VRIJE
UNIVERSITEIT
BRUSSEL

TECHNIQUES OF AI

Breast Cancer Classification

Deepika Chandrababu | 0572230

May 31, 2022

Sciences and Bio-Engineering Sciences

1 Introduction

Problem Statement:

The intent to classify whether the patient has breast cancer or not given from the 3500 instances and 150 features.This is a binary classification problem with the labels (0 and 1) where 0 means Benign who are non-cancerous patients and 1 means malignant for patients who have cancer. Micro-calcifications are calcium deposits in breasts that appear as a natural process of ageing. Some micros are non cancerous(benign) while some are cancerous(Malignant). It is very important to train and test our model without data leakage because the false positive cases and false negative cases are both harmful when dealing with life threatening diseases.

2 Data and Methods

2.1 Target variable

In this binary classification the labels in our data set has information as 0 (non cancer) and 1(cancer). From Figure1, it is seen that the variable is slightly imbalanced but the description says that there are 50 benign cases and 50 malignant cases. Of the 2800 instances in the training set, 1600 occurences are Non-cancerous patients are higher than the cancerous patients which has 1200 occurences. This could be because of the multiple micros that occurs for a single patient. From a brief overview of the dataset, I found that for labels [0] the micros are less and for labels [1] there are multiple micros. So, I grouped the information available for a single patient taking the mean value of the records of their micros every patient in the data set has multiple micros with the same label. Now, Figure2 shows that there are almost 48 non-cancerous benign patients and 50 malignant cancerous patients.

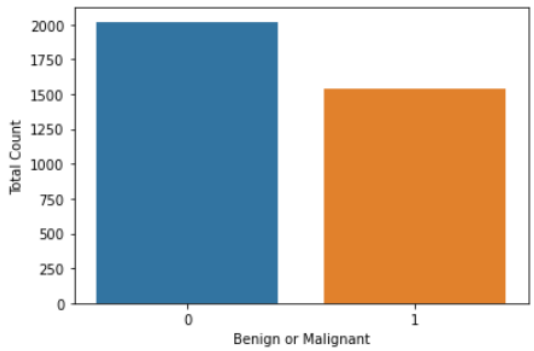

Figure 1: Target Variable with 3500 instances

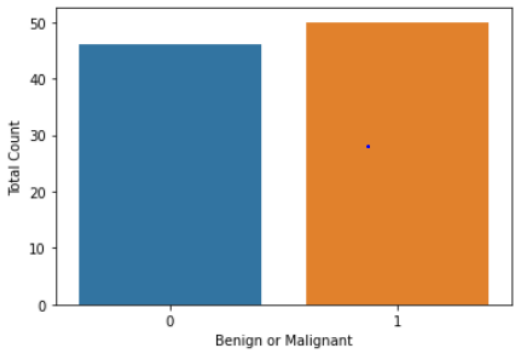

Figure 2: Target Variable with 50 instances

2.2 Univariate Analysis

Univariate analysis was done on the training set to show the distribution of the 150 features. The features has the shape and texture properties that indicate the presence of malignancy in the micros. The first graph in figure 3 is the first feature which shows the original shape elongation and the other features are about the skewness, root mean squared values, and variance in different micros. Since the column names are big and there are 150 instances, the plot is not readable. But from the bars of histograms of each feature it is seen that not all features are important. This is because for some of the features, the bars are left skewed meaning for the values in the range of 0 which will not be helpful to improve our model.

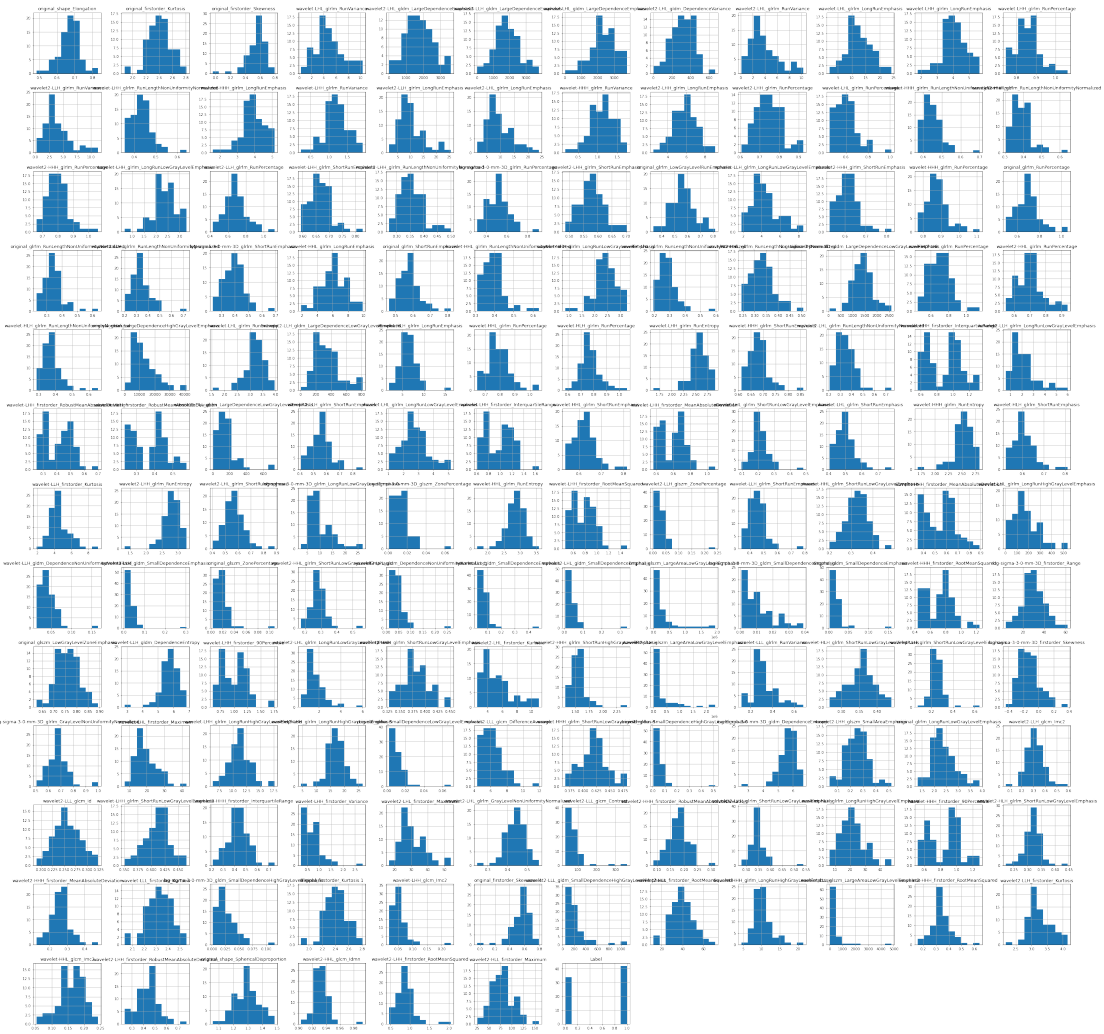

Figure 3: Univariate Analysis

2.3 Train/Test split

The data was divided into train and test set to avoid biases due to information leakage which can lead to incorrect results. 80 percent of the data was used to train the model and the 20 percent of the data was used to test the model performance.

3 Machine Learning Techniques

Two machine learning techniques were used in this data set. One is Extra trees classifier which was used to select the best features and Random forest classifier was used to train and test the model performance.

3.1 Feature Selection

Training data was used to fit the extra tree classifier model to give the best features based on their importance score. Extremely randomized Tree Classifier (Extra trees) fits a number of random decision trees and chooses a best one out of all the subset of branches in the tree. Extra trees adds random trees but improves the predictive accuracy and control over-fitting.

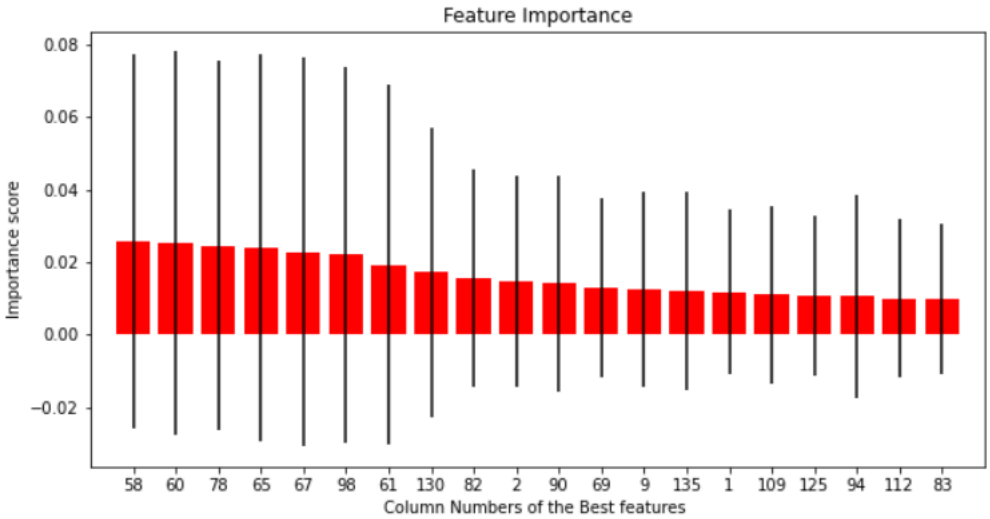

Figure 4: Important Features

In figure 4, the bar graph visualises the top 20 features based on their importance score after training a extra tree classifier model. The importance score is ordered in the scale of best to worst. The y axis shows the range of importance score and x axis shows the important features in terms of column numbers where the best column numbers are shown instead of the long feature names. Hence, the instance 58 has a score of about 0.03 being the highest. The top 6 features falls in the same range of 0.02 importance and then it is seen that the importance of the features decreases with increasing number of instances.

3.2 Model Training

For training the model, top 50 best features were used. The training data was further divided into training and validation set to evaluate the model on the validation set. The following model was used to train and test.

Random Forest Classifier:

The random forest classifier has many decision trees which outputs yes or no. As indicated in the name, it generates n number of random trees choosing a random sample from the dataset. It follows bootstrapping and aggregating methods which is also known as bagging. Using the original data, the model generates multiple random samples (decision trees) of the dataset called bootstrapping in such a way that they are different. After gathering each bootstrap, they are trained independently and the results are aggregated for each decision tree. The average or the highest of the classification results will be used for the final output.

3.3 Model Analysis

Cross Validation: It is crucial to consider the model performance by calculating the cross validated score using the performance metric "f1 score". The learning curve trains the random forest classifier on the training data set using 10 fold cross validation method. In each batch, 90 percent of the data will be used for training and 10 percent of the data will be used for cross validation. The scores of each each sample is aggregated and the mean of them is plotted in figure 5.

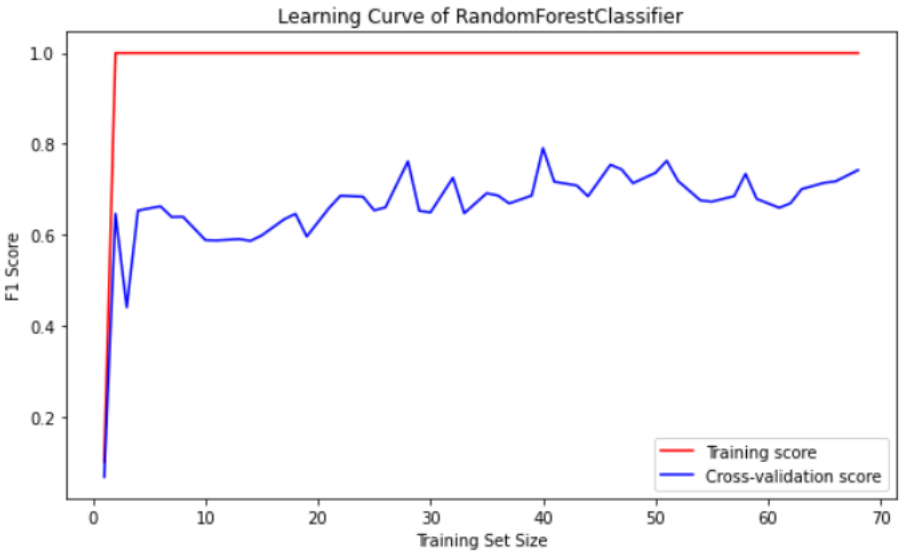

Figure 5: Learning Curve

Learning Curve:

The computed learning curves of the Random forest classier shows that the model suffers from high variance since the training score is higher than the validation score. It is clearly overfitting on the train data set since the score is at an f1-score of 1 constantly. But the validation curve appears to converge at the end with a f1-score close to 0.8 with the increase in sample size.

When I trained the model using top 75 features to increase the sample size, the model did not work better. So the final model was trained with 50 features.

4 Results

4.1 Performance Metrics:

Accuracy works by measuring the observations with respect to the false positive and false negative rates using the formula :

$$accuracy = (tp + tn)/(tp + fp + fn + tn)$$

where,
tp = true positive
fp = false positive
tn = true negative
fn = false negative

In this binary classification where we deal with cancerous and non cancerous classification, it is very crucial to avoid false positive and false negative cases. The f1 score treats the FP and FN equally being the best performance metric in this scenario. Hence, f1 score is calculated which works with the following formula:

$$f1score = tp/tp + 1/2(fp + fn)$$

4.2 Prediction Results:

The scores on the validation set:
Accuracy = 0.68
F1 score = 0.68
The scores on the test set:
Accuracy = 0.8
F1 score = 0.79

It clearly over fits on the validation set but performs well on the test set. Also, the log loss was computed for which the score decreased from 10.9 to 6.9 with respect to the validation and test set..

4.3 Confusion Matrix

The confusion matrix has the actual values vertically and predicted values horizontally. The below table shows the values corresponding to the confusion matrix. where TP = 0.82 meaning the model has predicted 80 percent of the times correctly. For example, Actual Benign and Predicted Benign are 82 percent correct but for 18 percent the actual value was benign and predicted was malignant. The same applies for false positive meaning the actual value was malignant but the predicted was benign and at last there is false negative at 0.78 meaning the actual and predicted values are malignant.

	Benign	Malignant
Benign	True Positive (TP) = 0.82	False Negative (FN) = 0.18
Malignant	False Positive (FP) = 0.22	True Negative (TN) = 0.78

Figure 6 is the matrix computed from the results which proves that most of the classification are true positive and true negative but there are some false negative and false positive rates which the model should minimize.

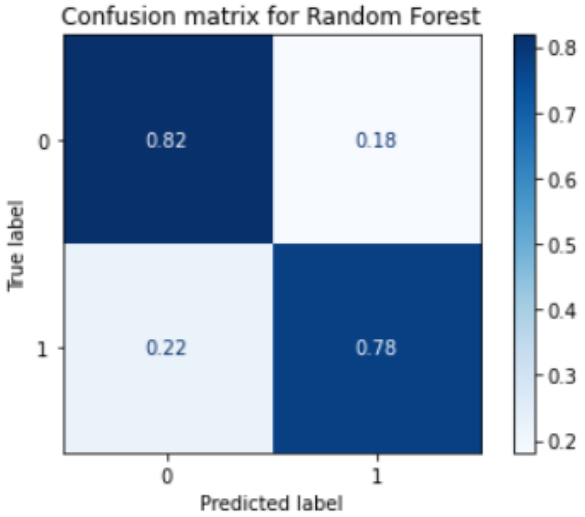

Figure 6: Confusion Matrix

4.4 ROC Curve

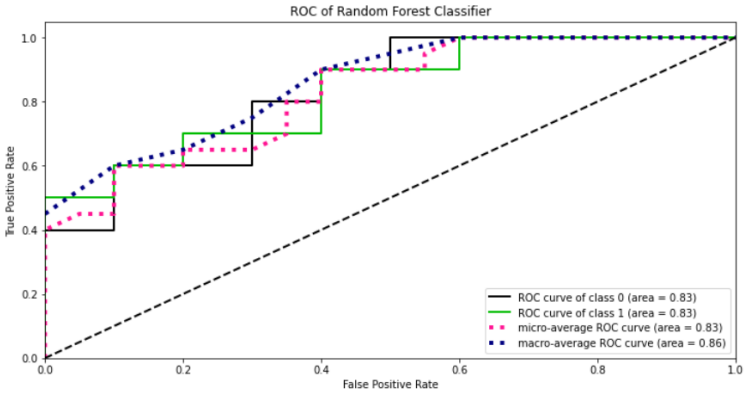

Figure 7: ROC Curve

As seen in the figure 7, the graph shows the trade-off between false positive rate(FPR) on the x axis and true positive rate(TPR) on the y axis. The baseline is the average(0.5) of TPR and FPR, if the predictions are random it indicated a straight line. The area below the dashed line(baseline) will show more false positive rate meaning the worst performance of a model. But for our model(binary classification), the model predicts the probability of the occurence of both

the classes. The ROC curve for both classes (0 and 1) appear above the baseline with an area of 0.83 proving most of the classification are true positives.

5 Discussion

5.1 Task 1

How well can you classify individual micros assuming all micros per subject have the same label?

As multiple rows had information about a single patient and their labels were identical, I grouped all the individual micros per subject having the same label with their mean values. The f1 score of 0.8 shows that 80 percent of the times the model was able to classify the benign and malignant labels correctly. Hence, the model is not over fitting or under fitting but predicts with a good accuracy.

5.2 Task 2

How well can you classify whether a subject has cancer based on your classification of the multiple micros per subject?

It is a good indication to note that on the test data the f1 score is 0.8 which is close to 1 after training a random forest classifier which proves that the model had performed well on the unseen patient records and classified their labels correctly with 80 percent accuracy.

6 Conclusion

The data set had details about multiple micros per subject ending with cancerous or non cancerous. This binary classification problem is solved using random forest classier with an accuracy of 0.8, which proves the model classifies all the micros with respect to their labels 80 percent of the times. One way to achieve the remaining 20 percent of the accuracy is by tuning its hyper parameters but in this case the model was over fitting with an accuracy of 0.99, hence I discarded the tuning and kept the default parameters.

The list of all the websites referred to do this project are cited here. [1] [2] [3] [4] [5] [6] [7]

References

[1] J. Brownlee. Feature selection for machine learning in python. [Online]. Available: https://machinelearningmastery.com/feature-selection-machine-learning-python

[2] Scikit-plot:visualizing machine learning algorithm results performance. [Online]. Available: https://coderzcolumn.com/tutorials/machine-learning/scikit-plot-visualizing-machine-learning-algorithm-results-and-performance3.1-Confusion-Matrix-https://coderzcolumn.com/tutorials/machine-learning/scikit-plot-visualizing-machine-learning-algorithm-results-and-performance3.1-Confusion-Matrix-

[3] Metrics module (api reference). [Online]. Available: https://scikit-plot.readthedocs.io/en/stable/metrics.html

[4] J. Czakon. F1 score vs roc auc vs accuracy vs pr auc: Which evaluation metric should you choose? [Online]. Available: https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc

[5] S. Vishwakarma. Different metrics to evaluate the performance of a machine learning model. [Online]. Available: https://medium.com/analytics-vidhya/different-metrics-to-evaluate-the-performance-of-a-machine-learning-model-90acec9e8726

[6] sklearn.ensemble.extratreesclassifier. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html

[7] S. E. R. Understanding random forest. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/