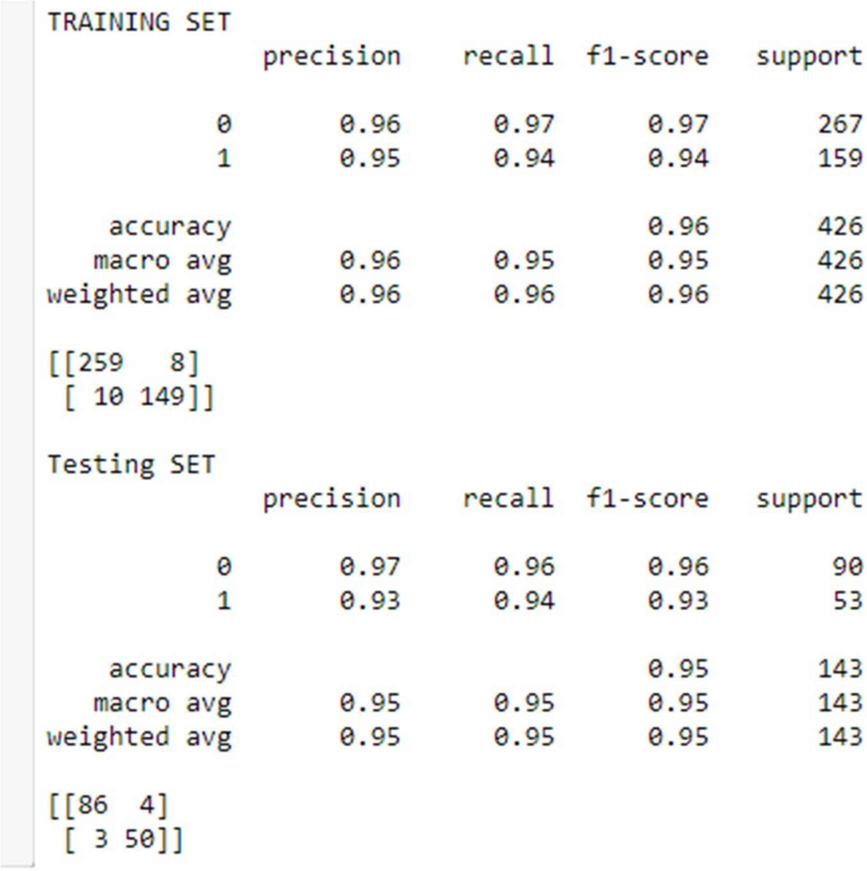
Samuel Slomowitz

Breast Cancer and Machine Learning

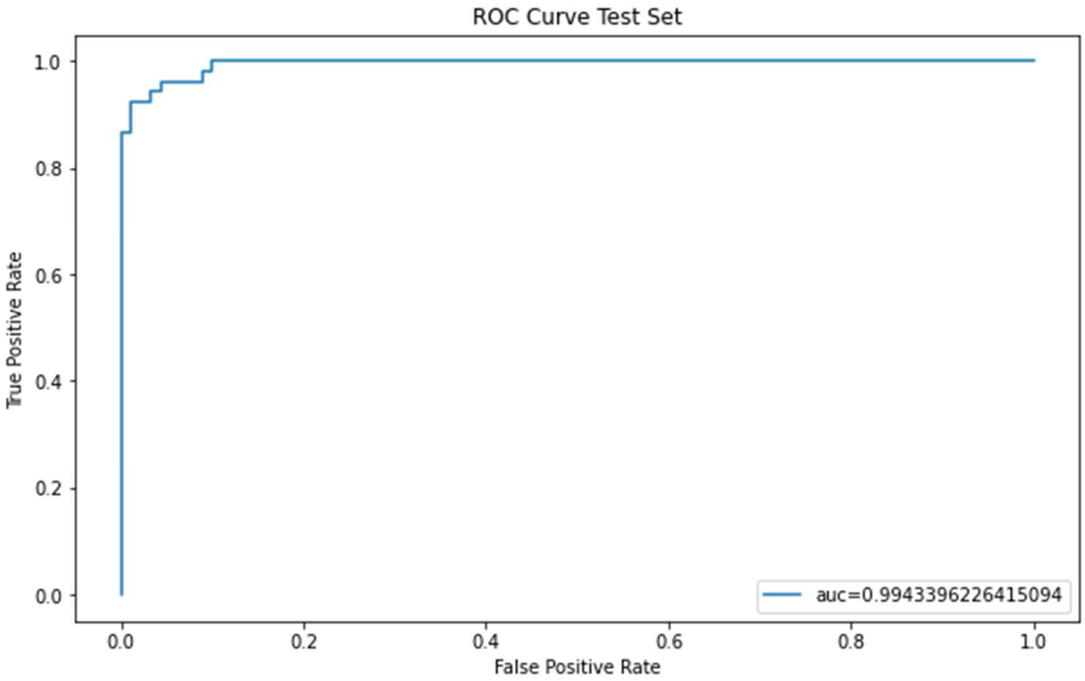
Breast cancer get much needed attention, yet tool to eradicate the surge in incidence per year are much needed. I decided to better understand the disease by delving into machine learning tools like classification methods in supervised learning. I found the top two marker for disease classification were concave points worst and perimeter worst. According to UC Irvine Machine Learning Repository, concave points worst is a cellular nucleus metric that approximates the number of concave portions of a “digitized image of a fine needle aspirate (FNA) of a breast mass” (UC Irvine ML Repository).

Logistic Regression out-performed Random Forest and K Nearest Neighbors (KNN) as well as Support Vector Classifier (SVC).



```
TRAINING SET
              precision    recall  f1-score   support

           0       0.96      0.97      0.97       267
           1       0.95      0.94      0.94       159

    accuracy                           0.96       426
   macro avg       0.96      0.95      0.95       426
weighted avg       0.96      0.96      0.96       426

[[259   8]
 [ 10 149]]

Testing SET
              precision    recall  f1-score   support

           0       0.97      0.96      0.96        90
           1       0.93      0.94      0.93        53

    accuracy                           0.95       143
   macro avg       0.95      0.95      0.95       143
weighted avg       0.95      0.95      0.95       143

[[86  4]
 [ 3 50]]
```

Figure: Confusion Matrix for Logistic Regression



ROC Curve for Logistic Regression Model

The logistic regression model was optimized by selecting the top 4 absolute value correlations scores for the diagnosis of breast cancer (binary: malignant or benign). Since logistic regression is a linear function, the extraneous dummy variable was dropped.

In conclusion, higher values of concave points worst, perimeter worst, concave points mean, and radius worst led to higher classification value of breast cancer. Machine learning remains a vital tool to better understand cancer diagnosis and prognosis.