RESEARCH

Introduction of Profile Areas of Data Science : Project 2

Aman Jain*]FMFrenny Macwan]SCSuresh Kumar Choudhary

Correspondence:
0
Full list of author information is available at the end of the article
*Equal contributor

Abstract

Goal of the Project: To build, and eventually evaluate and analyze the best possible classifier for Breast Cancer Diagnostics, along with which features are the most important to get the proper classification.

Main Result of the Project: The Random Forest classifier was accurate at predicting the Breast Cancer evaluation with an accuracy of about 95

Personal Key Leanings: We created four classifiers for the first time (Regression, Decision Tree, Xgboost Ensemble, Random forest classifiers), and learnt how to pick the best features.

Estimate working hours: 8

Project Evaluation: 1

Number of Words: ¡TD¿

Keywords: sample; article; author

Scientific Background

Breast cancer is a common disease. Fine Needle Aspirations (FNA) provide a way to examine a small amount of tissue from the tumor. Identifying the features of the nucleus (Like size, shape, texture etc), a classifier should be able to predict whether the samples are benign (Non Cancer) or malignant (Cancer).In order to increase the speed, correctness, and objectivity of the diagnosis process, we have used image processing and machine learning techniques. .

Goal

The goal of this project is to create classifiers and chose the best classifier which can predict whether the sample is benign or malignant along with the three most important features which help us determine about the sample.

Data

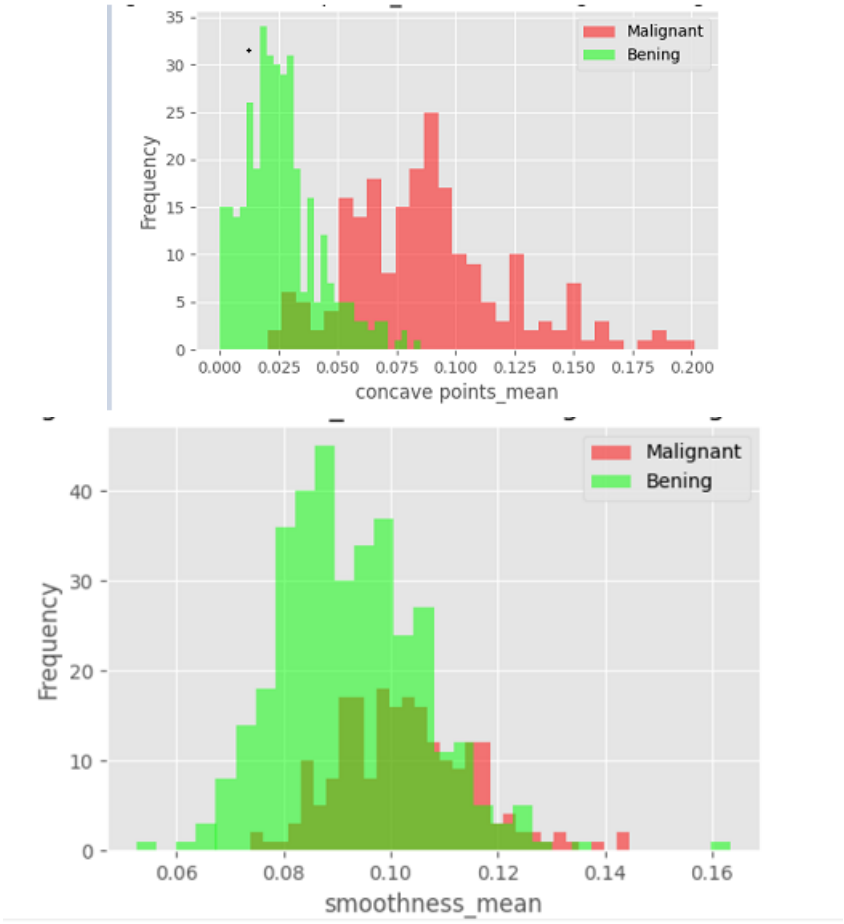
Raw Data was collected of 569 Patients from University of Wisconsin, where we filtered out the features which are not required using the overlapping histograms.

Result

Result obtained as part of this analysis.

Task 1

As part of Task 1 we gathered statistical data like histograms, outliers, summary statistics. We created a summary of the Malignant and Benign tumours and projected it on a histogram, through that we were able to conclude many things like mean radius of malignant tumours are bigger than mean radius of Benign tumours and we were able to detect features which we can use (concave points), features which we cannot use (smoothness) to accurately predict the tumour. After looking at the outliers we also found fractional dimensions have the most values which are out of range and creating a deviation from the mean.



Task 2

As part of Task 2, we built four classifiers namely, Logistic Regression, Decision Tree, Xgboost Ensemble, Random forest classifiers to classify the Malignant and Benign tumours and later on compared the accuracy. All these are supervised learning classifiers. **Logistic regression**: It is a supervised machine learning approach for binary classification. It uses gradient descent approach to converge and provides the weights for each feature after model training so it is easy to understand this model learning. **Decision Tree** : It is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data). **XGBoost** : It is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured

or tabular data. **Random forest** : It is also a very handy algorithm because the default hyperparameters it uses often produce a good prediction result. It selects subsets of feature and train the the model for individual tree

Task 3

As part of Task 3, we calculated the accuracy, precision/recall, ROC analysis and based on the output we evaluated Random Forest classifier performed the best between Random Forest and XGBoost Classifier. True Positive rate of Random Forest was better compared to XGBoost's True Positive Rate. We also generated the ROC Curve for all 4 classifiers and calculated their respective AUCs and found out RandomForest performed the best with AUC as 0.9939. Since Malignant samples are cancerous, and we require more precision because of it's sensitive nature, Random Forest gave a precision of 0.98 whereas XGBoost Classifier gave a precision of 0.95, hence we concluded Random Forest as a better classifier.

```
report(y_test,bst.predict(x_test))#report for XGClassifier
```

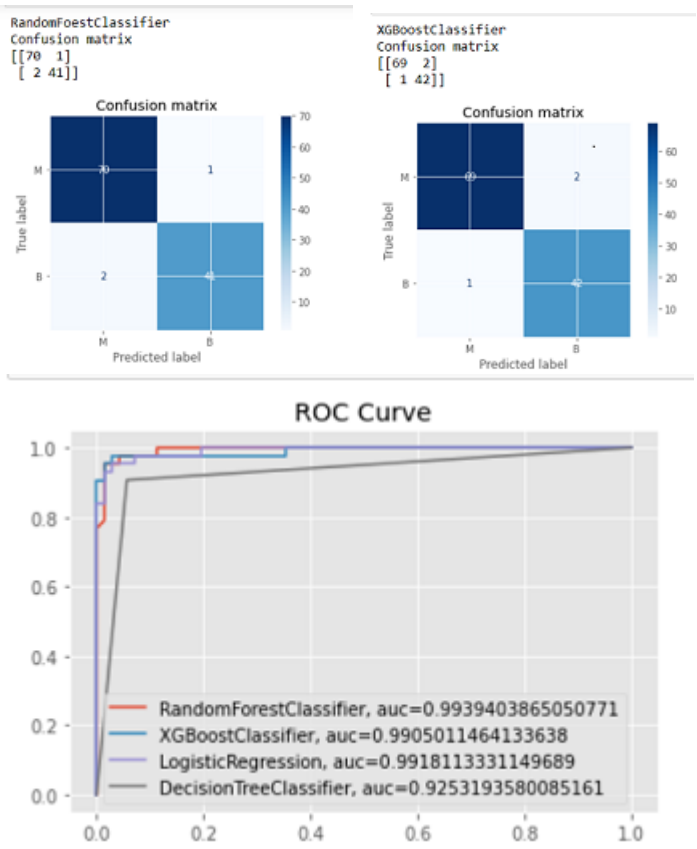
	precision	recall	f1-score	support
B	0.99	0.97	0.98	71
M	0.95	0.98	0.97	43
accuracy			0.97	114
macro avg	0.97	0.97	0.97	114
weighted avg	0.97	0.97	0.97	114

```
Accuracy: 0.9736842105263158
```

```
report(y_test,clf_rf.predict(x_test))#report for Random forest
```

	precision	recall	f1-score	support
B	0.97	0.99	0.98	71
M	0.98	0.95	0.96	43
accuracy			0.97	114
macro avg	0.97	0.97	0.97	114
weighted avg	0.97	0.97	0.97	114

```
Accuracy: 0.9736842105263158
```

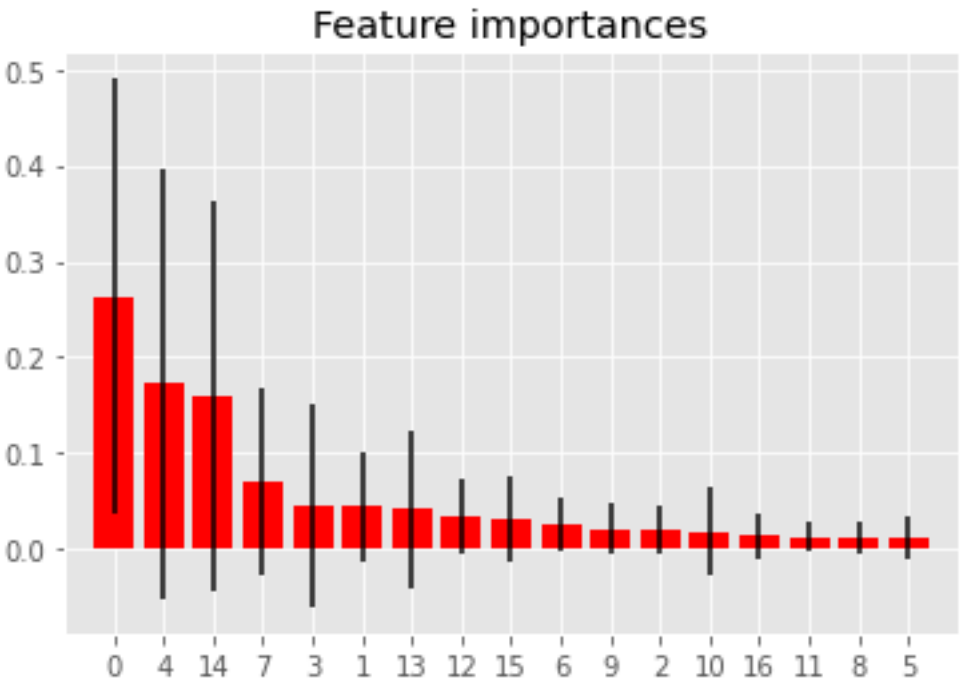


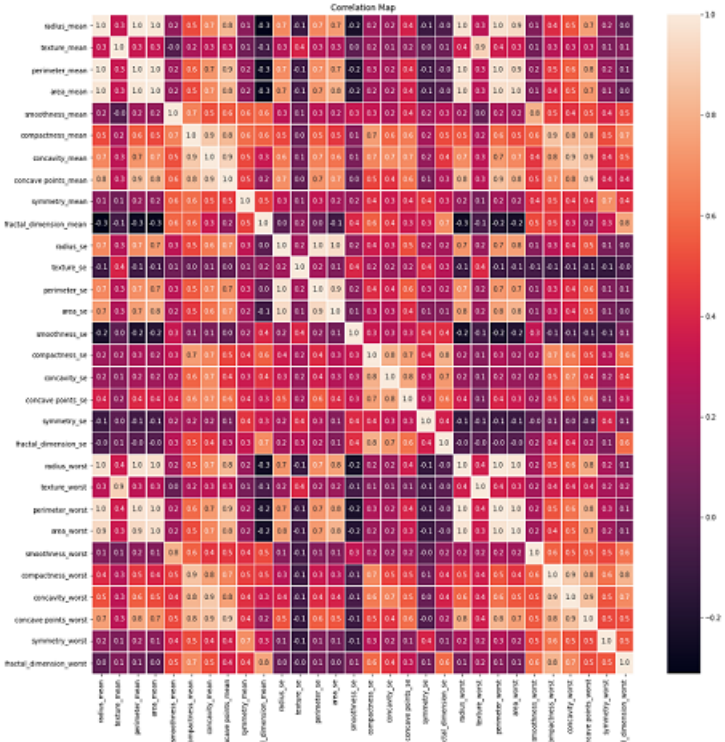
Task 4

As part of Task 4, we found out the five most important features for all the classifiers which are the biggest factors in our correct prediction of classifying Malignant and Benign tumours Also we removed certain columns like perimeter-mean, area-mean, concave-points-mean, perimeter-se, area-se, radius-worst, texture-worst, perimeter-worst, area-worst, concave-point-worst which were highly correlated and had a correlation of more than 0.9.

Random Forest		XGBoost	
radius_mean	0.26338	concavity_mean	0.43003258
concavity_mean	0.17249	radius_mean	0.1392054
concavity_worst	0.159644	concavity_worst	0.10569549
radius_se	0.070494	compactness_worst	0.06965829
compactness_mean	0.046035	smoothness_worst	0.042652305

Decision Tree		Logistic Regression	
concavity_mean	0.568569612	concavity_worst	3.284628259
radius_mean	0.176702041	compactness_worst	2.390223504
smoothness_worst	0.054995842	symmetry_worst	1.716005357
symmetry_se	0.047390151	radius_se	1.656064664
texture_mean	0.045084715	concavity_mean	0.910314102





Discussion

Breast cancer is the most common cancer amongst women in the world. It accounts for 25 percent of all cancer cases, and affected over 2.1 Million people in 2015 alone.Early diagnosis significantly increases the chances of survival. The key challenges against it's detection is how to classify tumors into malignant (cancerous) or benign(non cancerous).Machine Learning technique can dramatically improve the level of diagnosis in breast cancer. Research shows that experienced physicians can detect cancer by 79 percentage accuracy, while a 91 percentage (sometimes up to 97 percentage) accuracy can be achieved using Machine Learning techniques. To conclude, data scientists can play a vital role in Breast Cancer Analysis through designing strong and effective classifiers using proper features.

References

Code -

- https://www.kaggle.com/kanncaa1/statistical-learning-tutorial-for-beginners
- https://www.kaggle.com/kanncaa1/feature-selection-and-data-visualization
- https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e