Assessing Breast Cancer

Data Analysis and Machine Learning Study on the UCI ML Breast Cancer Wisconsin dataset.

Contents

1. Question and Hypothesis
2. Analysis and Approach
3. Finding Correlations
4. Testing Accuracy with Machine Learning
5. Conclusion

Question and Hypothesis

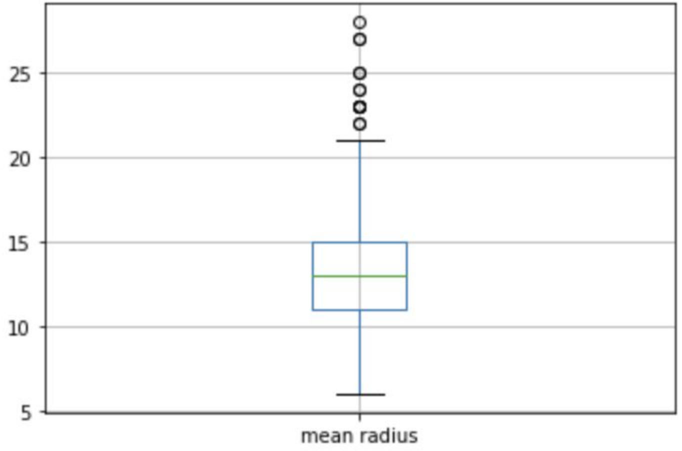
1. What do I want to learn from this data?
   - I want to analyze and determine any correlations in the quantitative data present that could predict the chances of a patient developing breast cancer.
2. Does radius and texture of a breast have an impact on how likely a patient could be at risk for developing breast cancer?
   - I believe there is a negative correlation between smaller radiuses, small texture values, and being at greater risk for developing breast cancer.
3. How will Machine Learning be used to test for accuracy of my analysis?
   - I will use three different machine learning models to test the accuracy of my findings to ensure my conclusion about the data shows correlation.

Analysis and Approach

1. Data table retrieved from sklearn.datasets
2. Formatted and converted into a Pandas dataframe to be analyzed with Seaborn and matplotlib.
3. Graphical visualizations will be employed to showcase analyzed data.
4. Machine learning models such as Logistic Regression, Support Vector Machine, ADABoost, and KNN will be implemented to find the most accurate classifier.

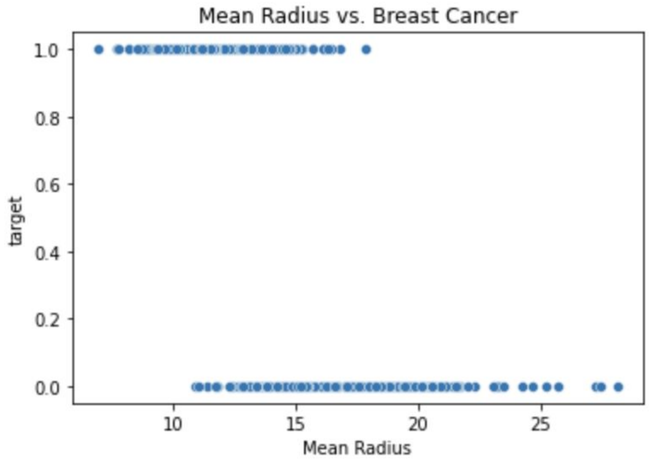
Finding Correlations - Determining Mean Radius

The mean radius of patients in this dataset is about 13cm with the minimum being at around 6cm and the maximum being around 28cm. This information will be useful when determining correlation between radius and whether a patient has breast cancer.



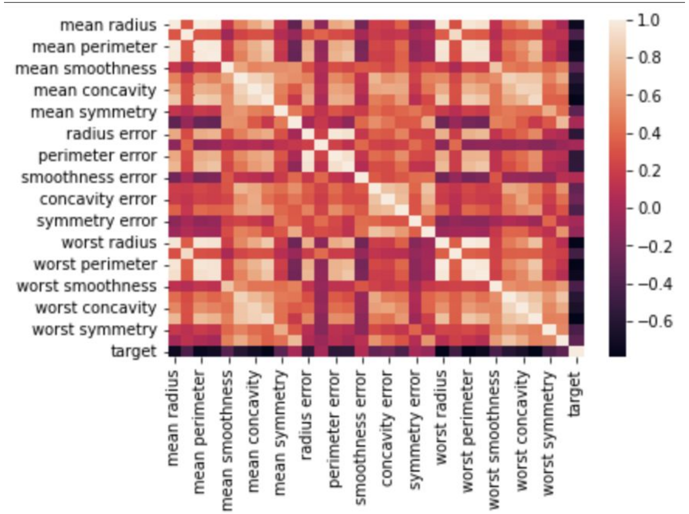
Finding Correlations - Correlation Between Mean Radius and Breast Cancer Diagnosis

Using a scatter plot, we can see that mean radius does have a negative correlation with breast cancer diagnosis. Smaller mean radiuses fall more within the target (1 being the patient has breast cancer), whereas larger mean radiuses fall more towards the patient not having breast cancer.



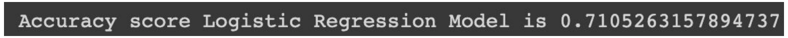
Finding Correlations - Determining Linear Relationships Between Features and Target

Using a heatmap, we can confirm the negative correlation of radius and the likelihood of the patient having breast cancer. With this data, we can confirm my hypothesis and go into further determining accuracy with machine learning models.

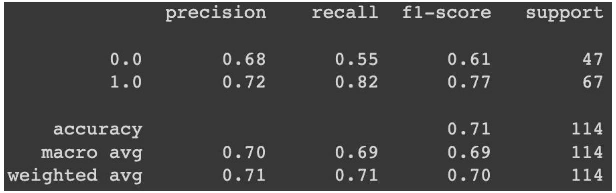


Testing Accuracy

Logistic Regression

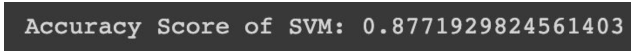


```
Accuracy score Logistic Regression Model is 0.7105263157894737
```

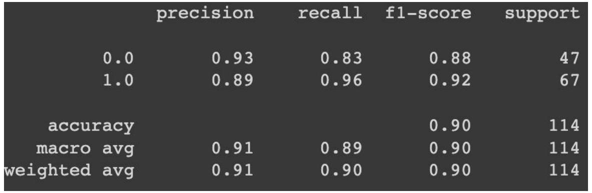


```
              precision    recall  f1-score   support

         0.0       0.68      0.55      0.61        47
         1.0       0.72      0.82      0.77        67

    accuracy                           0.71       114
   macro avg       0.70      0.69      0.69       114
weighted avg       0.71      0.71      0.70       114
```

Support Vector Machine

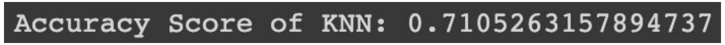


```
Accuracy Score of SVM: 0.8771929824561403
```

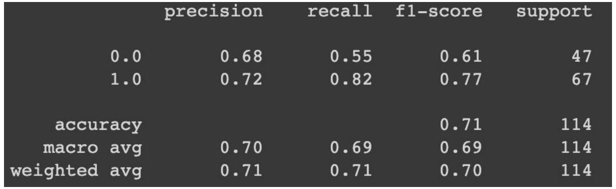


```
              precision    recall  f1-score   support

         0.0       0.93      0.83      0.88        47
         1.0       0.89      0.96      0.92        67

    accuracy                           0.90       114
   macro avg       0.91      0.89      0.90       114
weighted avg       0.91      0.90      0.90       114
```

K Nearest Neighbors



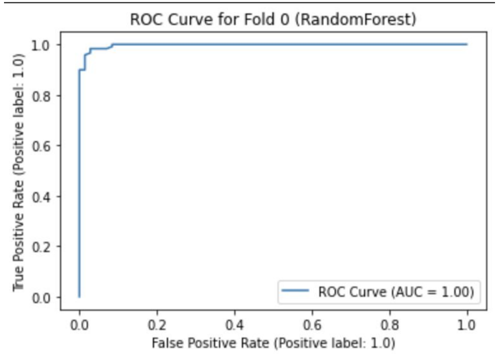
```
Accuracy Score of KNN: 0.7105263157894737
```

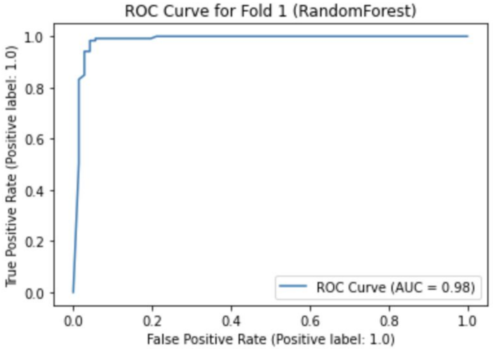


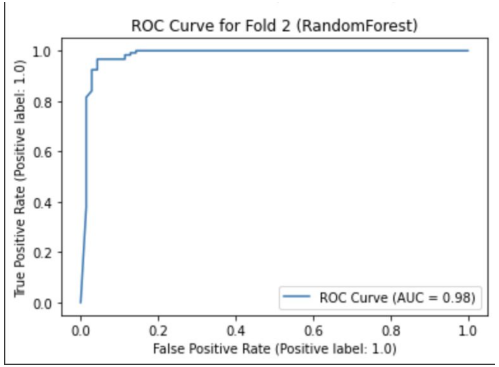
```
              precision    recall  f1-score   support

         0.0       0.68      0.55      0.61        47
         1.0       0.72      0.82      0.77        67

    accuracy                           0.71       114
   macro avg       0.70      0.69      0.69       114
weighted avg       0.71      0.71      0.70       114
```

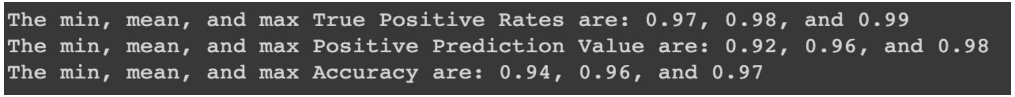
Testing Accuracy cont.

Random Forest





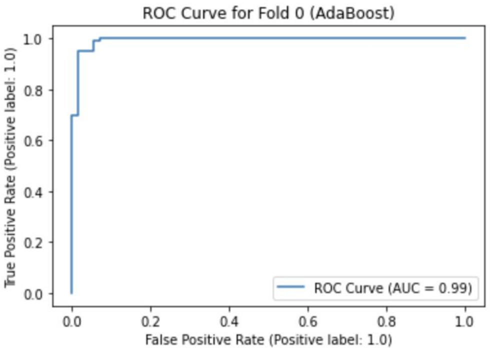


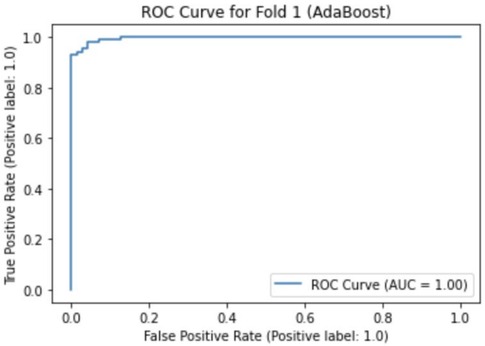


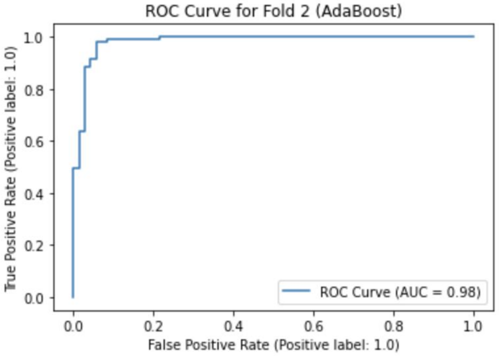
```
The min, mean, and max True Positive Rates are: 0.97, 0.98, and 0.99
The min, mean, and max Positive Prediction Value are: 0.92, 0.96, and 0.98
The min, mean, and max Accuracy are: 0.94, 0.96, and 0.97
```

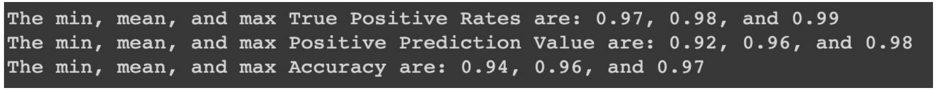
Testing Accuracy cont.

AdaBoost









```
The min, mean, and max True Positive Rates are: 0.97, 0.98, and 0.99
The min, mean, and max Positive Prediction Value are: 0.92, 0.96, and 0.98
The min, mean, and max Accuracy are: 0.94, 0.96, and 0.97
```

Conclusion

Using the different classifiers, I can infer the Adaboost classifier is more accurate as it gave us better prediction values and accuracy values. We can conclude there is a highly correlated relationship between radius, symmetry, and texture, when testing if a patient will have breast cancer.