for breast cancer detection. This work aimed to improve predictive models using Python in order to improve accuracy in forecasting accurate outcomes. The examination of the results indicates that the combination of data, feature scaling and various classification methods and analyses create a highly effective tool for prediction. It is concluded from the previous work, our research contributes more accuracy with the help of PCA and LDA as discussed in the paper. Even though we arrived the optimal accuracy, the most exact model is required to improve classification techniques' performance and predict accuracy as close to 100% as it is most important to save the human life.

REFERENCES

1. West, D., Mangiameli, P., Rampal, R., & West, V. (2005). Ensemble strategies for a medical diagnosis decision support system: A breast cancer diagnosis application. European Journal of Operational Research (162), 532–551
2. Barracliffe, L.; Arandjelović, O.; Humphris, G. A pilot study of breast cancer patients: Can machine learning predict healthcare professionals' responses to patient emotions? In Proceedings of the International Conference on Bioinformatics and Computational Biology, Honolulu, HI, USA, 20–22 March 2017; pp. 101–106.

3. A. J. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," Cancer Informatics, vol. 2, pp. 59–77, 2006.
4. M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," Expert Systems with Applications, vol. 36, no. 2, pp. 3240–3247, 2009.
5. Sharma, A.; Kulshrestha, S.; Daniel, S. Machine learning approaches for breast cancer diagnosis and prognosis. In Proceedings of the International Conference on Soft Computing and Its Engineering Applications, Changa, India, 1–2 December 2017.
6. Y. Sun, C. F. Babbs, and E. J. Delp, "A comparison of feature selection methods for the detection of breast cancers in mammograms: adaptive sequential floating search vs. genetic algorithm," in Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, pp. 6532–6535, Shanghai, China, September 2005.(15)
7. B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," Expert Systems with Applications, vol. 41, no. 4, pp. 1476–1482, 2014(17)
8. Alireza Osarech, Bita Shadgar, "A Computer Aided Diagnosis System for Breast Cancer", International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011(21)
9. Mandeep Rana, Pooja Chandorkar and Alishiba Dsouza, "Breast cancer diagnosis and recurrence prediction using machine learning techniques", International Journal of Research in Engineering and Technology Volume 04, Issue 04, April 2015.(22)
10. D.Dubey, S.Kharya and S.Soni, "Predictive Machine Learning techniques for Breast Cancer Detection", International Journal of Computer Science and Information Technologies, Vol.4 (6), 2013, 1023-1028.

Table 1. Comparison of Different Machine Learning Models Accuracy

Machine Learning Models	Training Set Accuracy	Testing Set Accuracy	Cross Validation Accuracy
Logistic Regression	96.94%	97.20%	96.70%
KNeighbors Classifier	96.24%	97.20%	96%
Support Vector Machine	96.94%	97.20%	96.46%
Naïve-Bayes Classification	96.94%	97.20%	96.93%
Decision Tree Classifier	100%	96.59%	94.83%
Random Forest Classifier	100%	96.59%	94.83%

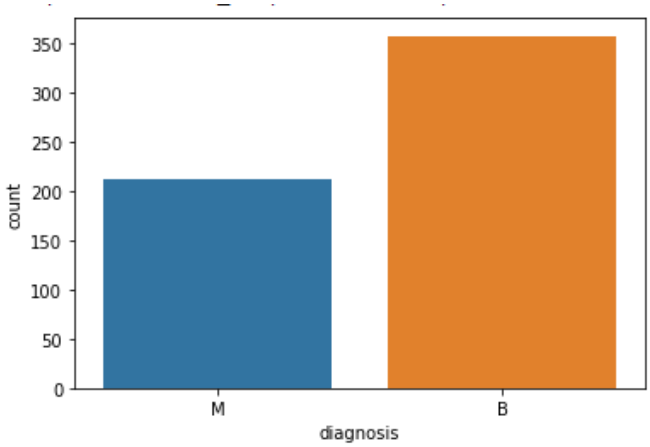

Fig.1. Count plot for Malignant & Benign patients