Breast Cancer Prediction using random forest method Document

by
Manideep Maddipatla

Introduction :-

Random forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random forests correct for decision trees' habit of overfitting to their training set.

It predicts the data by :-

1)Using out-of-bag error as an estimate of the generalization error.
2)Measuring variable importance through permutation.
It uses the decision tree.
So here we use decision tree for predicting the data.

Here in this assignment we should create a model that which can predict breast cancer by using the data given by the hospital members.

So here we are provided with the data that which contains the information that which helps to predict cancer. It contains the following attributes Unnamed: 0, id number, clump_thickness uniformity_of_cell_size, uniformity_of_cell_shape marginal_adhesion, epithelial_cell_size,bare_nuclei bland_chromatin, normal_nucleoli,mitoses.

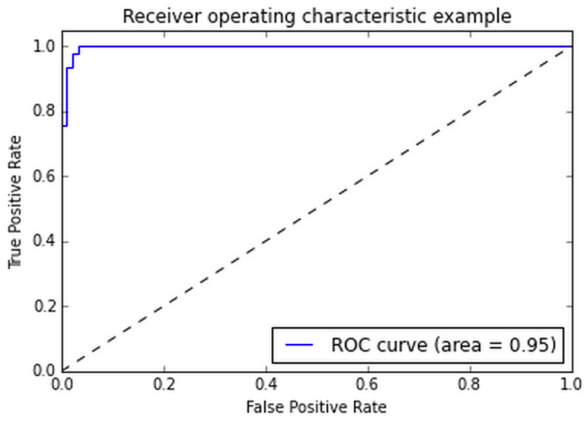
In the above given attributes we can select any of the which can be used to predict the cancer . According to me by using Unnamed: 0, id number we cannot predict the cancer so i had dropped those two tables . And we can select any of the attribute that had mentioned in the file. I had selected uniformity_of_cell_size, uniformity_of_cell_shape marginal_adhesion to predict the data.

Procedure :-

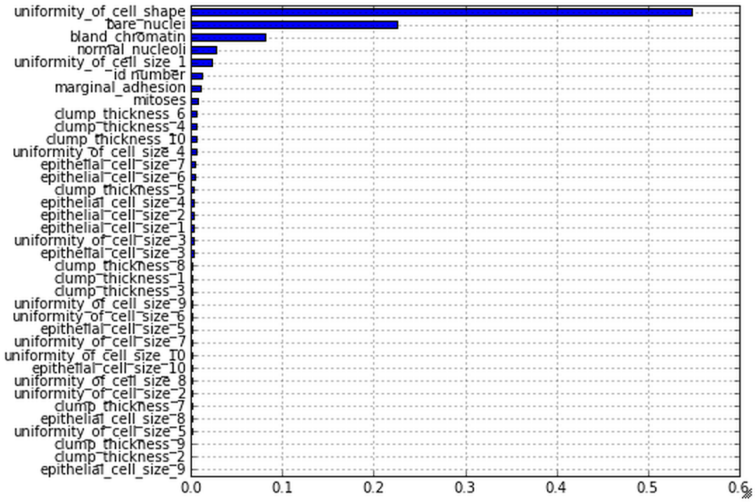
First i taught i can do two models and can check which model gives the best result . so i started with logistic regression method .
in this method i had used all the attributes except Unnamed: 0, id number and i got the following result .

```
---Base Model---
click to scroll output; double click to hide
             precision    recall  f1-score   support

          0       0.68      1.00      0.81        95
          1       0.00      0.00      0.00        45

avg / total       0.46      0.68      0.55       140



---Logistic Model---
Logistic AUC = 0.95
             precision    recall  f1-score   support

          0       0.96      0.99      0.97        95
          1       0.98      0.91      0.94        45

avg / total       0.96      0.96      0.96       140
```

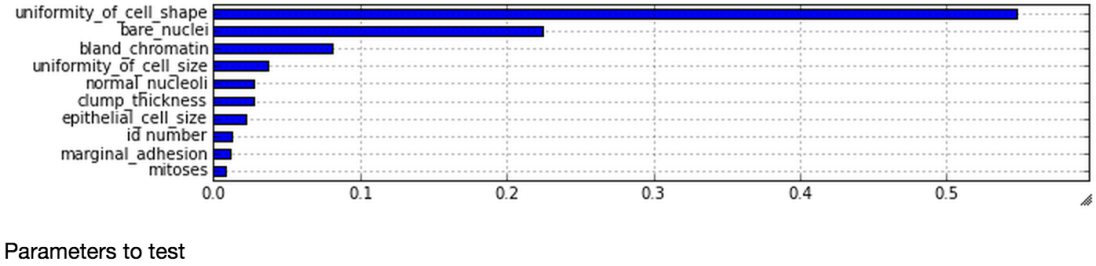
It shows that this model is 95 % accurate
And the roc curve for the above result is as follows
After this i did the Random Forest regression method , As this method is based on the decision tree the data is been divided into many parts and then the data is been compared in different levels we can predict the data by changing the values each time and testing it with test data. Here in this method i had dropped the same attributes that i did in the logistic regression method and tested the data by using few categorical values they are 'clump_thickness', 'uniformity_of_cell_size', ‘epithelial_cell_size’.
By doing this method i had got the following output .



So here i had used logistic regression method to predict the brest cancer and the this methos is 95% accurate.



And after scaling above graph it has bee re arranged as below



In this Random Forest regression model i got the C-Stat value as s C-stat: 0.991085180018 when the final values are as follows n_estimators=2000, oob_score=True,n_jobs=1,random_state=42, max_features="sqrt", min_samples_leaf=5).

Next i had testes the data with RandomForestClassifier model for this i had selected the same categorical values as in the regression method. And the result after testing this method is **C-stat: 0.96.** So by comparing the C-Stat values in both the methods we can say that the regression method is best with C-stat value 0.99.

For comparing these methods the terms we use is precision recall f1-score

In pattern recognition and information retrieval with binary classification, **precision** (also called positive predictive value) is the fraction of retrieved instances that are relevant, while **recall** (also known as sensitivity) is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of **relevance.**
In statistical analysis of binary classification, the **F1 score** (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results, and r is the number of correct positive results divided by the number of positive results that

should have been returned. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0.

so after applying the classifier method to the given data the comparision output obtained is as below.

	precision	recall	f1-score	support
0	0.93	0.99	0.96	87
1	0.98	0.89	0.93	53
avg / total	0.95	0.95	0.95	140

By this result we can say that this method gives 95 % accurate result in predicting the breast cancer.

The problem i observed while doing these methods are the logistic method is difficult and time taking when compared with the random forest method . We should compare each and every attribute separately where as in random forest we can include all the attributes at the same time and compare it with test data.

And the problem we have with this method is as this process gives 99 % of exact prediction in regression method and 96% of exact prediction in classifier method there is few percent of chances for this method to give wrong prediction.

So finally i state that the random forest method is best when compared to logistic regression .

Here are the links i used to write the document

http://en.wikipedia.org/wiki/Random_forest
http://en.wikipedia.org/wiki/Precision_and_recall
http://en.wikipedia.org/wiki/F1_score