Breast cancer Data Analysis using Machine Learning Approaches

Introduction

Breast cancer is the most common cancer affecting females worldwide. Correct and early diagnosis is an extremely important step in rehabilitation and treatment. Machine Learning (ML) techniques can be used to develop tools for physicians that can be used as an effective mechanism for early detection and diagnosis of breast cancer which will greatly enhance the survival rate of patients.

Objective

The objective of this study is to learn and compare different machine learning and statistical approaches for analysis of breast cancer data.

Dataset information

For this study we are using breast cancer data set supplied by University of Wisconsin, Clinical Sciences Center, and Madison.

It consists of attributes of digitized image of fine needle aspirate (FNA) of a breast mass. They describe features of the cell nuclei present in the image.

The Wisconsin breast cancer data set is real multivariate data set. It consists of 569 numbers of instances and 32 numbers of attributes are given in the data set.

Dataset attributes are:

- Id number
- Diagnosis
- Radius
- Texture
- Perimeter
- Area
- Smoothness
- Compactness
- Concavity
- Concave points
- Symmetry
- Fractal dimensions

Classification models used:

- Support Vector Machine
- Naïve Bayes
- K-Nearest Neighbor
- Logistic Regression

Methodology used:

The following Fig. 1 shows overview or work flow of methodology.

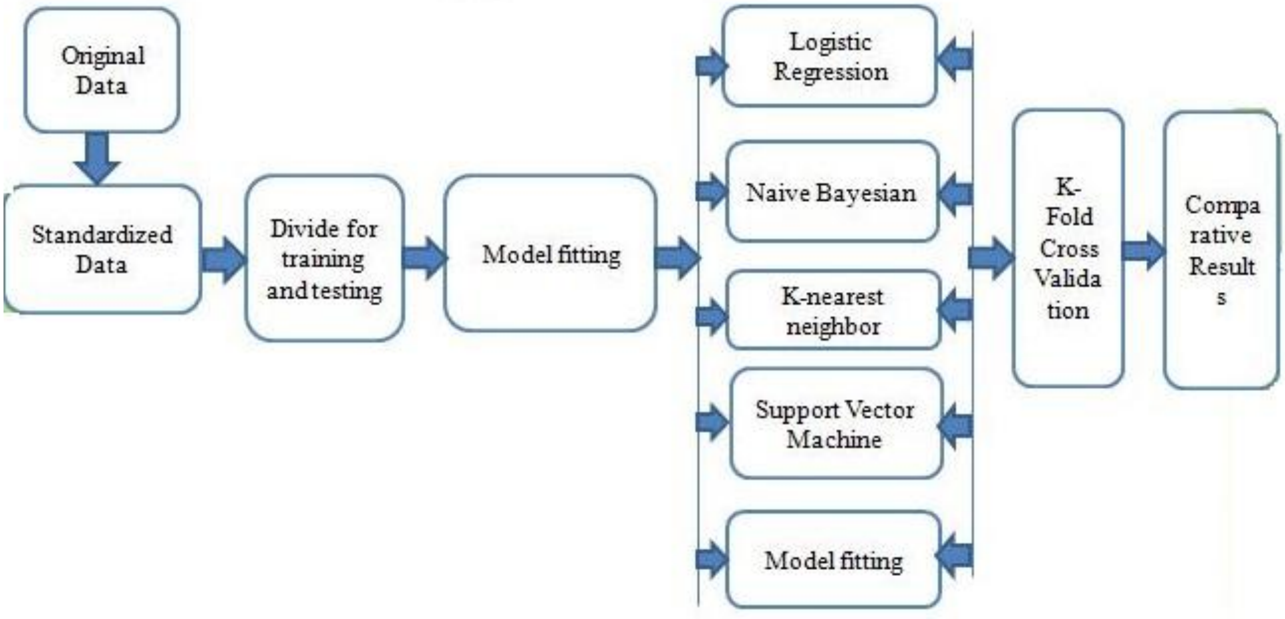

Fig. 1: Flow of work

Mathematical Modeling

For system S,

S= {I, O, F}

I= Input

O=Output

F=Functions to be implemented

In our case,

I= Feature rich dataset of breast cancer

O = diagnosis attribute(binary)

O = {0, 1}

Where,

0 = Benign

1 = Malignant

- F=Functions to be implemented(Classification algorithms)
- F= $\{F_{1,} F_{2,} F_{3,} F_4\}$

- F_1 = Support Vector Machine(SVM)
- F_2 = K-nearest neighbor(K-NN)
- F_3 = Naïve Bayesian
- F_4 = Logistic regression

F_1 is following,

F_1 = Support Vector Machine (SVM) algorithm

$$d(\mathrm{X}^T) = \sum_{i=1}^{l} y_i \alpha_i X_i X^T + \ b_0$$

Where y_i = class label of support vector X_i

X^T= Test tuple

α_i and b_0 = numeric parameters that were determined by SVM algorithm

l = number of support vectors

F_2 is following,

F_2 = K-nearest neighbor(K-NN) algorithm

$$dist(X1, X2) = \sqrt{\sum_{i=1}^{n} (x1i - x2i)^2}$$

Where, dist(X1,X2) = distance between two tuples

X1 = (x11,x12,…..,x1n)

X2 = (x21,x22,…..,x2n)

F_3 is following,

F_3 = Naïve Bayesian algorithm

$$\mathrm{P}(C_i|\mathrm{X}) = \frac{\mathrm{P}(\mathrm{X}\ |C_i)P(C_i)}{\mathrm{P}(\mathrm{X})}$$

Where, X= data tuple

Ci = class C

P(Ci|X) = posterior probability, Ci conditioned on X

P(Ci) = prior probability

P(X|Ci) = posterior probability, X conditioned on Ci

P(X) = prior probability of X

P_4 is following,

P_4 = Logistic regression algorithm

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

Where, p = probability of success (presence of characteristics)

1-p = probability of failure (absence of characteristics)

logit(p) = log odds of dependent variable

Analysis and Evaluation

- It is observed from the plot that there were 357 cases in which the diagnosis of the lump/tumor was benign, and there were about 212 cases in which the diagnosis was malignant out of total 569 subjects.

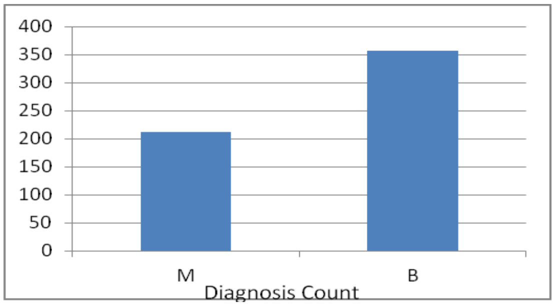

Fig.2: Count plot for split between the malignant and benign diagnosis based on the dataset

- The relationship between these 32 independent variables is shown by correlation matrix in Fig.3. This shows perfect positive correlation between 32(multiple) variables.

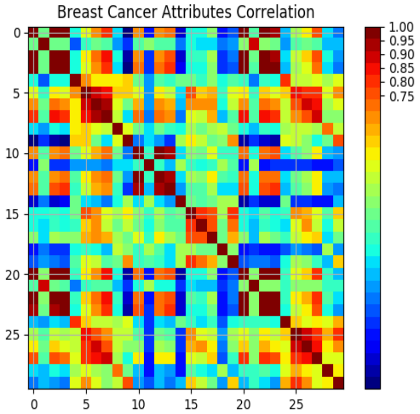

Fig 3: Correlation matrix for relationship between independent variables

- Total predictors available in data set are 32 out of that we are using 10 predictors. We applied correlation of coefficient method to find the impact of relationship between diagnosis and the independent variable (Predictor) The standard deviation and variance are also calculated for finding more details about the proportionally of predictor variables.

Table 4.1: Data analysis using statistics like Mean and Standard Deviation

Mean and Standard Deviation	**Mean and Standard Deviation**
Standard Deviation for radius_mean = 3.524 Mean for radius_mean= 14.127 Standard Deviation for texture_mean = 4.301 Mean for texture_mean = 19.29 Standard Deviation for perimeter_mean = 24.299 Mean for perimeter_mean = 91.969 Standard Deviation for area_mean = 351.914 Mean for area_mean = 654.889 Standard Deviation for smoothness_mean = 0.014 Mean for smoothness_mean = 0.096	Standard Deviation for compactness_mean = 0.053 Mean for compactness_mean= 0.104 Standard Deviation for concavity_mean = 0.08 Mean for concavity_mean = 0.089 Standard Deviation for concave points_mean = 0.039 Mean for concave points_mean = 0.049 Standard Deviation for symmetry_mean = 0.027 Mean for symmetry_mean= 0.181 Standard Deviation for fractal_dimension_mean= 0.007 Mean for fractal _dimension_mean= 0.063

Coefficient of Correlation

- Correlation between diagnosis & radius _mean = 0.730029
- Correlation between diagnosis & texture _mean = 0.415185
- Correlation between diagnosis & perimeter _mean = 0.74264
- Correlation between diagnosis & area _mean = 0.70898
- Correlation between diagnosis & smoothness = 0.35856
- Correlation between diagnosis & compactness _mean = 0.59653
- Correlation between diagnosis & concavity _mean = 0.69636
- Correlation between diagnosis & concave points _mean = 0.7766
- Correlation between diagnosis & symmetry _mean = 0.3305
- Correlation between diagnosis & fractal dimension _mean = -0.0128

The conclusion is major six predictors are highly correlated with the diagnosis are: Radius, perimeter, area, compactness and concavity and concave points. As correlation coefficient value is nearer to 1.

Observation for applicability of ML algorithms and Comparative study

- To validate each prediction model, we used a 10-fold cross validation. Firstly we compute a model performance without standardizing the dataset and then compute a model performance with standardizing the dataset.
- The performance comparison is provided using box plots for LR, SVM, NB, KNN model.

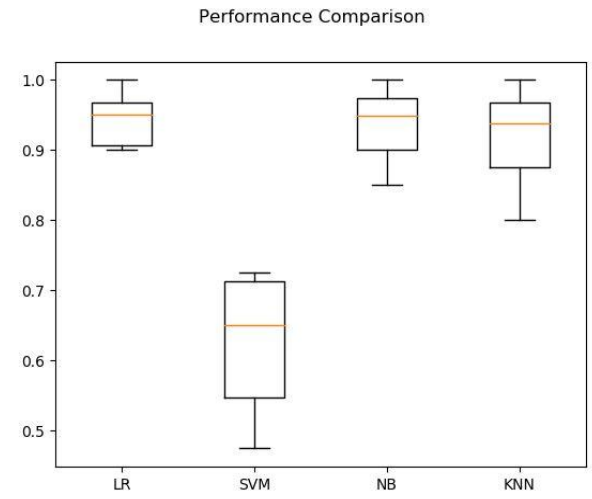

Fig. 4: Performance comparison using Box-plot without standardizing the dataset

- Form Fig. 4, conclusion is each of statistic (median, Q1 and Q3) for SVM is lower than LR, NB and KNN. The performance accuracies for LR and NB are same to some extent.

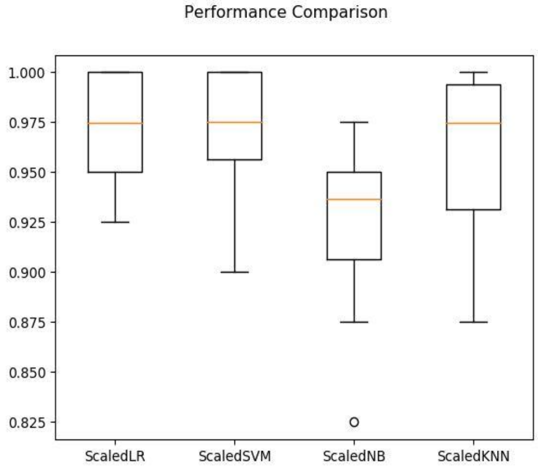

Fig. 5: Performance comparison using Box-plot after standardizing the dataset

- The Fig. 5 provides performance comparison after standardizing the dataset. Here SVM gives improved performance after standardizing the dataset. The one outlying observation is plotted or detected for ScaledNB.