BREAST CANCER PREDICTION

SNEHA KUMARAN

GEORGE MASON UNIVERSITY

INTRODUCTION

Breast Cancer occurs more often in women in general. It continues to be a major problem as many people tend to suffer from this cancerous disease. When the cells grow out of control, the cancer can begin or occur anywhere that is part of the breast. Most breast cancers happen in the duct or lobules of the breast. Not only is this cancer physically threatening for women, but it is psychologically threatening as well. Some women can get tumors that are noninvasive which will still mentally affect them. But after all the tests, people get to know it's not cancerous (benign). So, breast cancer can be divided into two types of tumors, which are benign and malignant. Some people can get misdiagnosed, and this mentally can affect them as well. A mass in the breast itself will classify some sort of warning in your body, so using the dataset from UCI Machine Learning, the Breast Cancer Wisconsin (Diagnostic) Data Set will provide a list of patients that have features of the tumor. Many of the machine learning algorithms have helped identify a set of patients whether they each have a certain type of tumor.   A classification method will be used to predict whether each patient has a benign tumor or a malignant tumor, then using the number of certain type of tumors it can help identify the answer for the hypothesis. To see how well the ML algorithm works, the results will be shown through the accuracy of the matrix.

OBJECTIVES

Overall Objective:

The overall objective is using the dataset, which has a set of 569 patients. In the dataset there are a total of 32 columns. There are 10 real-valued features that are computed and used in the Logistic Regression (Classification) algorithm to find the diagnosis. The attributes are the features computed from a digitized image of fine needle aspirate of a breast mass. They tend to describe the characteristics of the cell nuclei that present in the image. The attributes include of Patient ID, radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness (perimeter^2 / area - 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, and fractal dimension ("coastline approximation" - 1). Using these main attributes, the goal was to find the diagnosis of each patient.

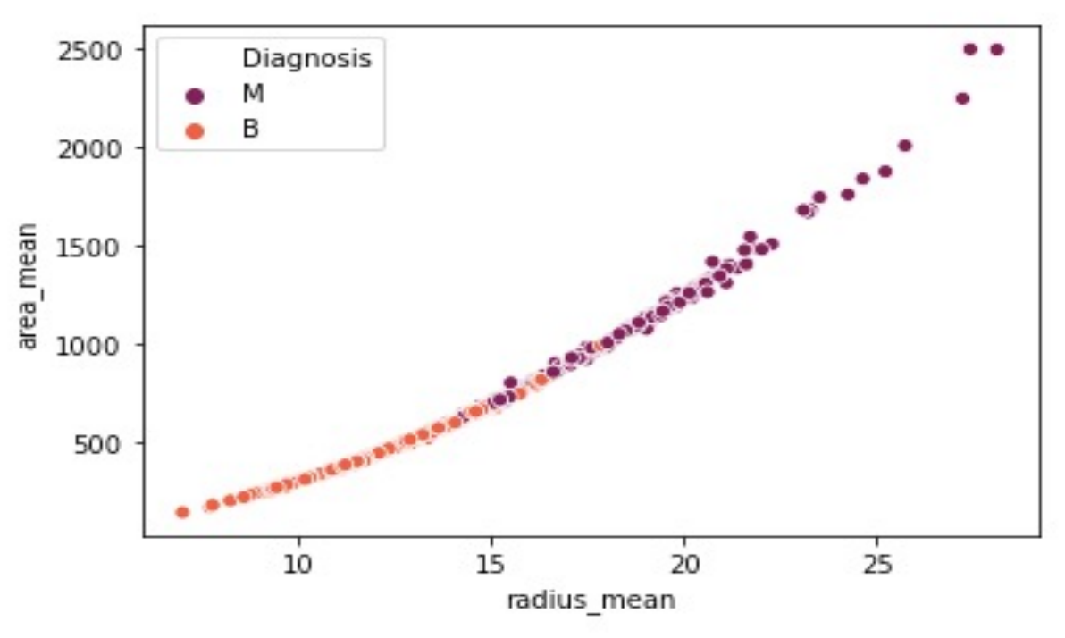
Hypothesis:

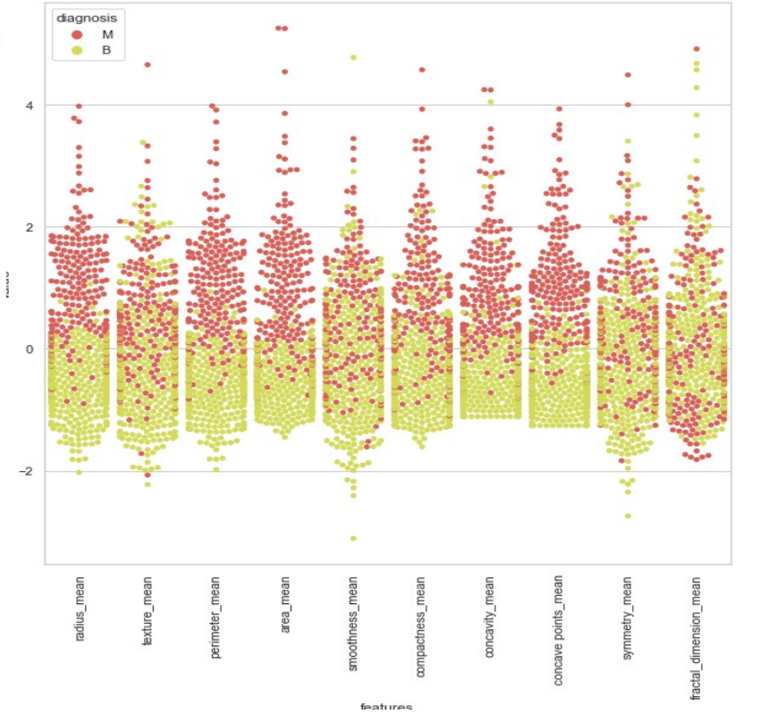
If the radius of the mass is longer than 19 cm, then the tumor is malignant.

After finding the diagnosis, the determination of the tumor will help identify the answer to the hypothesis. The hypothesis for this project after looking into the data has a lot of factors/features based on each patient's tumor. There was genuine curiosity to find out if the size or how long the mass is having to do anything with it being a cancerous tumor or noncancerous tumor. The average of the radius mean in the dataset was 19 cm, so beyond 19 cm would determine would be a malignant tumor and below 19 cm would be a benign tumor.

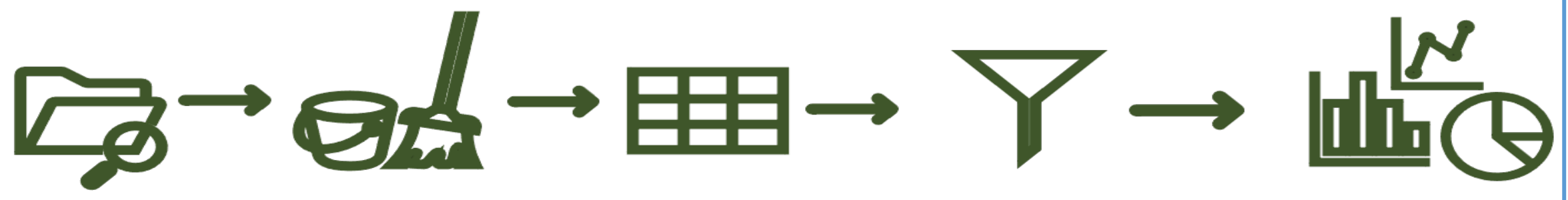
Classification ML Algorithm for the Dataset and Hypothesis:

Once the algorithm is applied, the analysis between the attributes will be depicted to show whether the algorithm worked. There will be bar plots, scatter plots, accuracy matrix, to determine how the radius and area attributes are strong enough and present the answer to the hypothesis. The visuals created will be shown for the results.



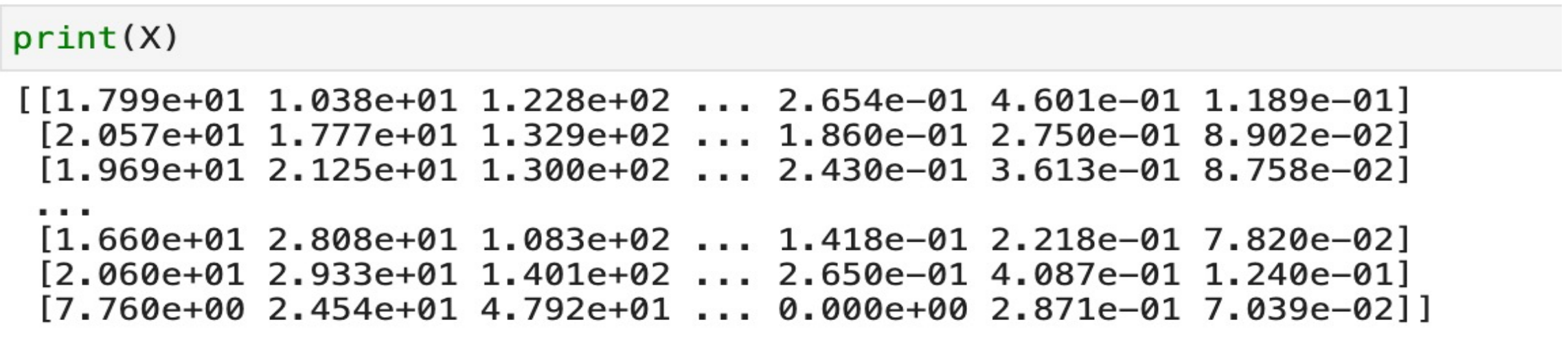


MATERIALS & METHOD



This is a preprocessing flow diagram, where the first icon demonstrates finding the data information. The second icon is about cleaning up the data. This is how the LabelEncoder in python was used to sort and clean up the data. The third icon is putting the cleaned data into a CSV file and importing the CSV file into Jupyter Notebook. The csv file was used with a DataFrame and Pandas Module. The fourth icon is about sorting the data to pull out the information we need to analyze. Finally, the fifth icon is taking the columns and data that was chosen to make visuals and graphs/plots.

After picking the data and finalizing the decision of using the breast cancer prediction dataset, the process of scanning the dataset to know whether there are any missing values is a crucial aspect of understanding the data and getting accurate results.

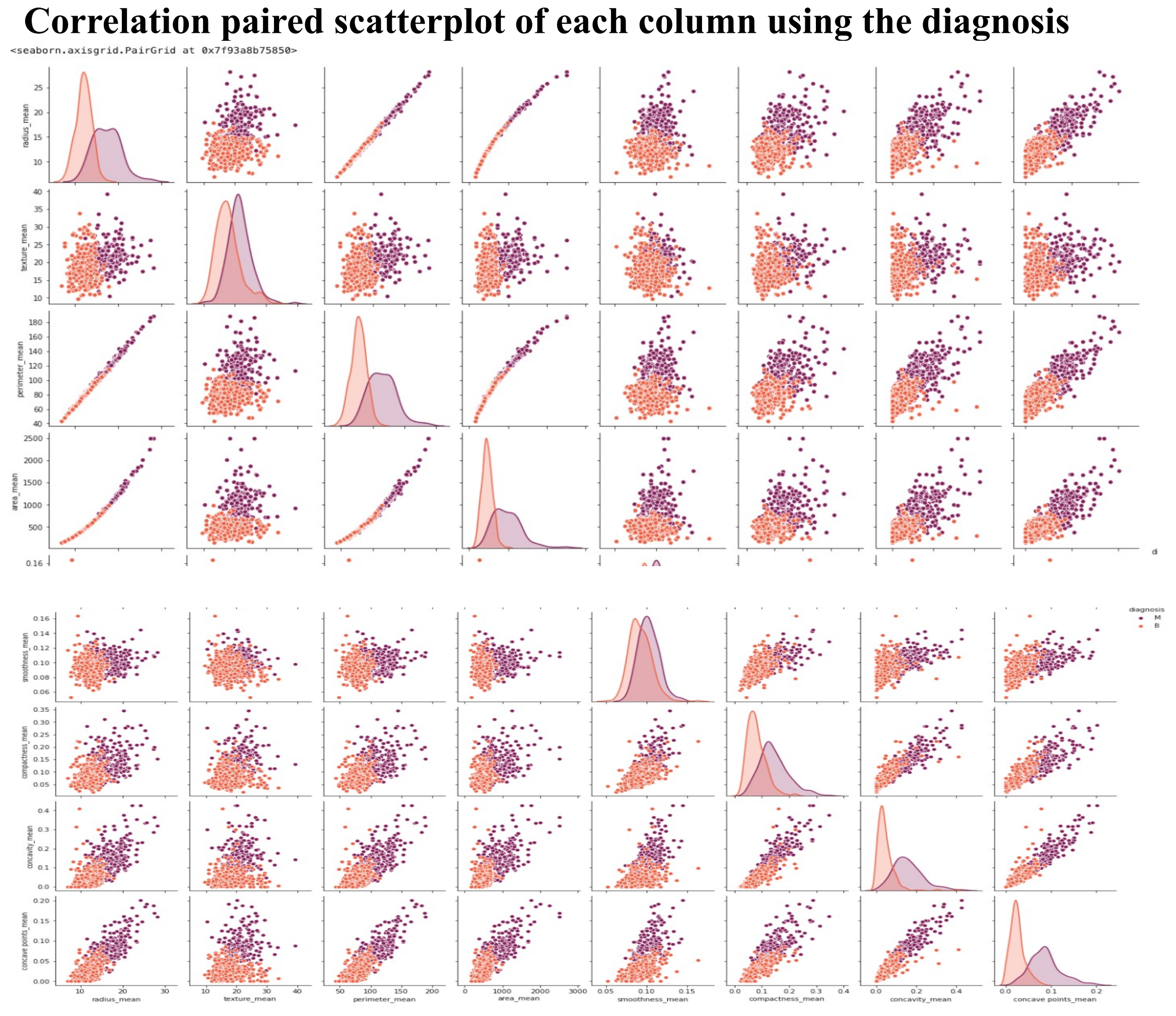


```
print(X)
```

```
[[1.799e+01 1.038e+01 1.228e+02 ... 2.654e-01 4.601e-01 1.189e-01]
 [2.057e+01 1.777e+01 1.329e+02 ... 1.860e-01 2.750e-01 8.902e-02]
 [1.969e+01 2.125e+01 1.300e+02 ... 2.430e-01 3.613e-01 8.758e-02]
 ...
 [1.660e+01 2.808e+01 1.083e+02 ... 1.418e-01 2.218e-01 7.820e-02]
 [2.060e+01 2.933e+01 1.401e+02 ... 2.650e-01 4.087e-01 1.240e-01]
 [7.760e+00 2.454e+01 4.792e+01 ... 0.000e+00 2.871e-01 7.039e-02]]
```

After applying the Simple Imputer, the Simple Imputer fills in the missing values if there is a missing value in the column. It will assign the strategy to equal to mean, then it will replace missing values using the mean along each column. This is not random, this is computation, based on the total mean per column it will fill in if anything is missing. Luckily, no values were missing in my dataset, this is just in case anything is missing. LabelEncoder will help code the tumors based on Malignant being 1 and Benign being 0. This makes it easier to classify as this is all preprocessing techniques. The splitting of the training set and test set code was split to 80% (which means test size = 0.2) of the data hoping for a better accuracy rate. Importing the sklearn.model_selection for using the train/test split module will help split the x's and y's into proper and designated sets. Then importing the Standard scalar module will compute all the missing values. The standardized features will rescale the features that have properties of a standard normal distribution within a mean of zero and a standard deviation of one. Therefore, this helped break down the training set of X's and the testing set of the X's.

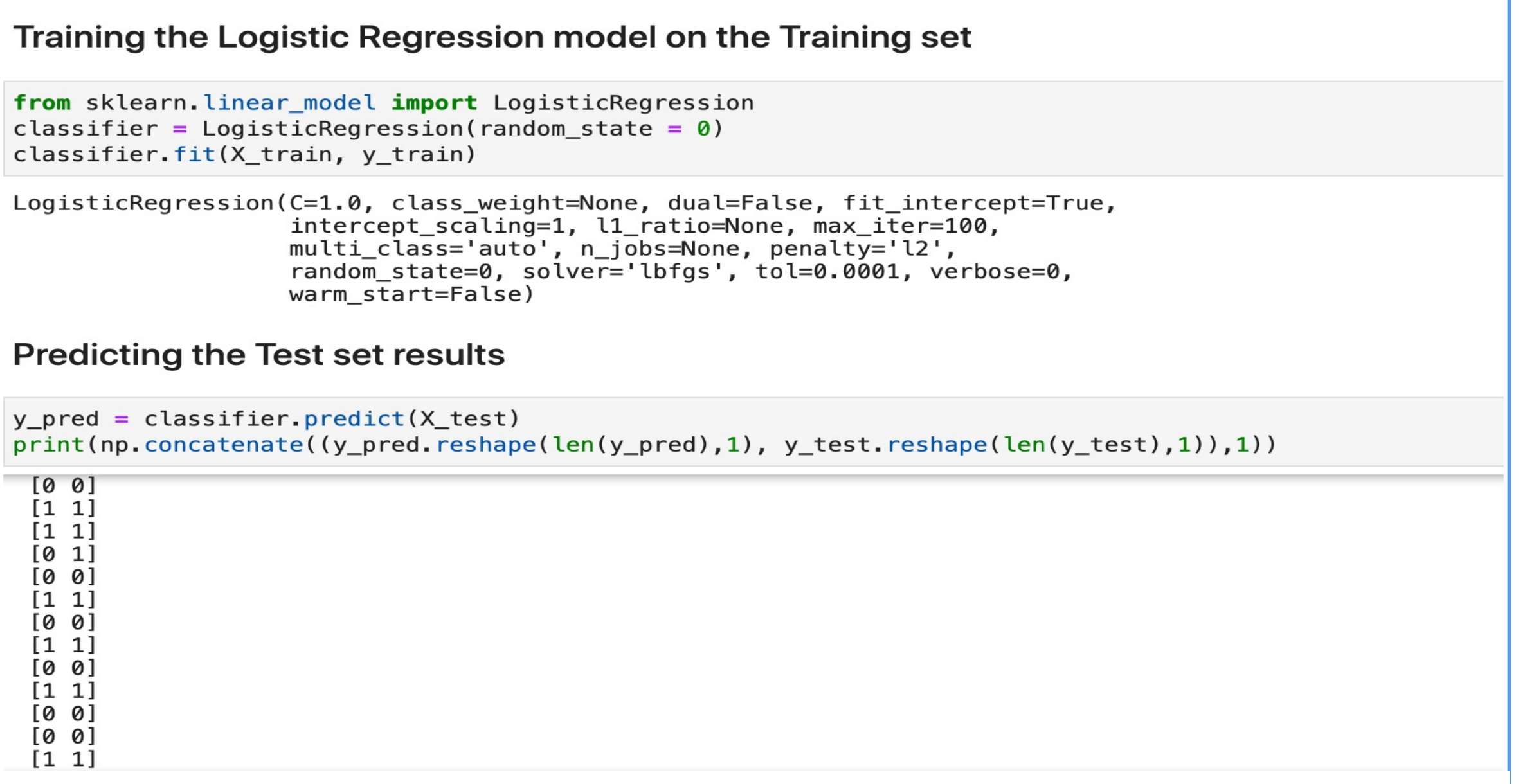
Correlation paired scatterplot of each column using the diagnosis



```
<seaborn.axisgrid.PairGrid at 0x7f93a8b75850>
```

RESULTS

Importing the logistic regression module and setting the random state to 0 (as the seed), will help split the data to be a constant integer. Then I made it fit into the logistic regression. I wanted to use the testing set to predict the results and set this into my y_pred for the results.



Training the Logistic Regression model on the Training set

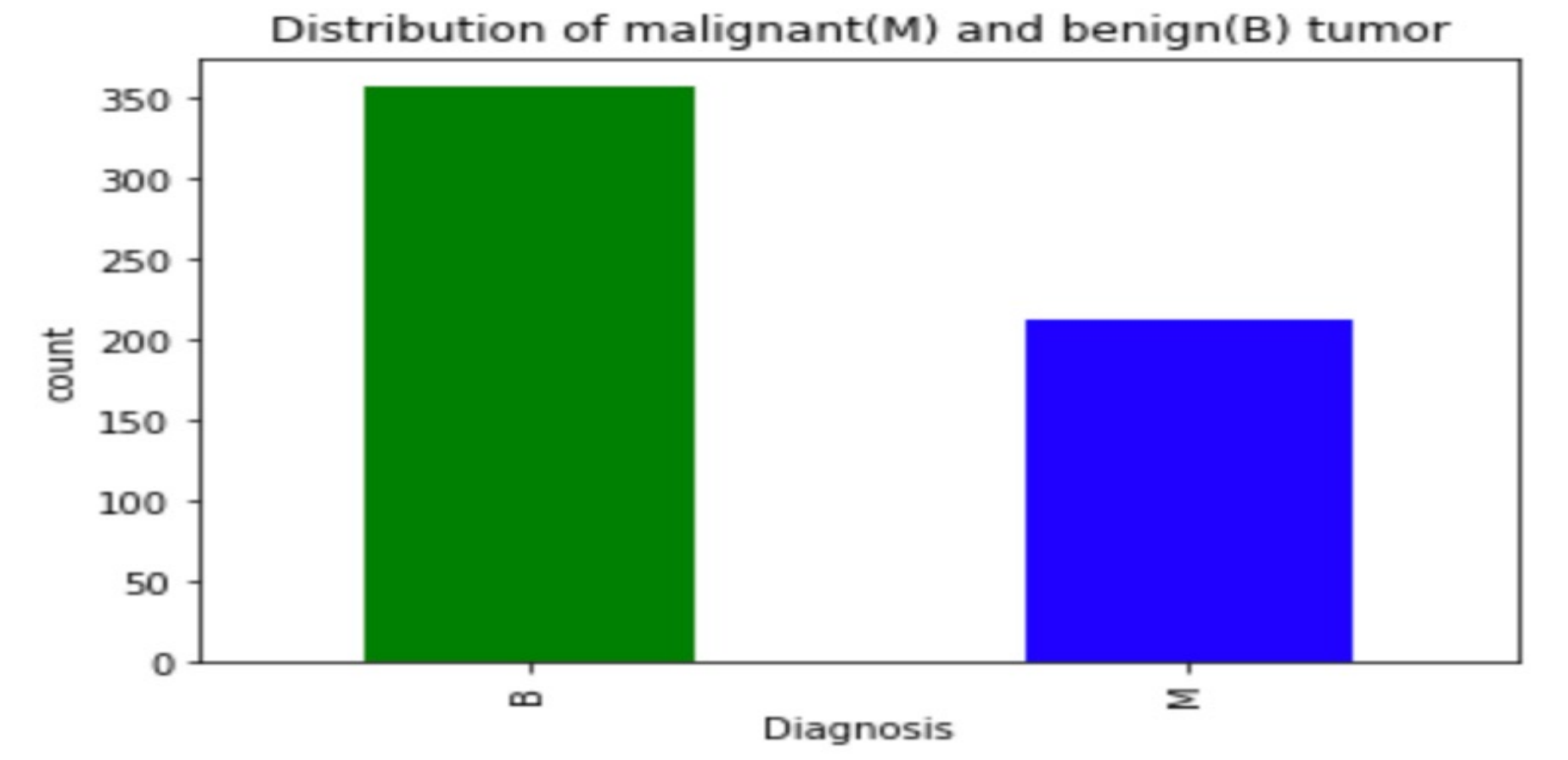
```
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression(random_state = 0)
classifier.fit(X_train, y_train)
```

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=0, solver='lbfgs', tol=0.0001, verbose=0,
                   warm_start=False)
```

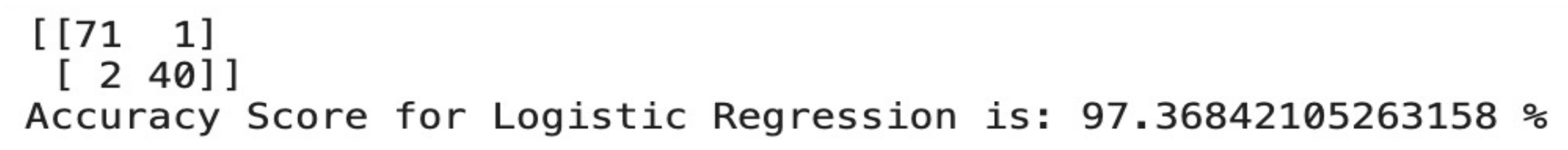
Predicting the Test set results

```
y_pred = classifier.predict(X_test)
print(np.concatenate((y_pred.reshape(len(y_pred),1), y_test.reshape(len(y_test),1)),1))
```

```
 [0 0]
 [1 1]
 [1 1]
 [0 1]
 [0 0]
 [1 1]
 [0 0]
 [1 1]
 [0 0]
 [1 1]
 [0 0]
 [0 0]
 [1 1]
```



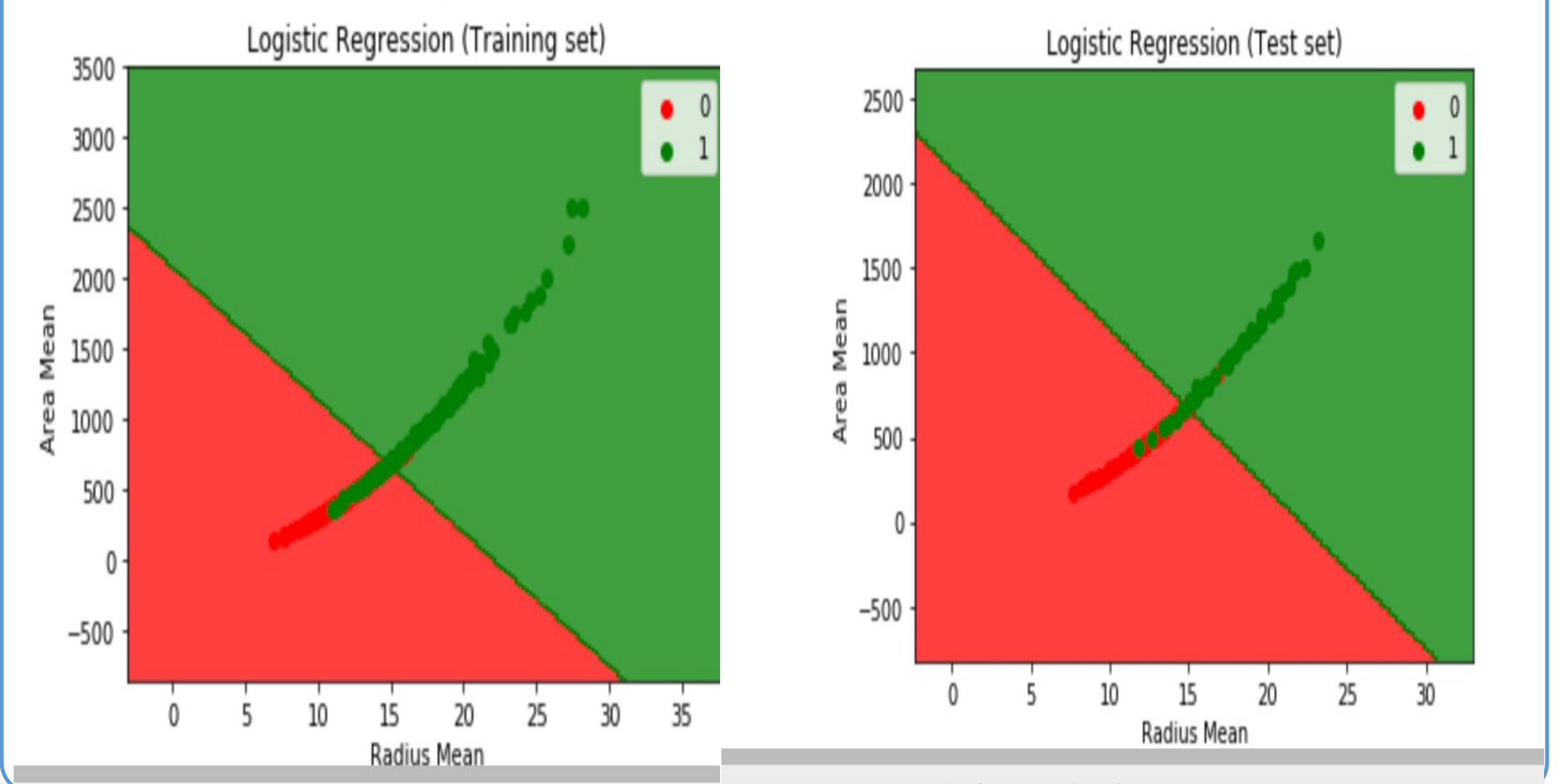
The accuracy score of the logistic regression of the overall prediction is 97.37% . As there are multiple independent variables this is the most efficient way to show the result. The ML algorithm is (Logistic Regression) classification, the accuracy is high and proves that the algorithm predicted well.



```
[[71  1]
 [ 2 40]]
Accuracy Score for Logistic Regression is: 97.36842105263158 %
```

Hypothesis ML Results

Two independent variables (Radius Mean and Area Mean) were chosen to test if the radius of the tumor is longer than 19 cm, then the tumor is malignant. 75% of the data was used for the training set and 25% of the data was used for the test set. These are the ML prediction results for the Logistic regression algorithm. The first scatter plot has benign tumors that are below of 11 radius mean and the second scatter plot is below 13 radius mean. The first scatter plot has most of the malignant tumors above 14 radius mean, and the second scatter plot is above 14 radius mean. So, the size of the tumor matters on whether the tumor is malignant or benign. The cut off line of the benign tumor is 17.5, that is where majority of the malignant tumors start occurring in general.



CONCLUSION

In conclusion, "breast cancer is one of the most common and leading causes of cancer among women" (Ravi ,Sivasangari, and Yarabarla 1). Using machine learning algorithm, it will "train the machines to learn and perform by itself without any explicit program or instruction." (Ravi ,Sivasangari, and Yarabarla). Logistic regression will classify the tumors that each patient has based on dataset features. As the algorithm was applied into the data and splitting of the data occurred, this distinguished to provide results based on how accurate the algorithm was applied. There were a total of 357 patients that have benign tumors and 212 patients that have malignant tumors. The confusion matrix gave an accuracy score of 97.37%. For the hypothesis picking the area and radius means were the best fit attributes because it depicts the measure of how big the tumor is. Majority of the malignant (cancerous) tumors are approximately above 17.5 cm. There aren't any benign tumors after that. So, that means the hypothesis is wrong due to the radius being longer than 17.5 cm , which will determine that the tumor is malignant. Therefore, the size of the mass does play a role in breast cancer, as the scatter plots shown in results have depicted that the tumors after 17.5 are malignant and tumors below 17.5 are benign. This is the analysis done for the breast cancer prediction using a Machine Learning algorithm of classification.

REFERENCES

Breast Cancer Dataset from UCI Machine Learning Repository

Website: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

Breast Cancer Dataset from UCI Machine Learning Repository

Kaggle: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Article*: Breast Cancer Prediction via Machine Learning*

M. S. Yarabarla, L. K. Ravi and A. Sivasangari, "Breast Cancer Prediction via Machine Learning," 20193rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 121-124, doi:10.1109/ICOEI.2019.8862533.

Article: *Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction*

Li, Y., & Chen, Z. (2018). Performance evaluation of machine learning methods for breast cancerprediction. Appl Comput Math, 7(4), 212-216.

CONTACT

SNEHA KUMARAN

GMU EMAIL : SKUMARA2@GMU.EDU

CDS 490 PROFESSOR: PROFESSOR KINSER

GMU EMAIL: JKINSER@GMU.EDU