Prediction of Disease is Malignant or benign for breast cancer

1Yogesh Mangal, 2Ishika Gupta, 3Garvit Kashyap
Students
Galgotias University

Abstract: Breast cancer is the second most leading cancer occurring in women compared to all other cancers. Around 11 lakh cases were recorded in the year 2004. These rates of this cancer increases with the main cause of industrialization and urbanization. This problem is most common in high-income countries but now it is increasing rapidly in both middle and low income countries that includes most of Asia, and Latin America. Breast cancer is much fatal in under half of all cases and is the leading cause of death from cancer in women, accounting for 16% of all cancer deaths worldwide. In this Research Paper, the aim is to predict the breast cancer effect i.e. it is going to be malignant or benign. This prediction help Doctor’s and many peoples to save precious life of patients.

Keywords: Numpy, Pandas, Scikit-Learn library, Breast cancer data, Principal Component Analysis, Machine Learning

II. Introduction

The number and the size of databases recording breast cancer data are increasing rapidly. Breast cancer dataset, produced from measurements, experiments, and researches, etc., are stored in different databases on a daily bases. So the size of dataset increases day by day. This enormous amount of data exceeds the ability of traditional methods to analyze and search for interesting patterns and information that is hidden in them due to the clash of larger size of dataset. Therefore, new techniques and tools for discovering useful information in these datasets are becoming more demanding. So here comes the most emerging technology called Machine Learning that helps in analyzing these data with new analytical methods to find interesting patterns and hidden knowledge in these data sources.

III. Literature Surveys

Several studies have been reported that have focused on breast cancer survival persons. These studies have applied different approaches to the given problem and achieved high classification accuracies. Some of the information of previous research works are given in the following:

Liu et al, a data scientist used decision table (DT)-based predictive models for breast cancer survivability, and concluded that the survival rate of patients was 86.52%. They employed the under-sampling C5 technique and bagging algorithm to deal with the imbalanced problem, thus improving the predictive performance on breast cancer.

Tan and Gilbert demonstrated the usefulness of employing ensemble methods in classifying micro array data and presented some theoretical explanations on the performance of trail methods. The final result, they suggest that ensemble machine learning should be considered for the task of classifying gene expression data for cancerous data samples.

Chaurasia and Pal compare the performance criterion of supervised learning classifiers, such as Naïve Bayes, SVM-RBF kernel, RBF neural networks, Decision Tree (D) (J48), and simple classification and regression tree (CART), to find the best classifier in breast cancer datasets. The experimental result shows that SVM-RBF kernel is more accurate than other classifiers; it scores at the accuracy level of 96.84% in the Wisconsin Breast Cancer (original) datasets.

Chaurasia and Pal proposed three popular data mining algorithms: CART, ID3, and DT for diagnosing heart diseases, and the results comes out demonstrated that CART obtained higher accuracy within less time. Chaurasia and Pal conducted an experiment to identify the most common data mining algorithms for prediction, implemented in modern Medical Diagnosis, and evaluate their performance on several medical datasets. Five algorithms that they were chosen: Naïve Bayes, RBF Network, Simple Logistic, J48 and Decision Tree. For the evaluation two Irvine Machine Learning Repository (UCI-UC) databases were used: heart dis-ease and breast cancer datasets. Several performance metrics were utilized: percentage(%) of correct classifications, True/False Positive rates, area under the curve (AUC), precision, recall, F-measure, and a set of errors.

IV. Methodology

As this is a Research Paper that focuses on Problem that a disease is malignant or benign, the first and foremost requirement is the knowledge of machine learning, types of machine learning algorithms and their working. Another thing needed is the knowledge of Python. One should be familiar with the Jupyter notebook environment.

The following methodologies are used to evaluate a model:

(a) Collecting the data
(b) Training a model based on data
(c) Evaluating a model
(d) Improving the performance
(e) Model Validation

(a) Collecting the data:
For any analytical process, one needs to take the data carefully. And then fix the issues in data like impurity and missing values. There are several datasets available for Breast cancer disease and one such dataset is available on Scikit-Learn library.

(b) Training a model based on data:
In this step, appropriate algorithm and data representation is to be done in the form of a model. The final data is divided into 2 parts. The first part is called Training Part which is used for developing model and another part is called Testing part which is used as a reference.

(c) Evaluating a model:
For evaluation, the second part of data i.e. testing data is used. This step determines the accuracy in the choice of the algorithm based on the results. The best way to check accuracy of model is to see its performance on that dataset which was not used at all during the model build.

(d) Improving the performance:
This step might involve choosing a different model altogether or introducing more variables to increase the efficiency. This needs significant amount of time to be spent in data collection and preparation.

(e) Model Validation:
The validity of the model can be checked by many ways. The common and most used way is Confusion Matrix.

In confusion matrix, the accuracy of the model is checked by comparing Actual values vs Predicted values.

V. Flow Chart
The flowchart proposed below represents the methodology of the model. The data taken in this model is collected from Scikit-learn library.
Also, the data flow diagram that is a graphical diagram that represents the data flow of a system for this model is given below:

VI. Algorithm applied

Basically, both supervised and unsupervised machine learning algorithms can be applied to build this model, but the most suitable algorithm for this model is unsupervised algorithm because this can helps to find all kind of unknown patterns in data and also in dimensionality reduction.

Unsupervised learning is also of many types. One of them is “Principal Component Analysis”.

Principal Component Analysis also abbreviated as PCA is a statistical procedure that uses an orthogonal transformation which converts a set of correlated variables to a set of uncorrelated variables.

PCA is a most widely used tool in exploratory data analysis and in machine learning for predictive models.

Moreover, PCA is an unsupervised statistical technique used to examine the interrelations among a set of variables. It is also known as a general factor analysis where regression determines a line of best fit.

The need of PCA is because of the problem of over fitting. To reduce the problem of over fitting PCA is needed. PCA try to convert the high dimensionality into a set of low dimensionality.

The graph below here depicts the dimensionality reduction with the help of PCA.
VII. Implementation and Results

For the implementation of the model, one needs to import the machine learning libraries such as numpy, pandas, matplotlib and seaborn. Then the dataset should be imported from scikit-learn library. The dataset should be converted into the dataframe by using pandas library. The data in the dataframe is not in standard format. To make the data in standard format, one needs to import StandardScaler library from scikit preprocessing library. The data is fit in standard form by using fit and transform method. The dataset is now ready for applying PCA and that should be done by import PCA library from sklearn. The PCA is scaled on 2 components, the one component represents the disease is Malignant and the other one represents the disease is Benign. The graph is plotted between these 2 components. The dots on graph for these 2 components is comes of different color that means one color represent the disease is malignant and the other one represents the disease is benign.

Results:
A model is developed on the dataset (collected from scikit learn) that will predict the disease is malignant or benign by the use of Principal Component Analysis algorithm.

VIII. Conclusion and Future Scope

This paper focuses on the model that predicts the disease is malignant or benign on breast cancer data by using PCA algorithm of unsupervised learning. This analyze the breast cancer data in multiple ways. The main focus of this model is to analysis on the breast cancer data with accuracy and provide better diagnostic accuracy. By the prediction of disease at early stage, it will be really helpful to save the precious life of human beings. This paper conclude that the Principal Component Analysis Algorithm can be helpful in disease prediction at early stage.

References: