Capstone 3: Project Report

Del Wester

3/16/2021

Breast Cancer

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 and cytology, FNAC) 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 FNAC 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. This dataset was obtained from the University of Wisconsin Hospitals, Madison from https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic).

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

Output variable:

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

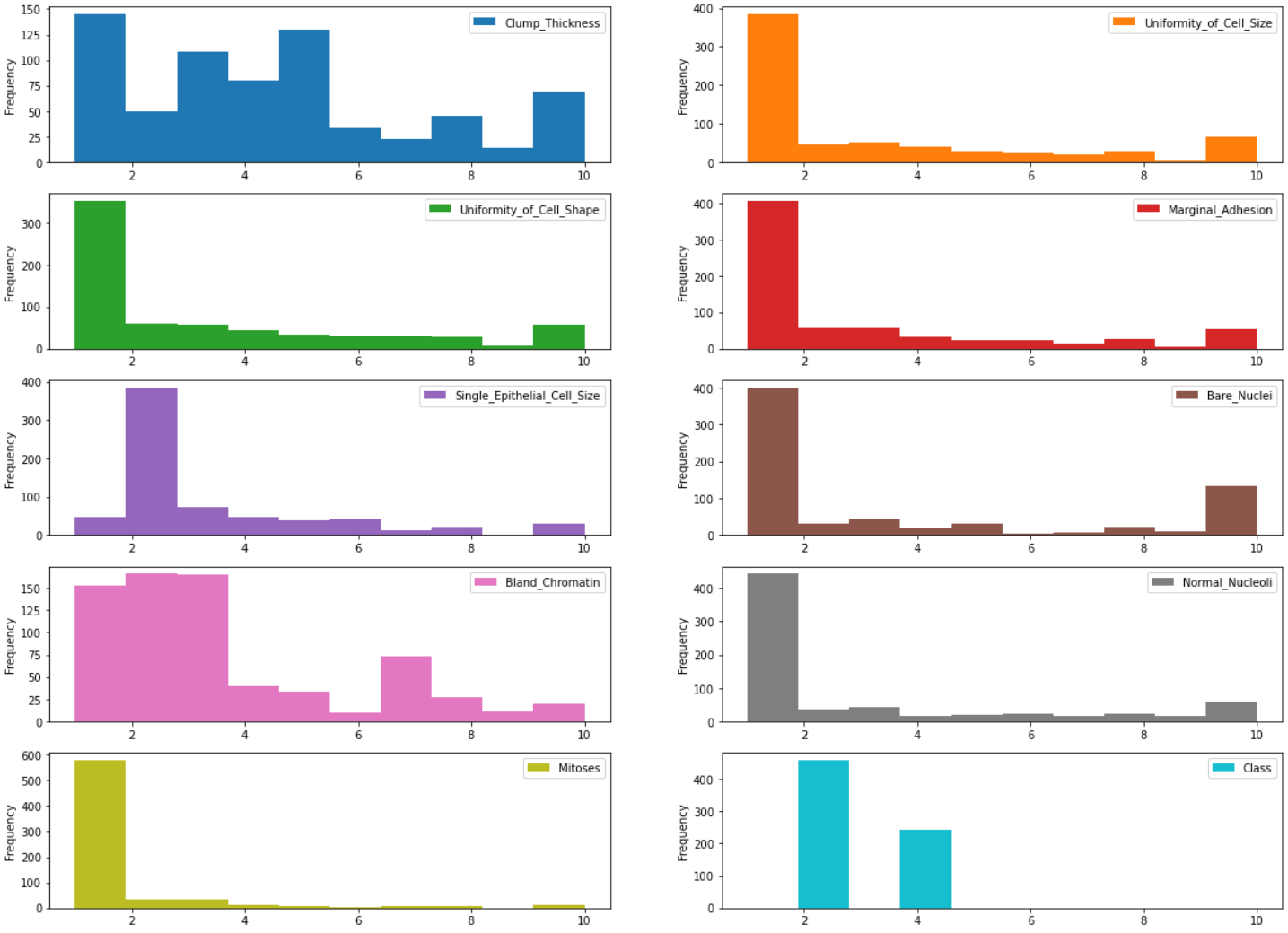
The Breast Cancer data includes mostly continuous data with a single categorical column. . Exploratory data analysis was used to derive relationships between the class and the various features available from the data profile. The Class feature of this set was determined by the image of a fine needle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

"Class" Distribution:

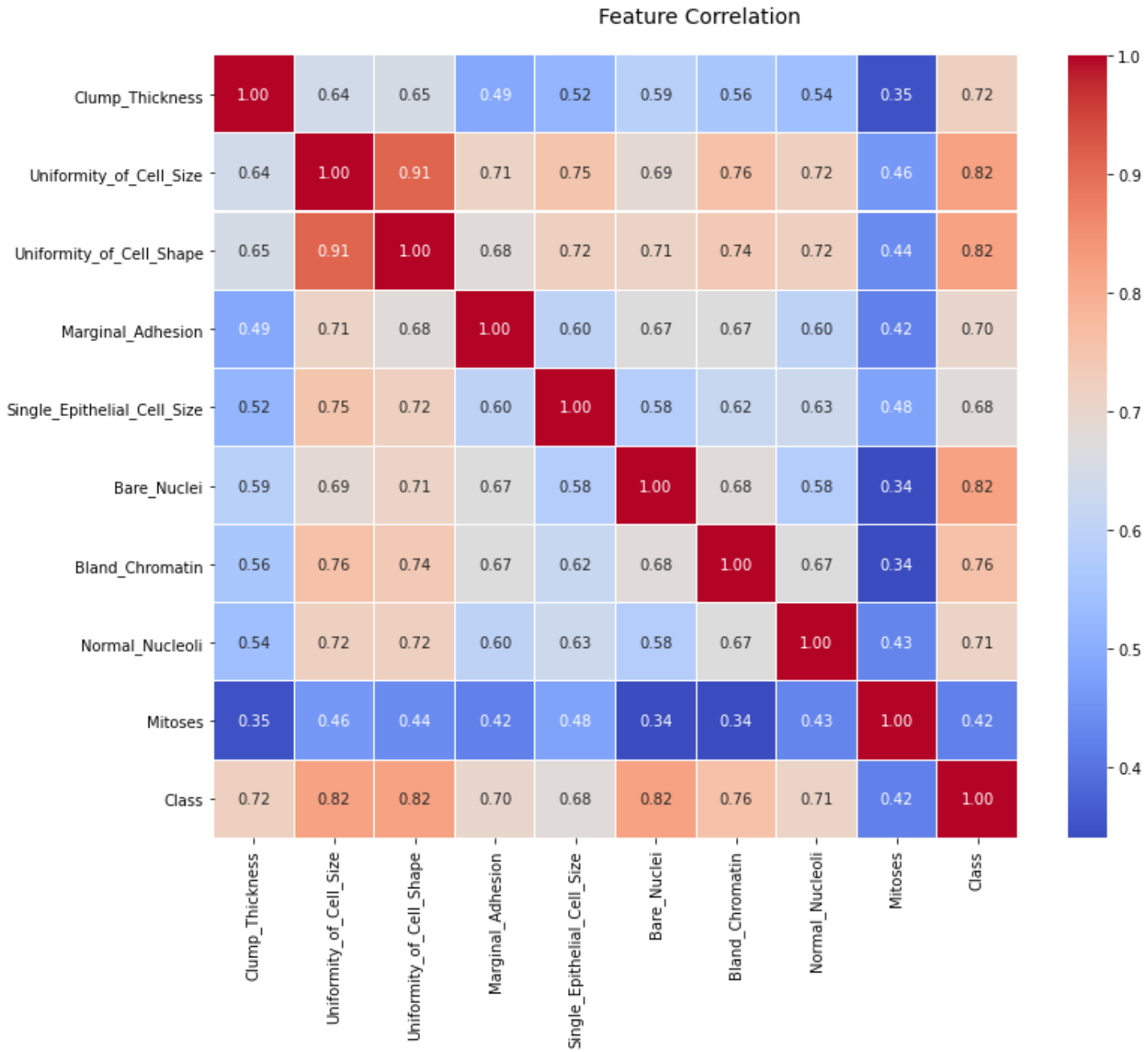
Benign: 458

Malignant: 241

The plots for the output variable vs each of the features are below.



Correlation Matrix

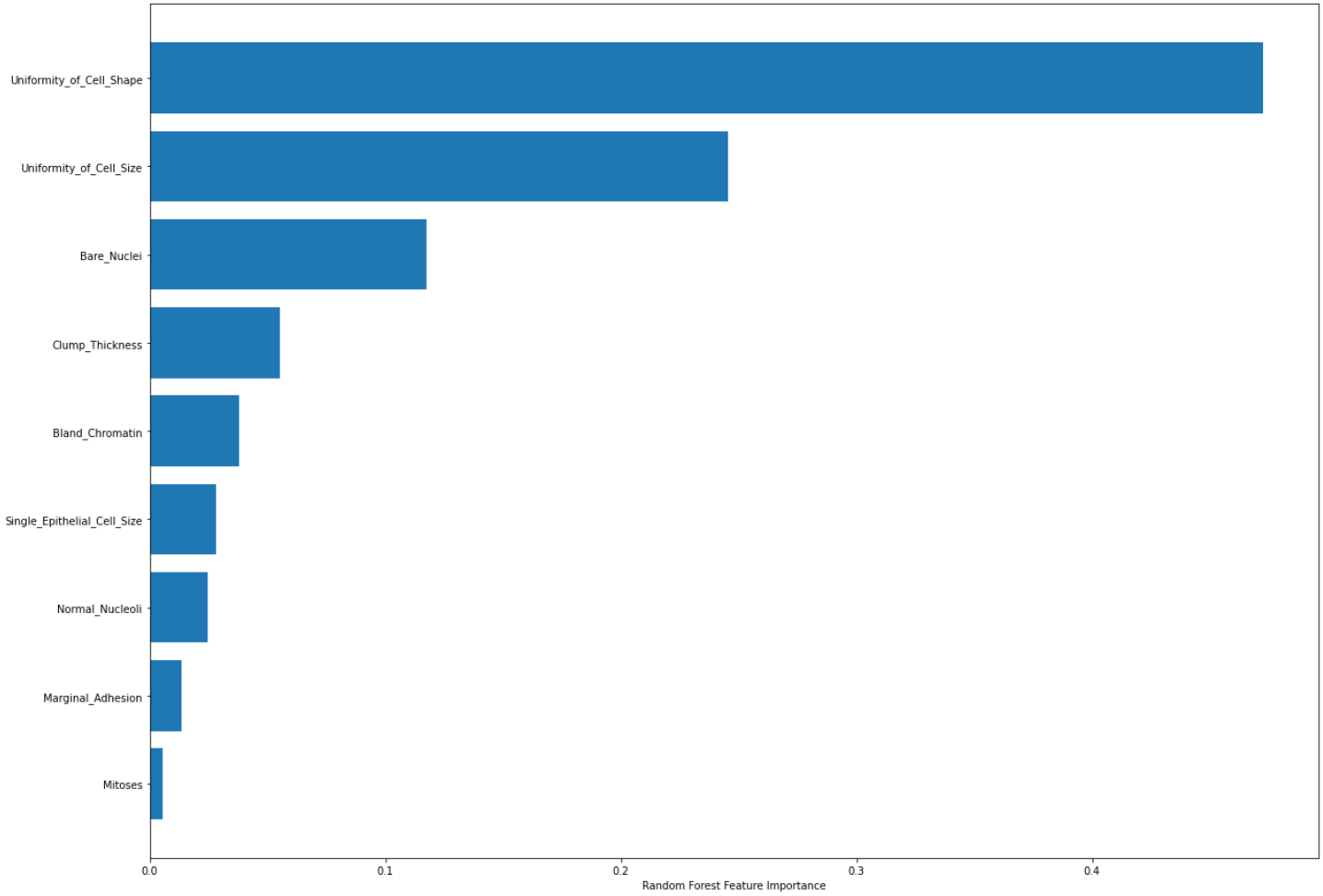


Comment here

Preprocessing and Training

In this section, the data was scaled using a Robust Scaler and I ran a Random Forest Regressor to find the Feature Importance which is displayed below. I didn't drop any features.

Top Features - Uniformity_of_Cell_Shape, Uniformity_of_Cell_Size, and Bare_Nuclei.



Modeling

I tested and scored the following models: DecisionTreeClassifier, RandomForestClassifier, KNeighborsClassifier, GradientBoostingClassifier, GaussianNB and SVC. The KNeighbors Classifier scored higher than the others and was what I went with for the final model.

Decision Tree Classifier
Score: 95.7

Random Forest Classifier
Score: 97.6

KNeighbors Classifier
Score: 98.1

Gradient Boosting Classifier
Score: 96.2

GaussianNB Classifier
Score: 97.6

SVC Classifier
Score: 97.1

I used Gridsearch hyperparameter tuning for the KNeighbors model as well as determined the ROC_AUC scores and ROC curve.

dict_keys(['algorithm', 'leaf_size', 'metric', 'metric_params', 'n_jobs', 'n_neighbors', 'p', 'weights'])

{'n_neighbors': [10, 20, 50, 100, 150], 'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'], 'weights': ['uniform', 'distance']}

.best_estimator_: n_neighbors=150, weights='distance'

.best_score_: 98.7

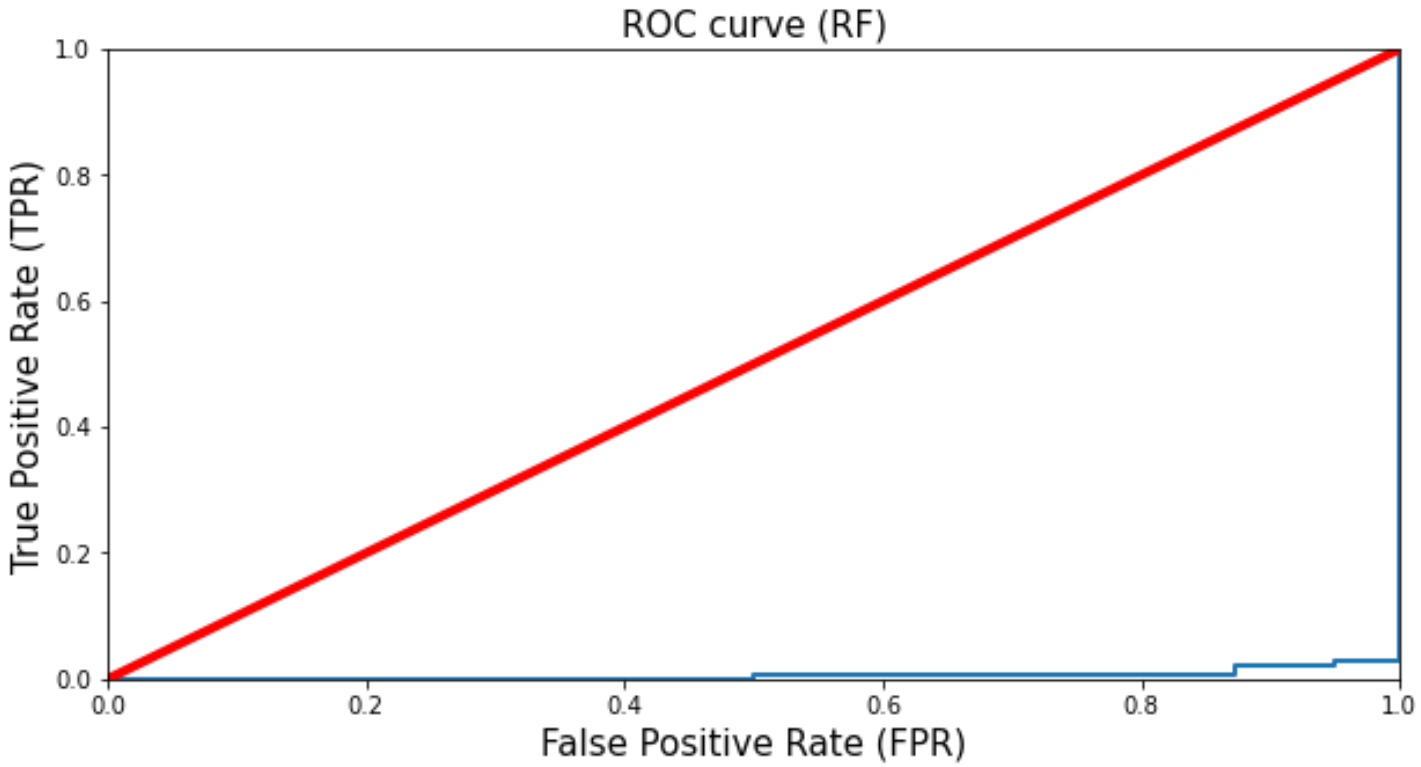
I then scored the KNeighbors model with the new hyperparameters and ran the ROC AUC scores and curve.

Accuracy Score: 94.3

CV Score: 99.0

ROC-AUC Score: 99.4

A very low FPR is good considering you don’t want to tell someone they have cancer when they don’t.



Classification Report

Class	precision	recall	f1-score	support
Benign	0.92	0.99	0.96	132
Malignant	0.99	0.86	0.92	78
Accuracy			0.94	210
Macro avg	0.95	0.93	0.94	210
Weighted	0.95	0.94	0.94	210

Sources & Citations:

https://en.wikipedia.org/wiki/Breast_cancer#Diagnosis