Capstone 3: Breast Cancer

Del Wester

4/26/2021

Overview

Most types of breast cancer are easy to diagnose by microscopic analysis of a sample - or biopsy - of the affected area of the breast. The two most commonly used screening methods, physical examination of the breasts by a healthcare provider and mammography, can offer an approximate likelihood that a lump is cancer, and may also detect some other lesions, such as a simple cyst. When these examinations are inconclusive, a healthcare provider can remove a sample of the fluid in the lump for microscopic analysis (a procedure known as fine needle aspiration, or fine needle aspiration, FNA) to help establish the diagnosis. A needle aspiration can be performed in a healthcare provider's office or clinic. Together, physical examination of the breasts, mammography, and FNA can be used to diagnose breast cancer with a good degree of accuracy. The features for this dataset are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. I will use this dataset to determine which model has the highest Recall score, which is to say the model that finds the most True Positives.

Data Wrangling

I loaded the dataset into Excel and replaced the numbered columns with the column names. I then saved as a csv file and loaded into a pandas dataframe.

There are 699 records in this dataset.

Number of Attributes: 10 + output attribute

Attribute information: except for ID and Class, all columns had values ranging from 1 – 10.

Input variables

- 1 - ID
- 2 – Clump_Thickness
- 3 - Uniformity_of_Cell_Size
- 4 - Uniformity_of_Cell_Shape
- 5 - Marginal_Adhesion
- 6 - Single_Epithelial_Cell_Size
- 7 - Bare_Nuclei
- 8 - Bland_Chromatin
- 9 - Normal_Nucleoli
- 10 - Mitoses
- 11 - Class (2 for benign, 4 for malignant)
- Missing Attribute Values: Bare_Nuclei was missing 16 values which I replaced with the mean.

Exploratory Data Analysis

The Breast Cancer data includes mostly continuous data with a single categorical column. . Exploratory data analysis was used to derive relationships between the class and the various features available from the data profile. The Class feature of this set was determined by the image of a fine needle aspirate (FNA) of a breast mass and takes into consideration the characteristics of the cell nuclei present in the image.
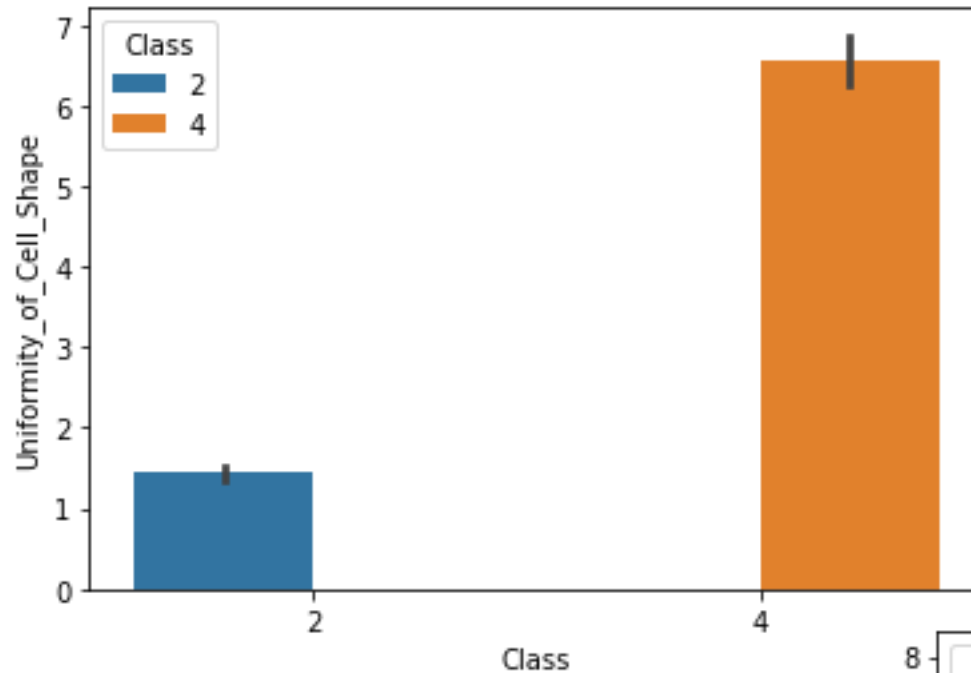
Class Distribution

Benign(2): 458   Malignant(4): 241
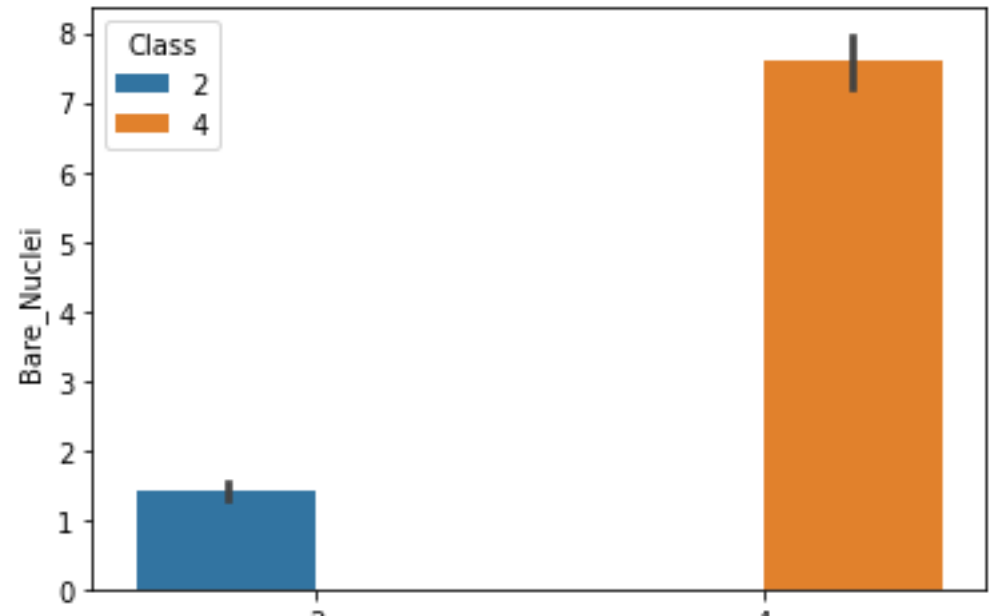

Class Count Plot
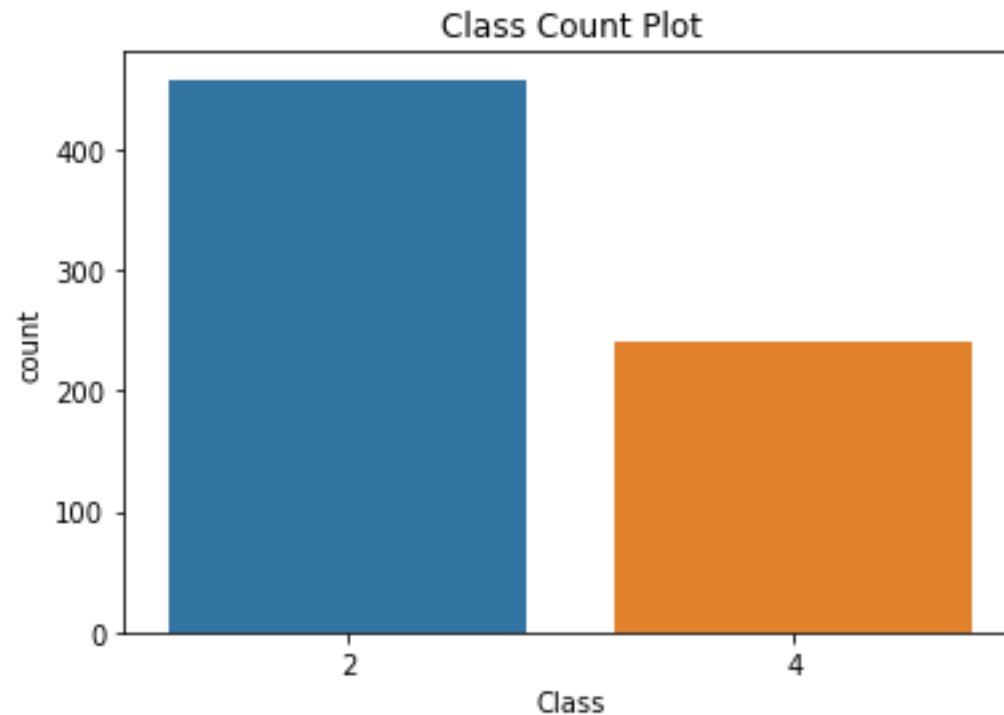
The plots for the output variable vs each of the features

Uniformity of Cell Shape
is much lower in malignant
cells (2-blue)

Bare Nuclei
Is also much lower in
malignant cells.

Correlation Matrix

Multicollinearity is a problem as it undermines the significance of independent variables. It can be fixed by removing the highly correlated features from the model.

Uniformity of Cell Size and Mitosis were both removed before modeling for this reason.
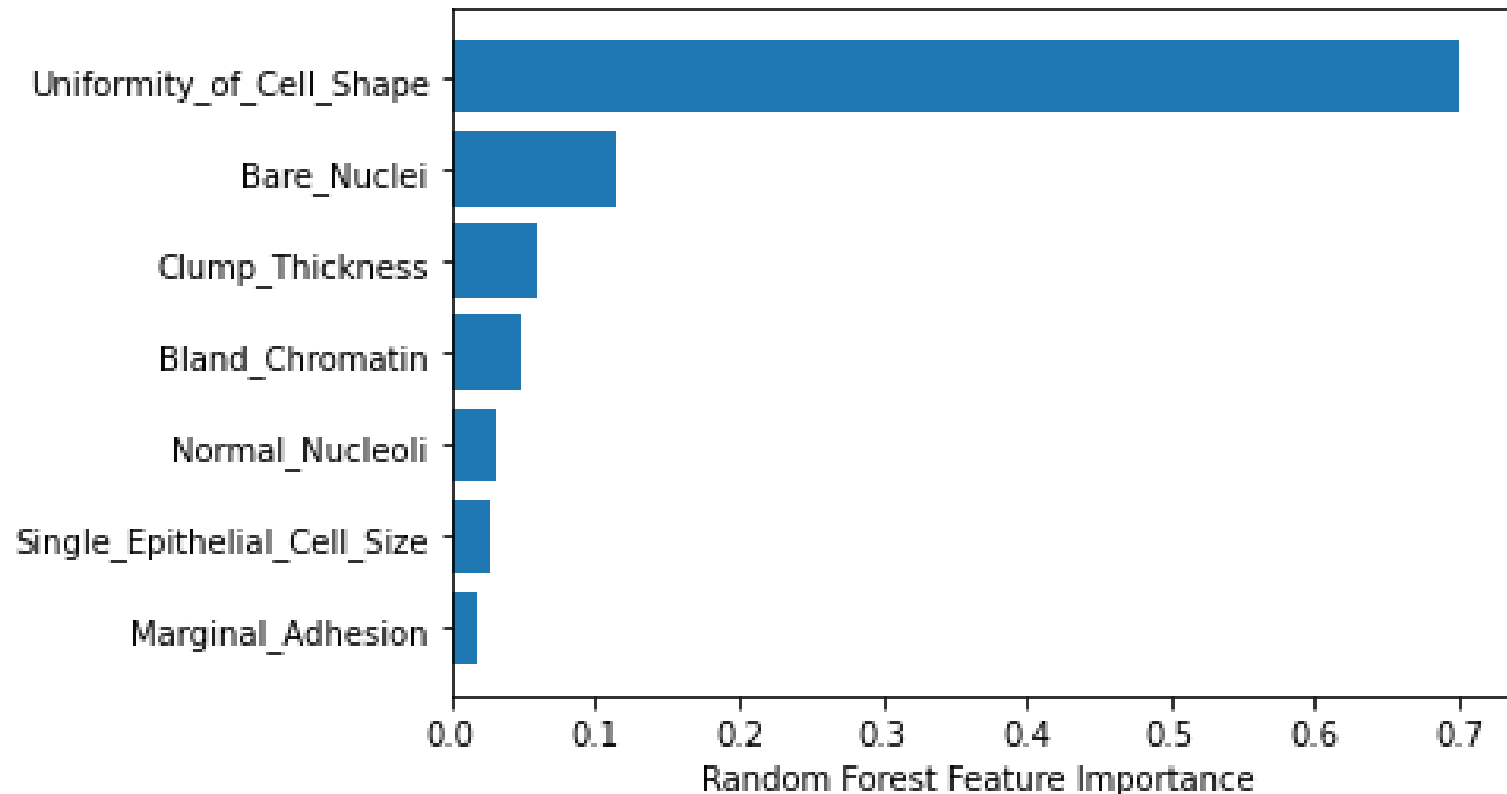

Feature Correlation
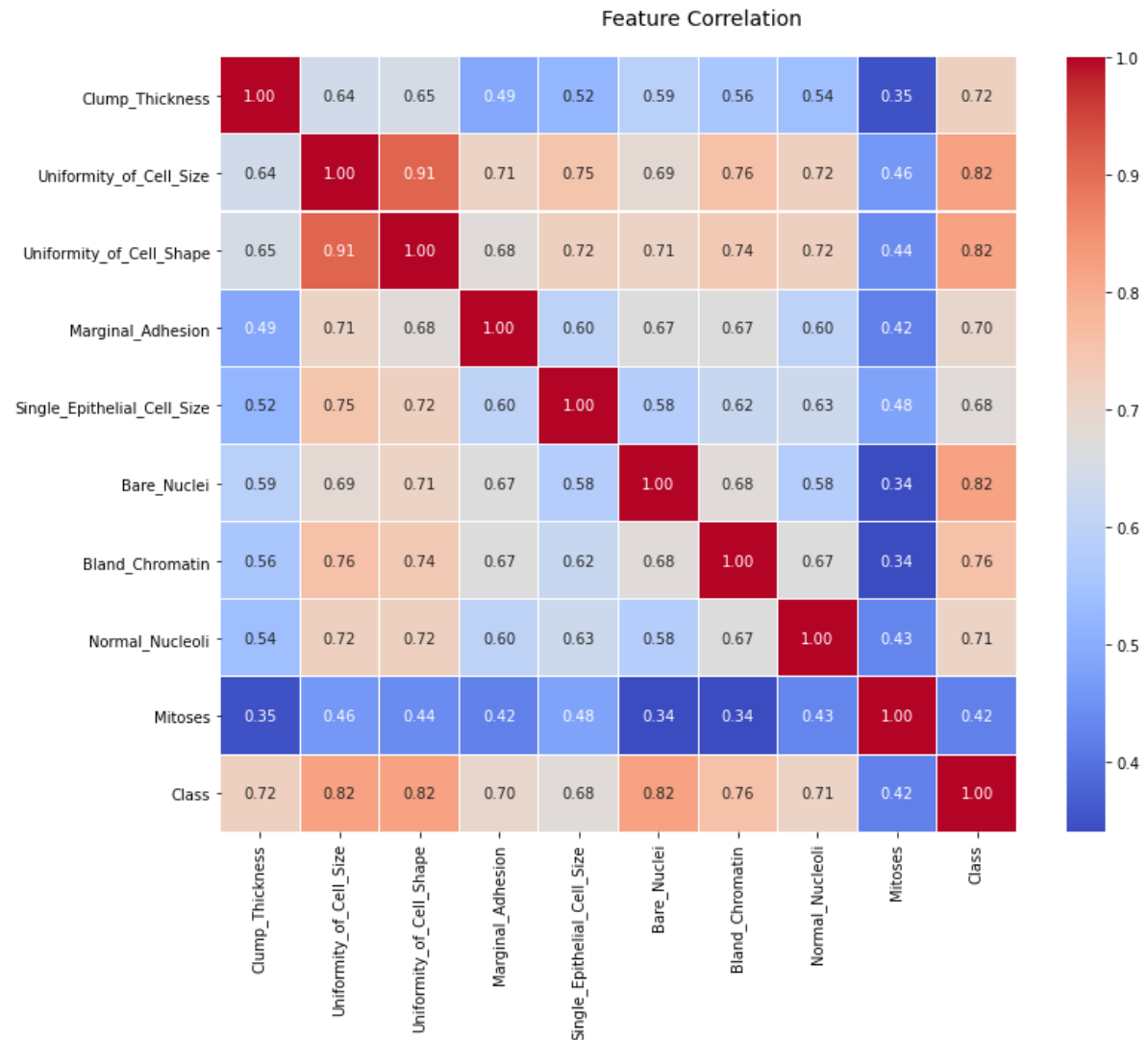
New Correlation Matrix

Clear high correlation between the target variable (Class) and Uniformity of Cell Shape and Bare Nuclei.

Feature Correlation
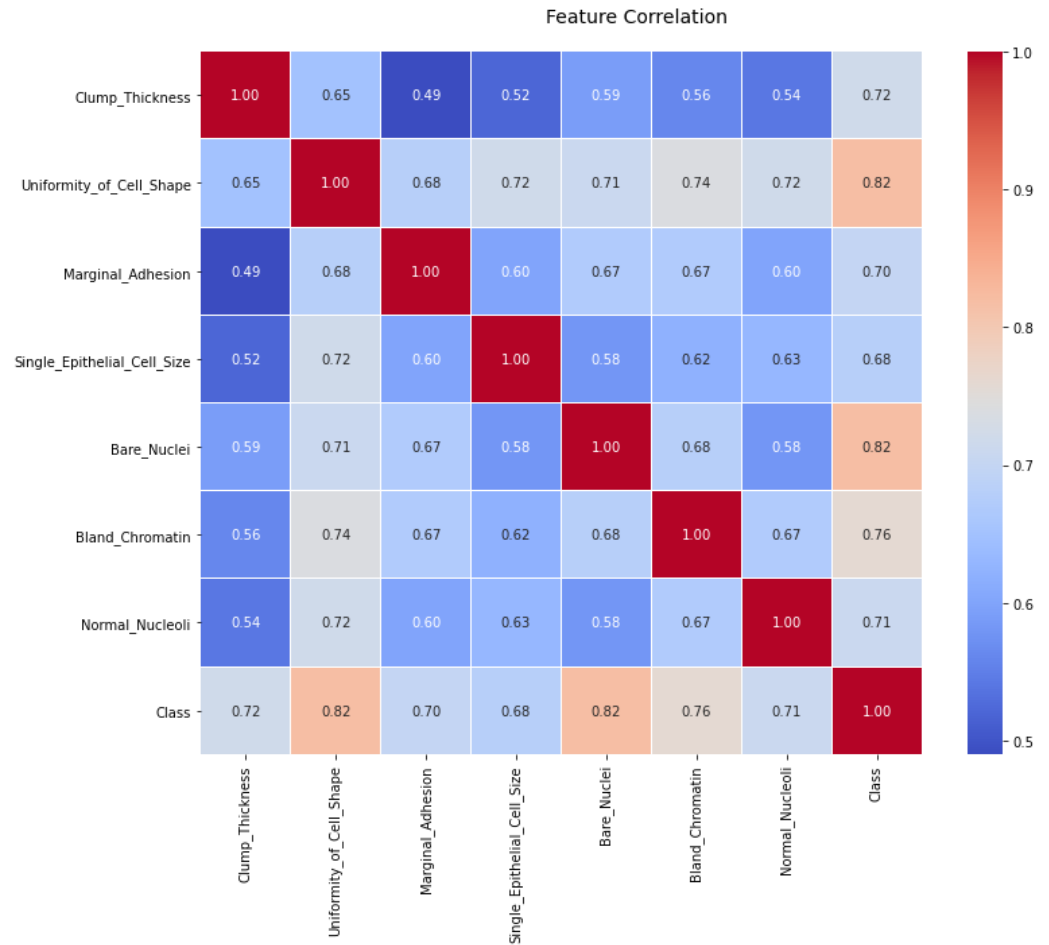
Preprocessing and Modeling

The data was scaled using a Robust Scaler and I ran a Random Forest Regressor to find the Feature Importance which is displayed below.

Modeling

I tested and scored the following models: RandomForestClassifier, KNeighborsClassifier, and Logistic Regression. I used GridsearchCV hyper parameter tuning for the models as well. I then ran the ROC_AUC scores/curve and Classification Reports. Logistic Regression scored higher than the others and was what I recommend for the final model.

With this model, we want a high TPR considering you don't want to tell someone they don't have cancer when they do.
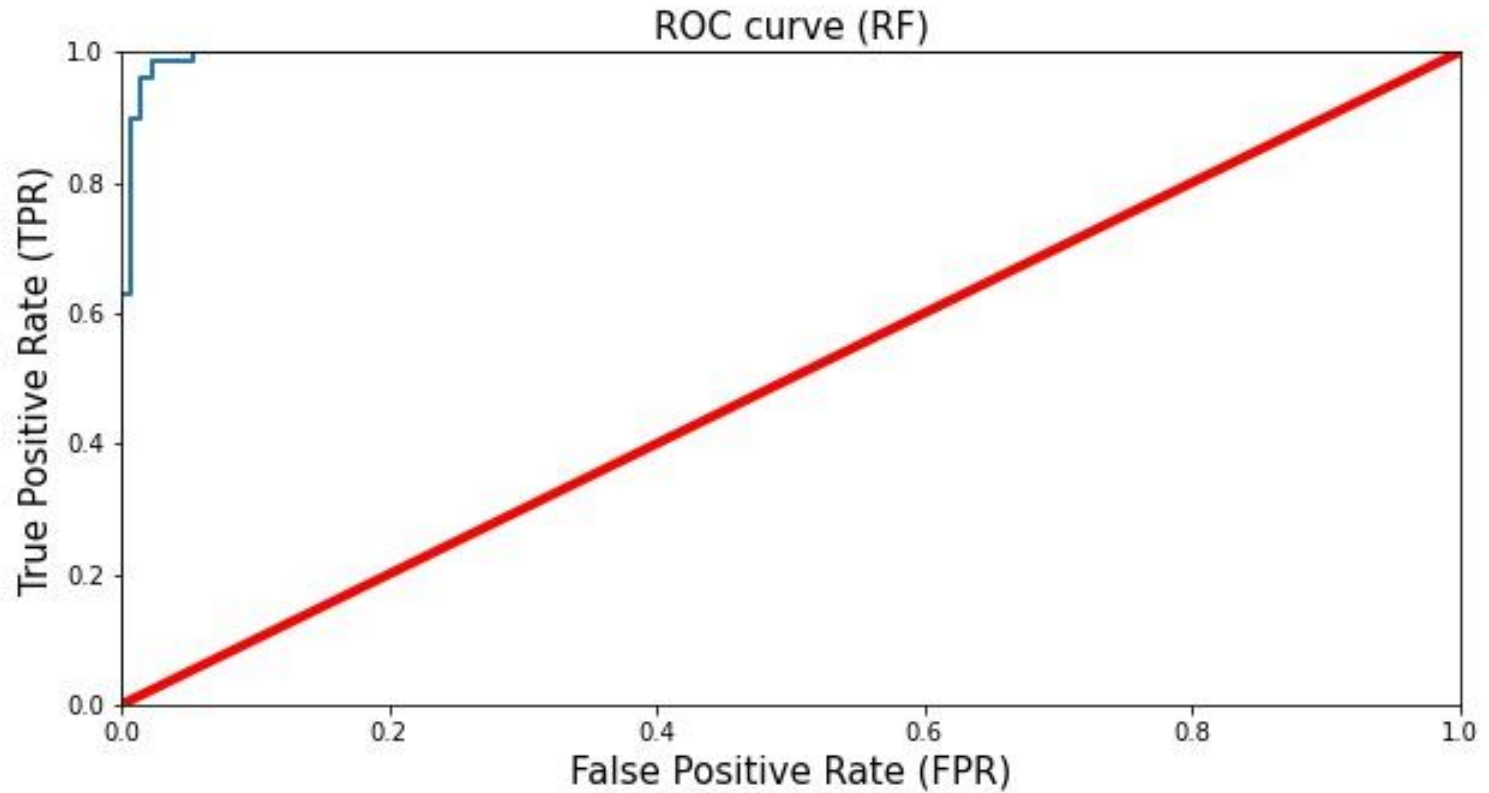
Models

With GridsearchCV hyper parameter Tuning

Logistic Regression	99.6
C = 1, solver = 'liblinear'	
Random Forest Classifier	99.4
max_depth = 4, max_features = 'sqrt', n_estimators = 17	
KNeighbors Classifier	99.3
Algorithm = 'ball_tree', n_neighbors = 100, weights = 'distance'	

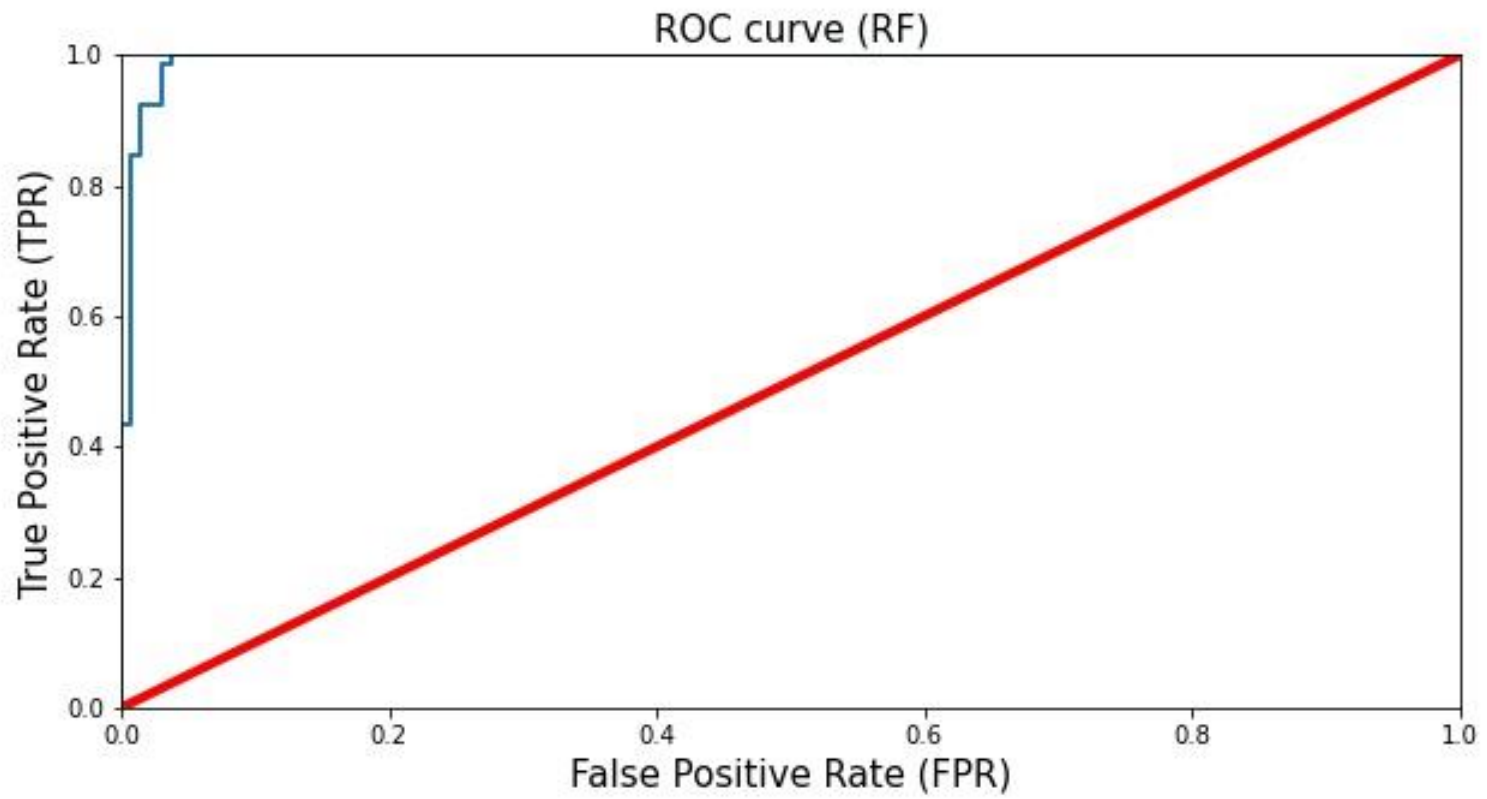# ROC-AUC Scores/Curve

Logistic Regression

ROC curve (RF)

Random Forest Classifier

KNeighbors Classifier

ROC curve (RF)
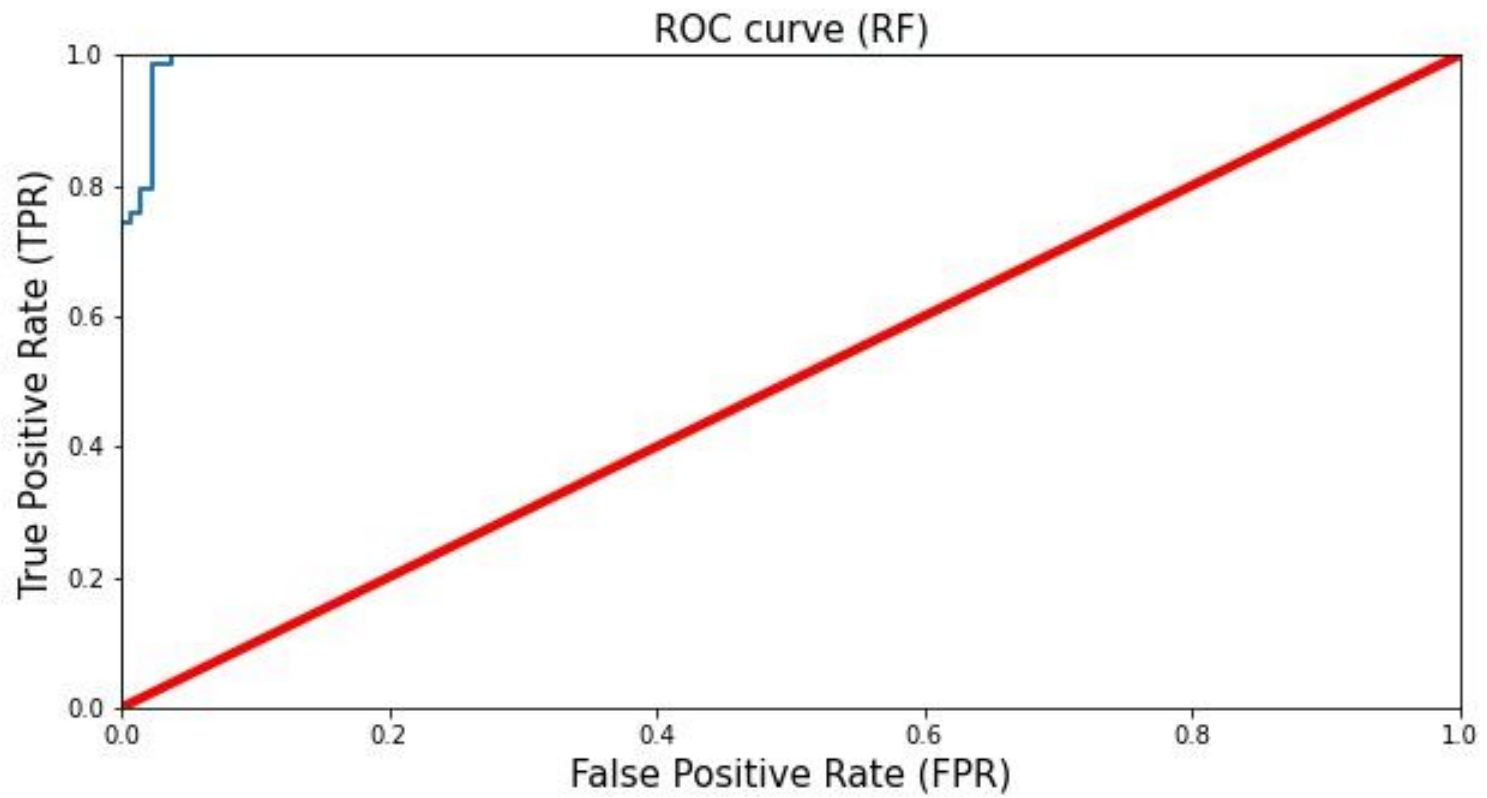
Classification Reports

Logistic Regression

Class	precision	recall	f1-score	support
2	0.99	0.98	0.98	132
4	0.96	0.99	0.97	78
accuracy			0.98	210
macro avg	0.98	0.98	0.98	210
weighted avg	0.98	0.98	0.98	210

Random Forest Classifier

Class	precision	recall	f1-score	support
2	0.99	0.96	0.98	132
4	0.94	0.99	0.96	78
accuracy			0.97	210
macro avg	0.97	0.97	0.97	210
weighted avg	0.97	0.97	0.97	210

KNeighbors Classifier

Class	precision	recall	f1-score	support
2	0.97	0.97	0.97	132
4	0.95	0.95	0.95	78
accuracy			0.96	210
macro avg	0.96	0.96	0.96	210
weighted avg	0.96	0.96	0.96	210

Conclusion

There are different techniques that can be used for the prediction of breast cancer. I've analyzed breast cancer data using three classification techniques to predict the type of cancer and compared the results. They indicate that Logistic Regression is the best classifier with this dataset, followed by Random Forest and KNeighbors, with all three models near perfect accuracy. Fine needle aspirate (FNA) appears to be a very practical methodology for determining breast cancer malignancy with extremely high accuracy.

Further studies could be conducted to improve performance of these classification techniques by identifying more features that could be used in the analysis and/or using a threshold that sacrifices accuracy to increase Recall to 100%.