Breast Cancer Prediction

CAPSTONE 2

BY ADRIANA THAMES

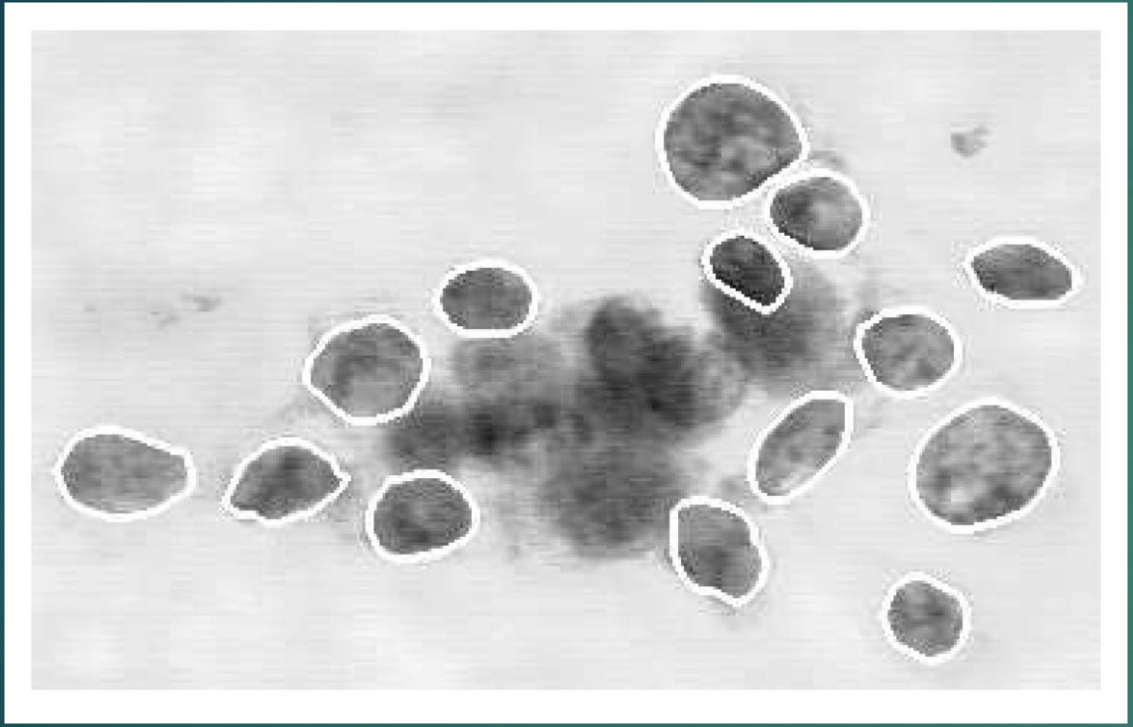
Introduction

- Breast cancer diagnosis can be performed on a breast tumor by the procedure Fine Needle Aspiration (FNA) biopsy.

- Benign and malignant cells can be differentiated based on certain cell nuclei characteristics, and these can be visualized on a digitized image of a FNA biopsy.

- The purpose of this project is to enhance the sensitivity and specificity of the FNA biopsy test by building a predictive model which can classify a tumor as benign or malignant based on cell nuclei characteristics seen on digitized images of FNA biopsies.

Description of the Data Set



A magnified image of a malignant breast FNA

A curve-fitting algorithm was used to outline the cell nuclei.

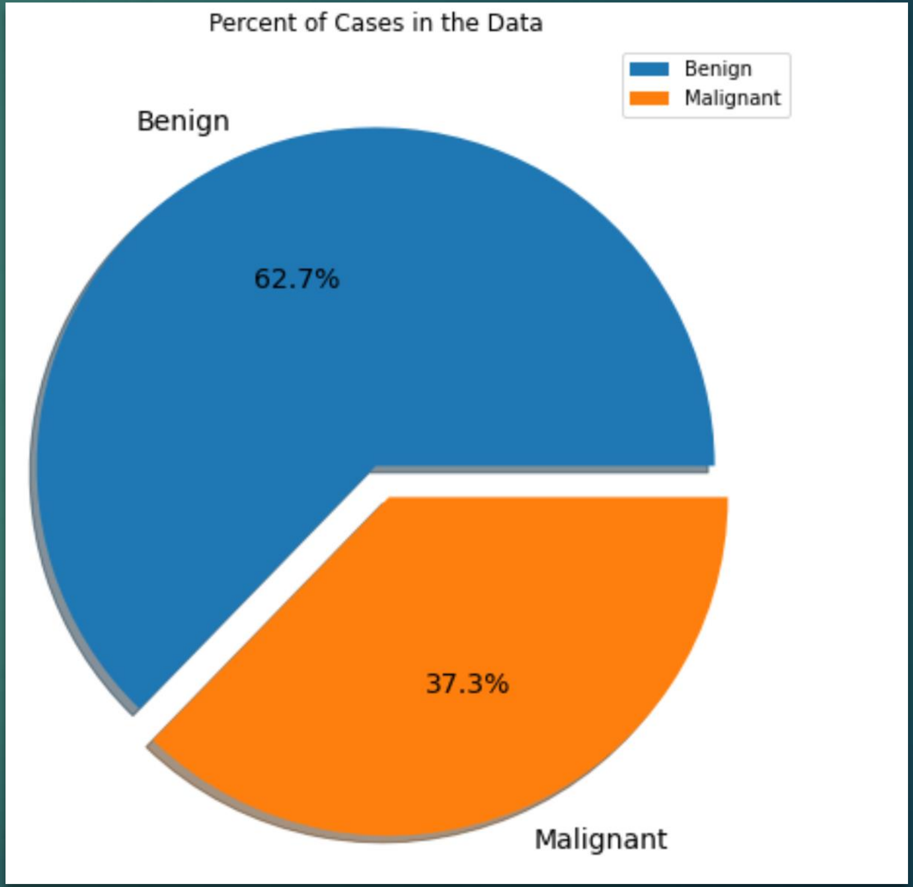
Each digitized FNA biopsy image is analyzed for 10 specific cell nuclei features including size, symmetry, and density.

EDA (Exploratory Data Analysis)

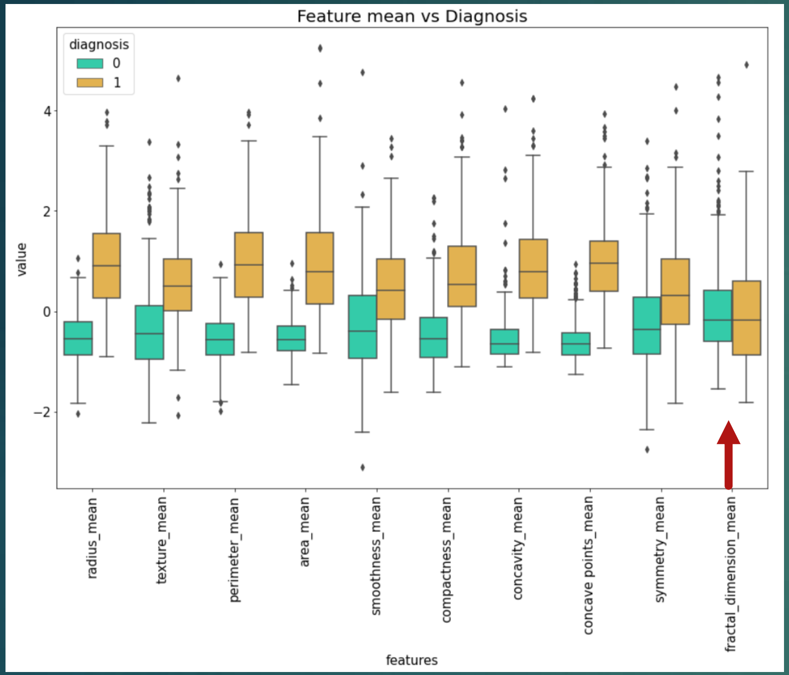
Total of 569 samples.

Ten real-valued features computed for each cell nucleus:

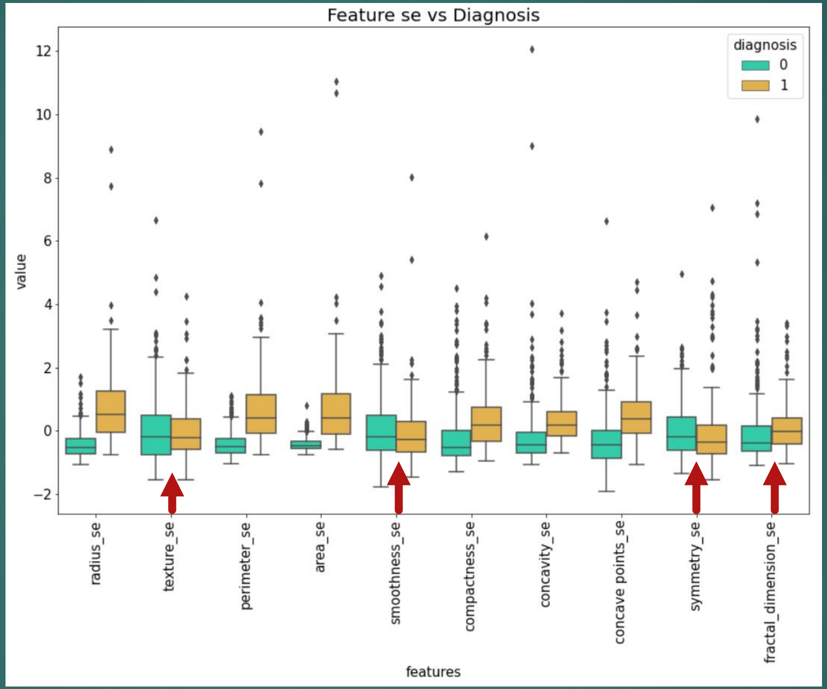
- a) radius (mean of distances from center to points on the perimeter)
  b) texture (standard deviation of gray-scale values)
  c) perimeter
  d) area
  e) smoothness (local variation in radius lengths)
  f) compactness (perimeter^2 / area - 1.0)
  g) concavity (severity of concave portions of the contour)
  h) concave points (number of concave portions of the contour)
  i) symmetry
  j) fractal dimension ("coastline approximation" - 1)
- Total of 30 numerical independent features calculated from the real values as mean, standard error and worst (mean of the 3 largest values)
- One categorical feature (target) which is the diagnosis
- The majority are Benign cases with 62.7%. Malignant are 37.3%.



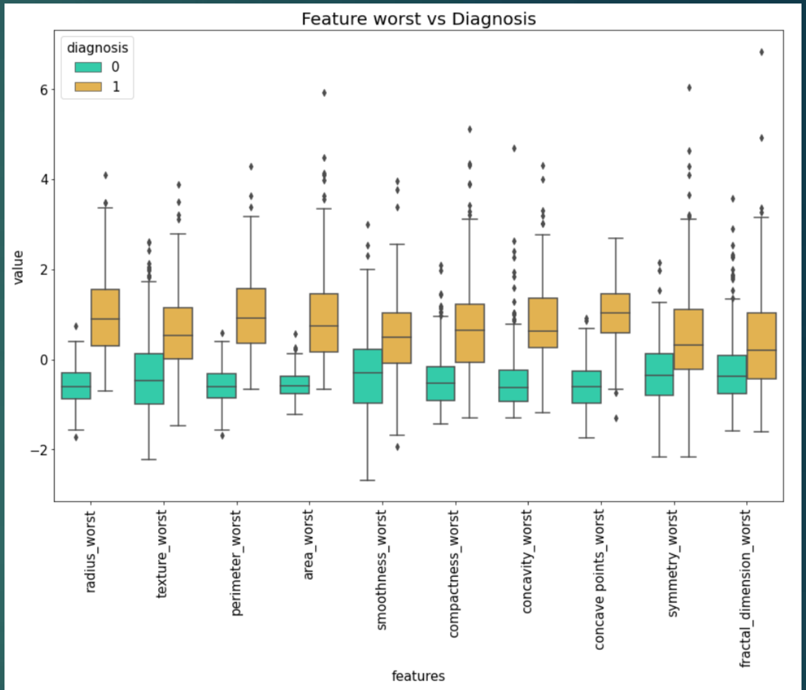
Data Visualization



A: Feature Mean vs Diagnosis



B: Feature Mean vs Diagnosis



C: Feature Mean vs Diagnosis

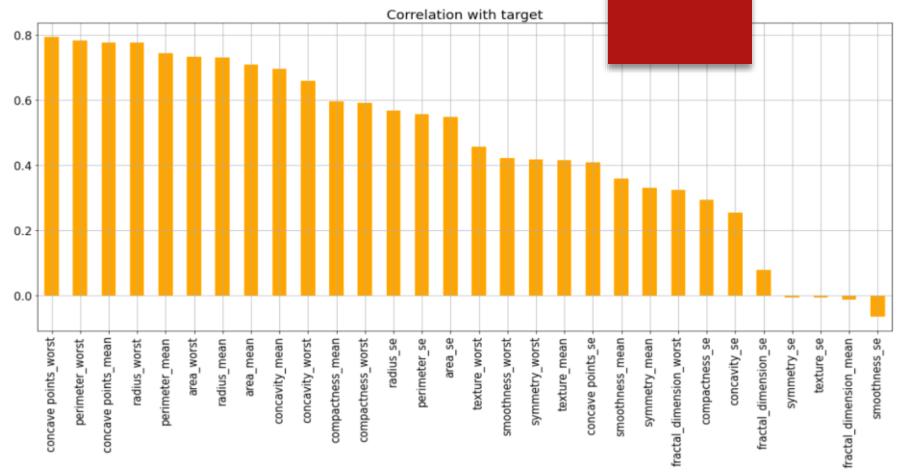
The box plot includes markers indicating the median and the interquartile range, which allows clear visualization of malignant tumors which have higher median values in all the features except for Fractal dimension_mean, texture_se, smoothness_se, symmetry_se and fractal dimension_se

Multivariate Analysis

- The correlation with the target shows that size (perimeter, area, radius) and concave points are highly correlated to the categorical target feature.

  Features such as smoothness_se, symmetry_se, texture_se, fractal_dimension_mean and fractal_dimension_se are not correlated to the target.

- The heatmap shows a high correlation (value =1.0) between independent features like radius, perimeter, and area. This can be an indication of collinearity.



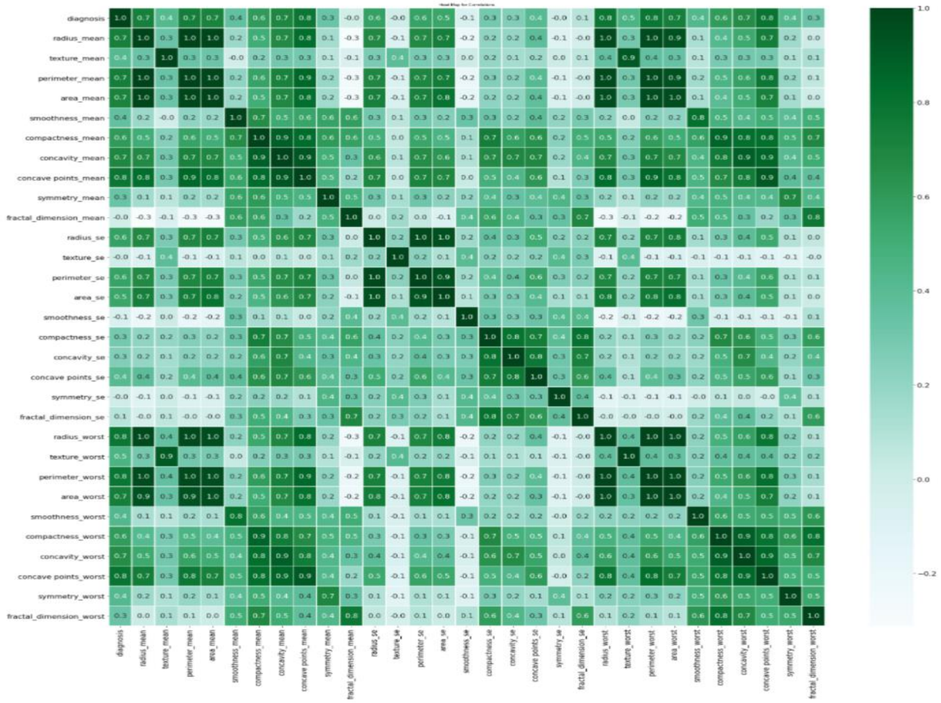
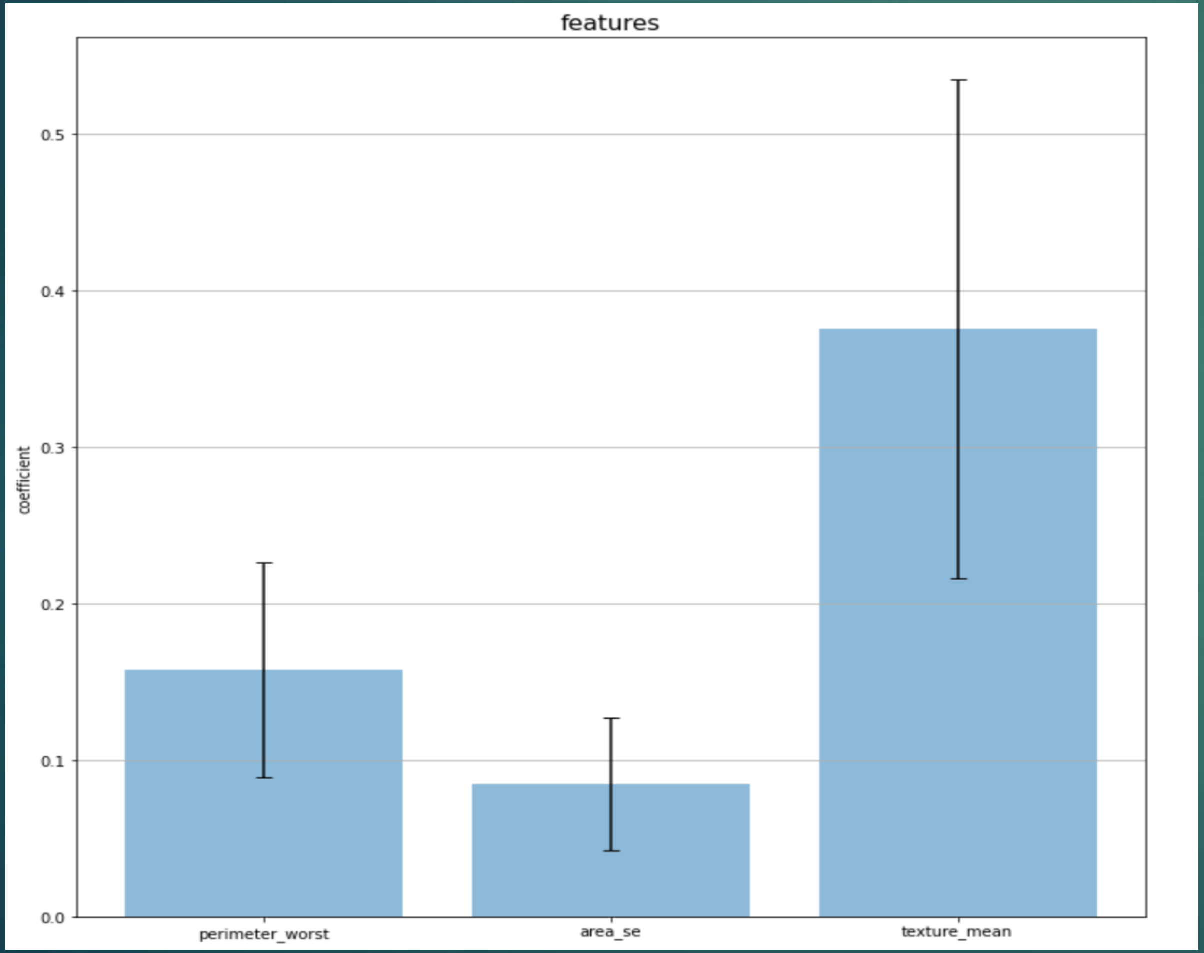


Figure 10. Correlation Heat map

Regression Analysis of Feature Effect on Target



Example of some of the feature effect on target:

Perimeter_worst has a coefficient β of 0.1575 which correspond to (e^{β} -1) that is 17 % increased chance of malignancy with single unit increase

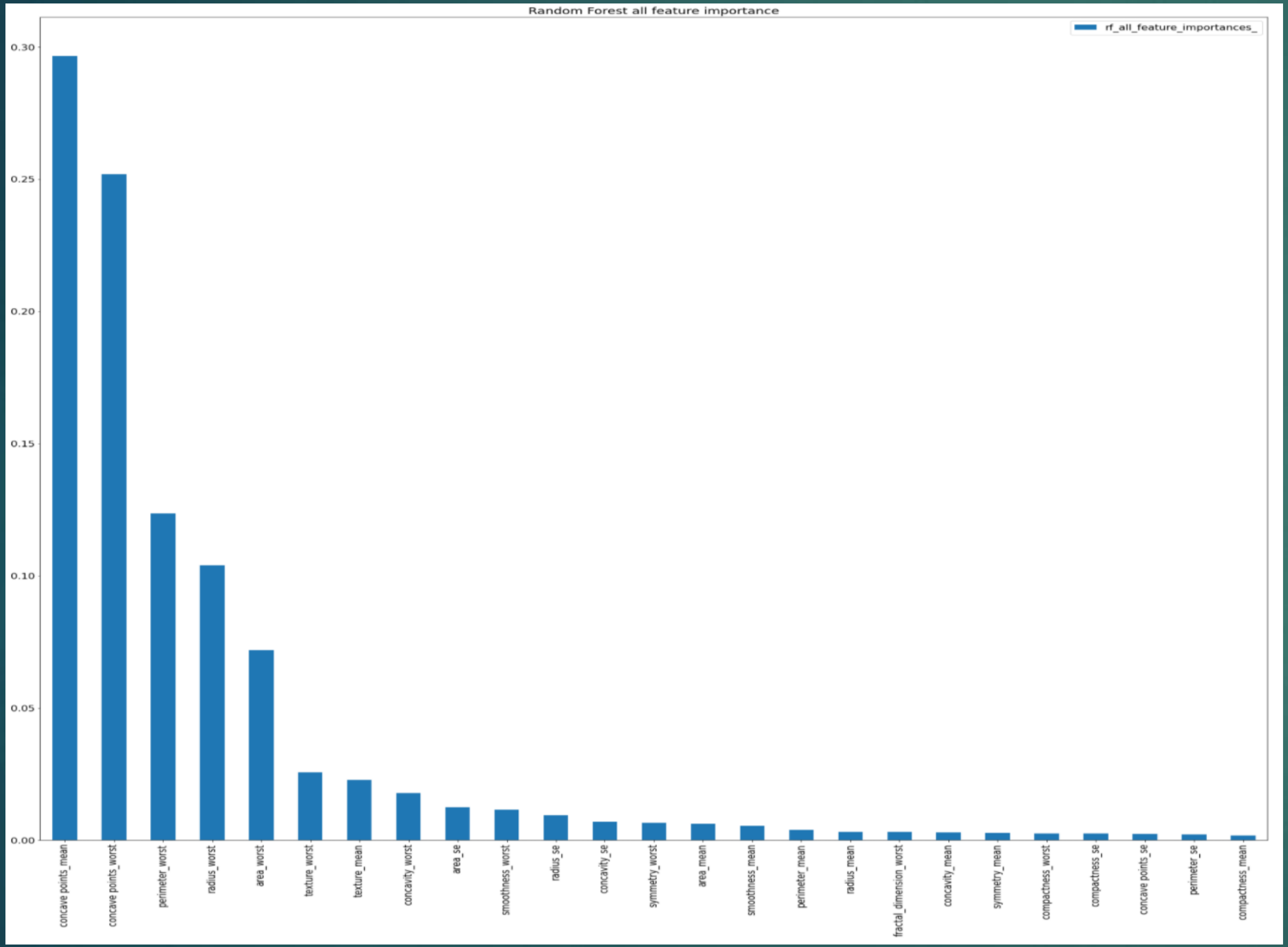
For area_se is 8 %

For texture_mean is 45 %

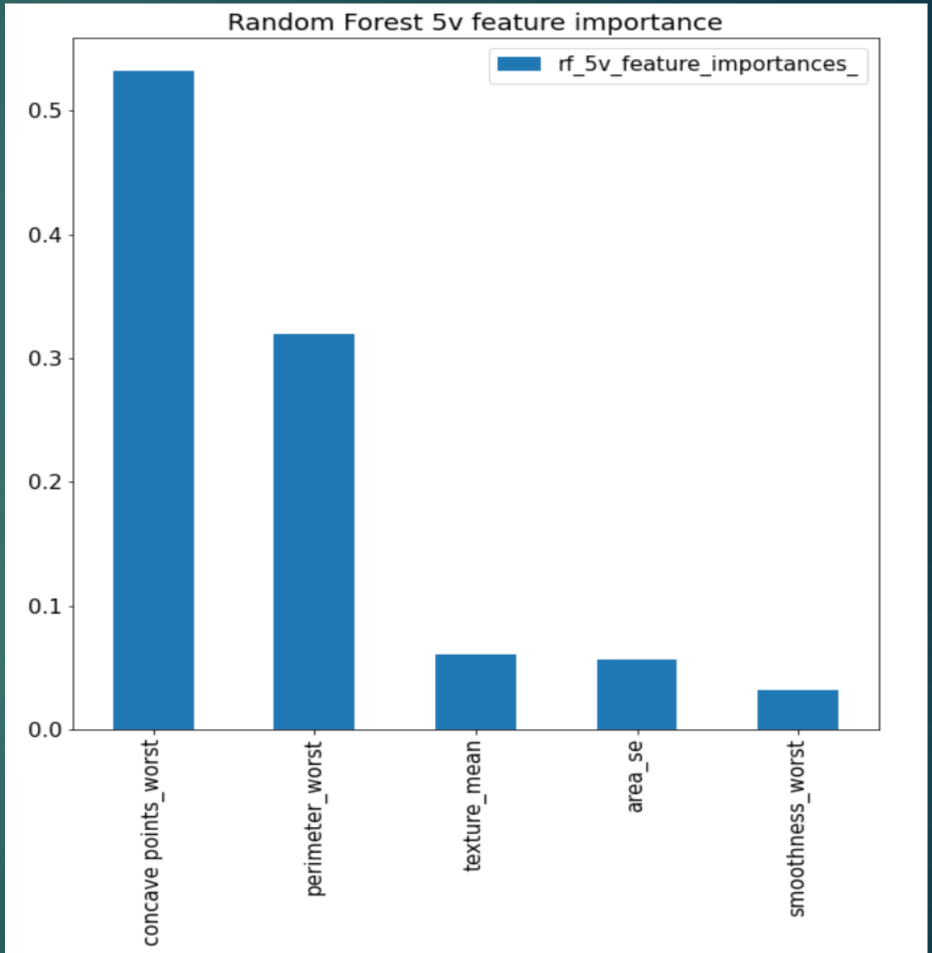
Random Forest Feature Importance and Feature Selection

Set	ROC_AUC Score	Brier Score
ALL Features	**0.995443857**	**0.025637427**
20 Features from Feature Importance	0.995149912	0.026006631
10 Features from Feature Importance	0.995149912	0.023722679
5 Features from Feature Importance	0.993239271	0.033512618
5v Features from VIF and Feature Importance	**0.995443857**	**0.025676023**

Selected Features for modeling



Feature Importance for ALL features



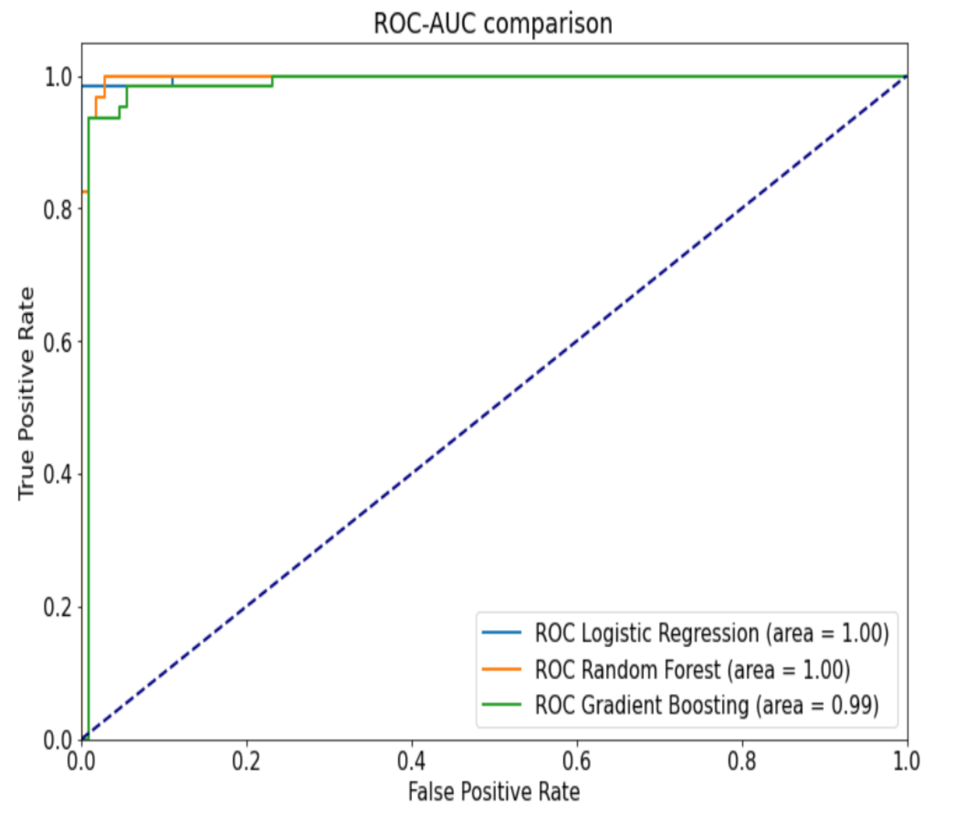
Feature Importance for 5v features

Modeling ROC_AUC Scores

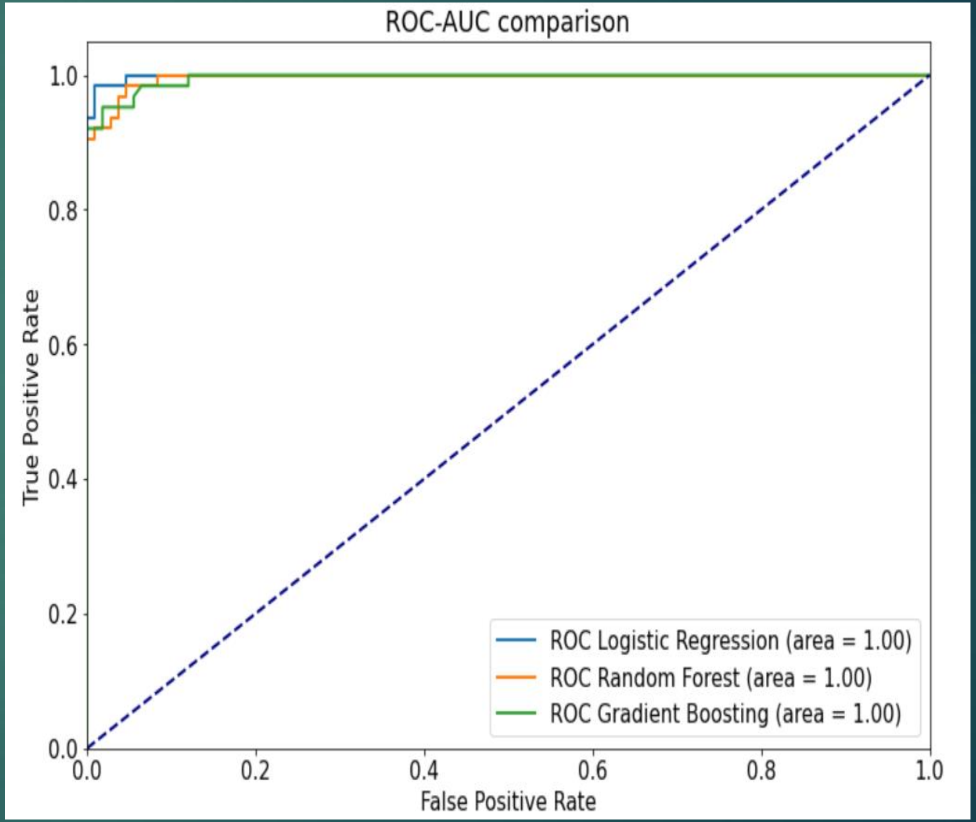
- Model Testing for the two final feature sets- All features and 5v features (Selected from random forest feature importance and VIF)
- The Logistic Regression model provided the highest ROC_AUC score

Model	Optimal Features	Best Parameter	ROC_AUC Score
Random Forest	ALL	max depth=10, No estimators=100	0.995443857
Boosting Gradient	ALL	max_depth=1, No_estimators=500	0.990550223
Logistic Regression	**ALL**	C=1	**0.998236332**
Random Forest	5v Features	max depth=None, No estimators=100	0.995443857
Boosting Gradient	5v Features	max_depth=1, No_estimators=100	0.994708995
Logistic Regression	**5v Features**	C=1	**0.998824221**

Visualization of the ROC-AUC for the tested models



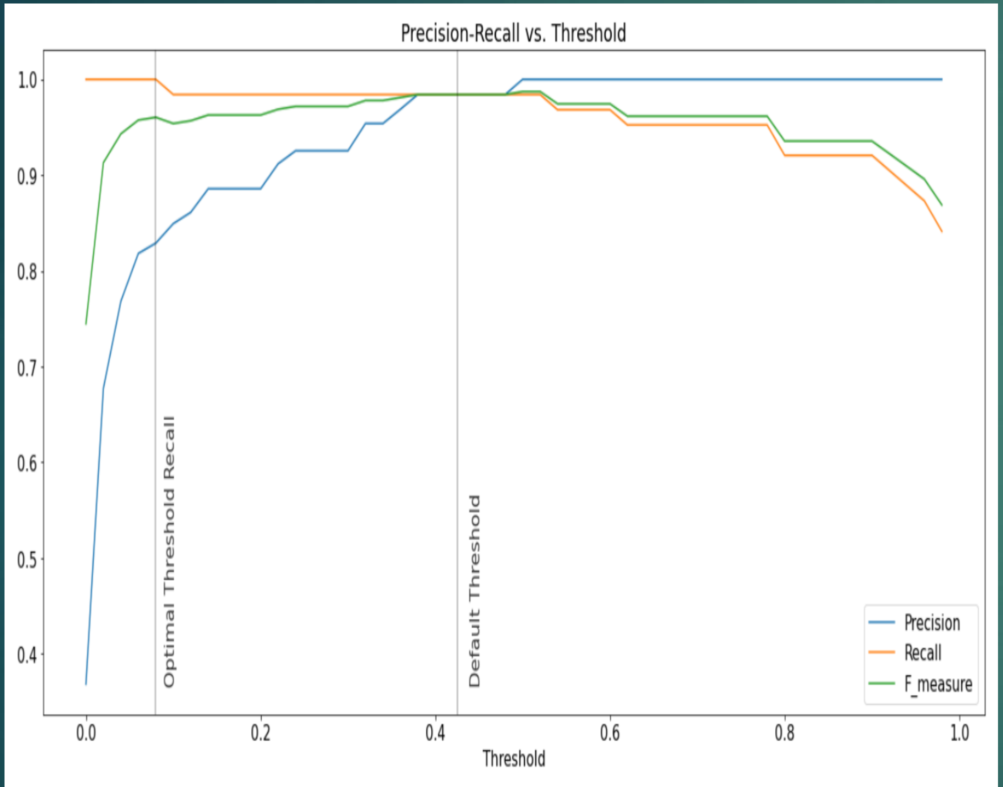
ALL features



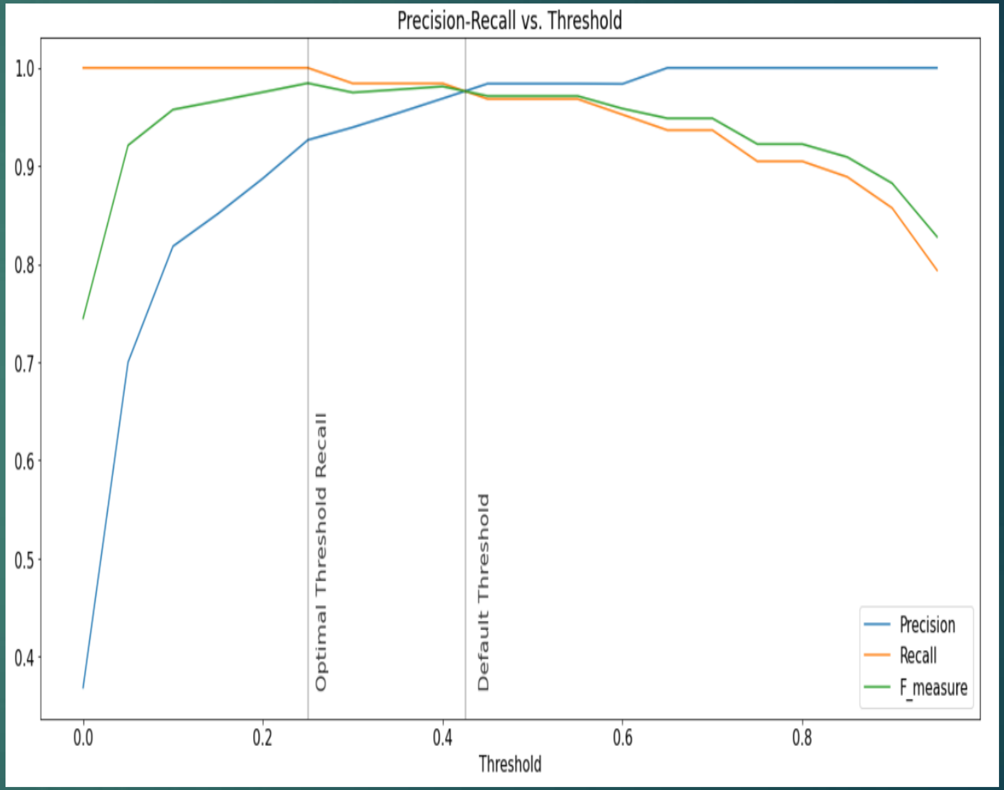
5v features

Thresholding the Model

Finding the threshold probability above which the prediction will be malignant

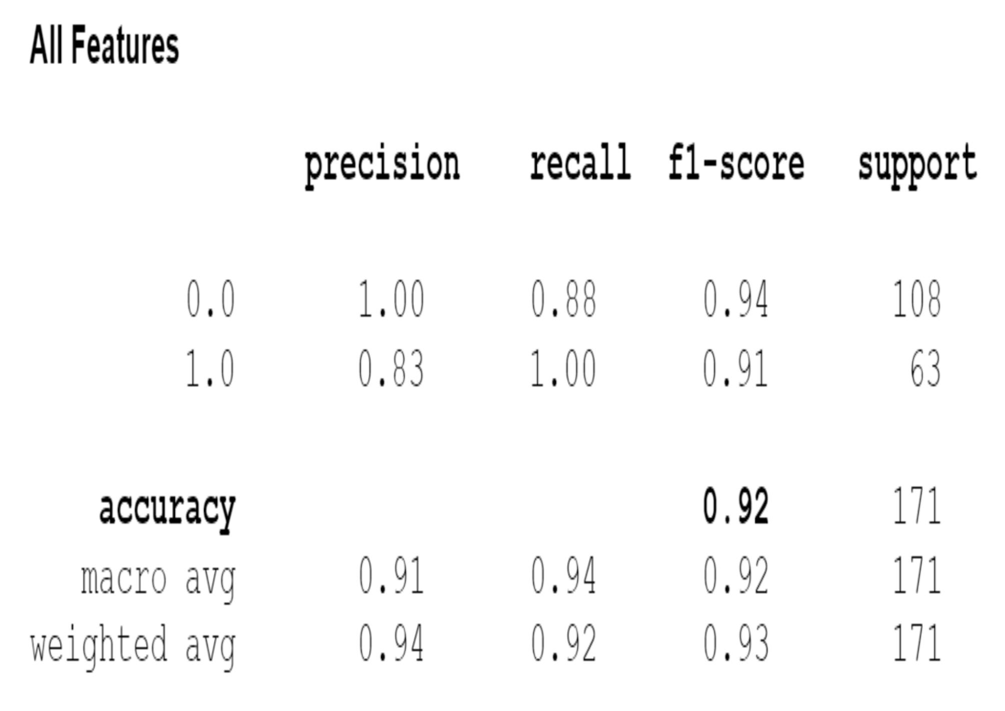


ALL features



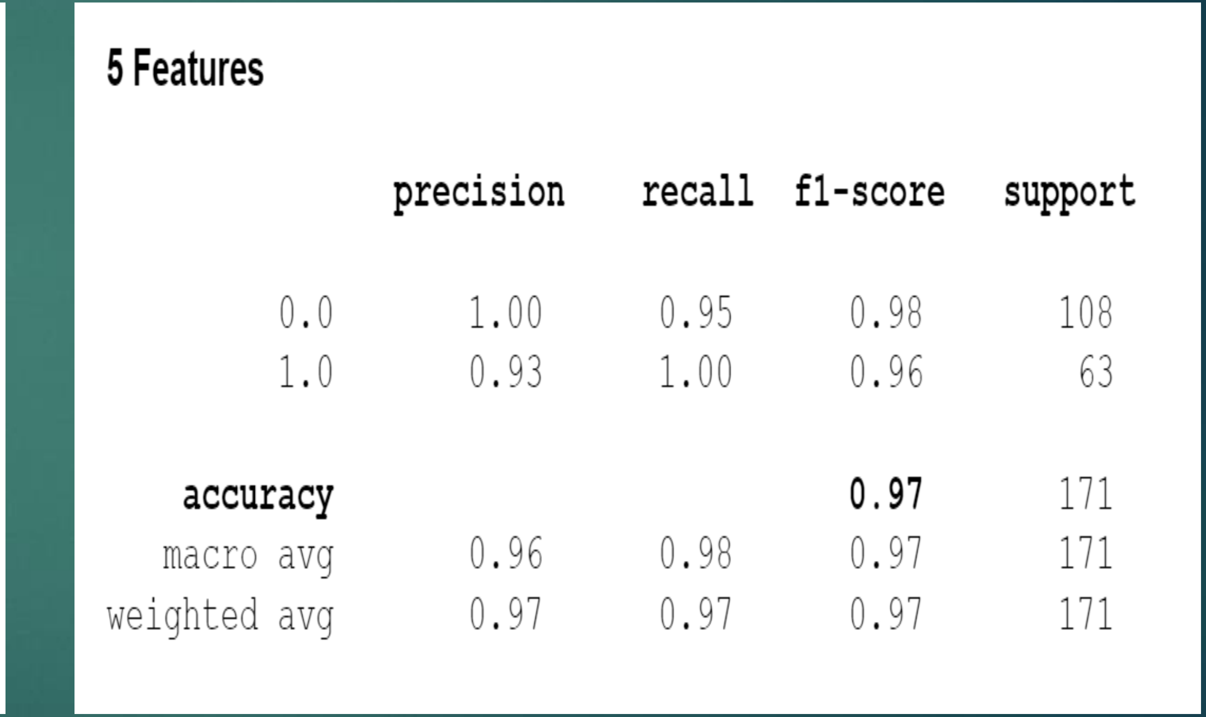
5v features

Classification Report



All Features

```
              precision    recall  f1-score   support

         0.0       1.00      0.88      0.94       108
         1.0       0.83      1.00      0.91        63

    accuracy                           0.92       171
   macro avg       0.91      0.94      0.92       171
weighted avg       0.94      0.92      0.93       171
```



5 Features

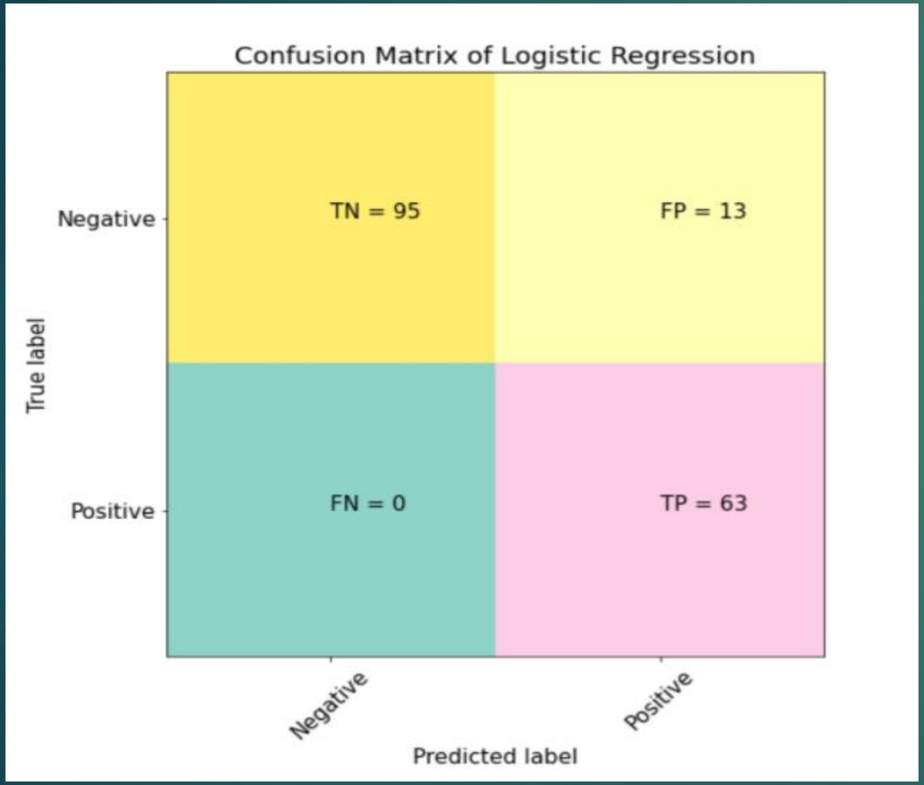
```
              precision    recall  f1-score   support

         0.0       1.00      0.95      0.98       108
         1.0       0.93      1.00      0.96        63

    accuracy                           0.97       171
   macro avg       0.96      0.98      0.97       171
weighted avg       0.97      0.97      0.97       171
```

ALL features

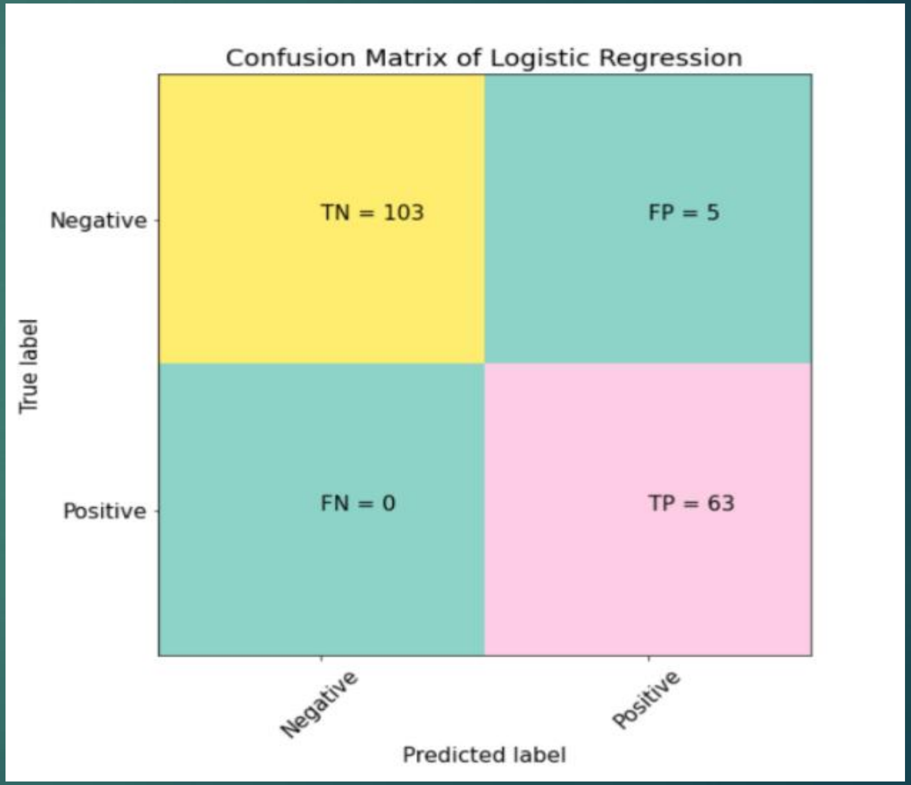
5v features

Confusion Matrix

Threshold adjusted for recall equal to 1



ALL features



5v features

Conclusion

The purpose of this project is to enhance the sensitivity of the FNA biopsy test by building a predictive model which can classify a tumor as benign or malignant based on these cell characteristics as seen on the digitized images.

The predictive model which can best classify a tumor as benign or malignant in my analysis was the Logistic Regression model with the following 5 predictors:

1. perimeter_worst
2. smoothness_worst
3. area_se
4. concave points_worst
5. texture_mean

An accuracy of 97% was achieved using this model with emphasis in obtaining a higher recall.

The future work for this project would involve the study of larger data sets and expansion of the number of classifiers to better understand which are most efficient in model prediction.

The benefits from the FNA procedure leverages the advantage of being quick, cost-effective, minimally invasive, and leaves no scar. This can become a standard practice as the model accuracy improves.