Wisconsin Breast Cancer Prediction

Adriana V Thames

November 6th, 2021

Table of Contents

Introduction

Breast cancer is a major public health concern in the United States, where the average risk of a woman developing breast cancer within her lifetime is about 13%. Breast cancer is also the second leading cause of cancer death in women, with a 2.6% mortality rate. The long-term survival of cancer victims depends greatly on the early detection and accuracy of the diagnosis. [1]

To diagnose breast cancer, a surgical biopsy or Fine Needle Aspiration (FNA) biopsy can be performed on a breast tumor. In an FNA biopsy, the doctor uses a very thin, hollow needle attached to a syringe to withdraw or aspirate a small amount of tissue or fluid from a suspicious area. The biopsy sample is then analyzed to see if cancer cells are detected.

Using the FNA has the advantage of being quick, cost-effective, minimally invasive, and leaves no scar. FNA provides a sufficient pathologic diagnosis to avoid open surgical biopsy in 63-85% of the cases. Estimation of cost savings based on the distribution of cases and indications for surgery suggested a savings of $250,000 to $750,000 per 1000 FNA performed, or approximately 5500 Relative Value Units.[2]

However, there are instances where the FNA results are inconclusive for making a definitive diagnosis of breast cancer.

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.

1. Description of the Data Set

The Breast Cancer Prediction Wisconsin dataset consists of information about tumor features derived from digitized images of the FNA. Using an image analysis software called Xcyt, the boundaries of a cell nuclei can be determined from a digitized 640×400, 8-bit-per-pixel grayscale image of the FNA. Each digitized FNA biopsy image is analyzed for 10 specific cell nuclei features including size, symmetry, and density. For each of the 10 specific cell nuclei features, 3 statistics are calculated - the mean, the standard error, and the "worst" mean (i.e. mean of the three largest values). These statistical values are ultimately correlated to the categorical target feature which defines the tumor type - malignant or benign.

The data used for this project is available at the following website:
https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

The dataset contains 569 samples from patients with a breast mass. Each sample is assigned the following information:
1) **ID number**
2) **Diagnosis:** (M = malignant, B = benign)
3) **Radius:** (mean of distances from center to points on the perimeter)

4) **Texture:** (Standard Deviation of gray-scale values). Each pixel of an image is represented by the 8-bit integer, or a byte, from 0 to 255 providing the amount of light, where 0 is clear black and 255 is clear white. The darker the image is the lower is the mean of intensity level of a pixel, i.e., byte. So, the SD of gray-scale values means how intense levels are spread for individual cells. The higher the SD the more image contrast is found.

5) **Perimeter:** The total distance between the snake points constitutes the nuclear perimeter.

6) **Area**: Nuclear area is measured simply by counting the number of pixels on the interior of the snake and adding one-half of the pixels in the perimeter.

7) **Smoothness**: (local variation in radius lengths)

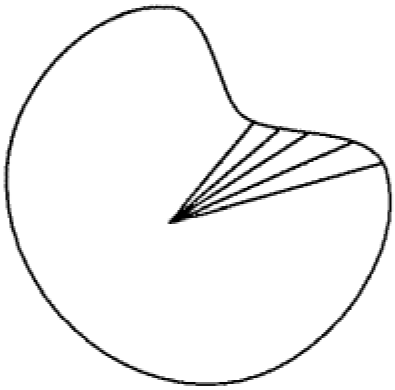


Figure 1: Radial Lines Used for Smoothness Computation

8) **Compactness**: ($\frac{perimeter^2}{area - 1}$) This dimensionless number is minimized by a circular disk and increases with irregularity of the boundary.

9) **Concavity:** (Severity of concave portions of the contour). The concavity is captured by drawing chords between two boundary points, which lie outside the nuclear. For the concavity mean the mean value of these lengths is calculated.

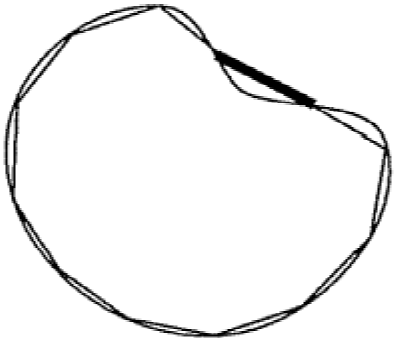


Figure 2: Chords Used to Compute Concavity

10) **Concave Points:** (number of concave portions of the contour). This feature is similar to concavity but measures only the number, rather than the magnitude, of contour concavities.
11) **Symmetry:** (measured along the major axis, or longest chord through the center). We then measure the length difference between lines perpendicular to the major axis and the nuclear boundary in both directions.

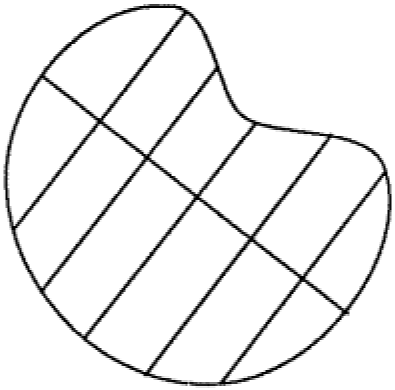


Figure 3: Segments Used in Symmetry Computation

12) **Fractal Dimension:** ("coastline approximation" - 1) The “coastline approximation” is described by Mandelbrot.[3]

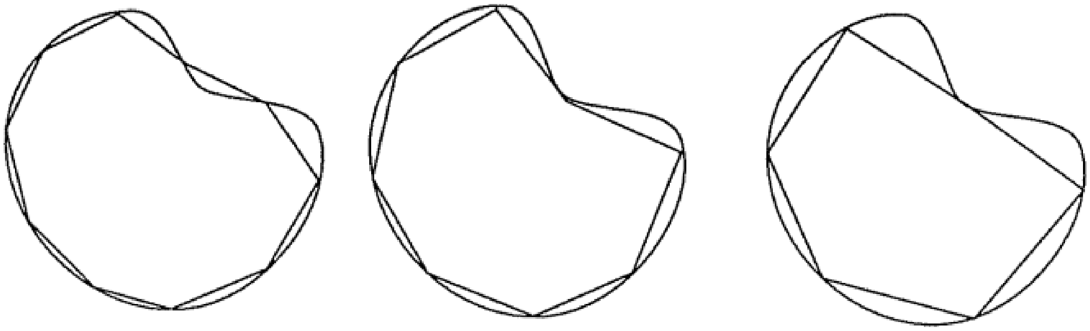


Figure 4: Sequence of Measurements for Computing Fractal Dimension

The 10 real-valued features correspond to the Mean (values from columns 3 to column 12), to the Standard Errors (values from columns 12 to 21), and the Worst, mean of the three largest values, (columns from 22 to 31).

Depicted below is an image of a malignant breast sample

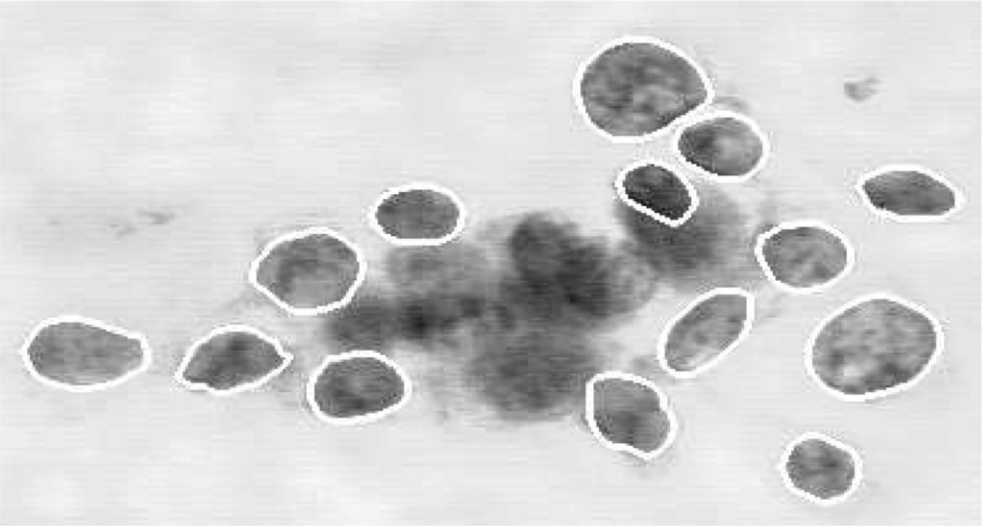


Figure 5: A magnified image of a malignant breast FNA. A curve-fitting algorithm was used to outline the cell nuclei. (Figure from Mangasarian OL., Street WN., Wolberg. WH. Breast Cancer Diagnosis and Prognosis via Linear Programming. Mathematical Programming Technical Report 94–10. 1994 Dec)

2. EDA (Exploratory Data Analysis)

The first column of the data set corresponds to the patient ID with a total of 569 samples. The second column represents the diagnosis (“Benign” or “Malignant”).

No of rows: 569

No of Columns: 31

There were no missing or NAN values. The percent of the outcome of the cases is shown below.

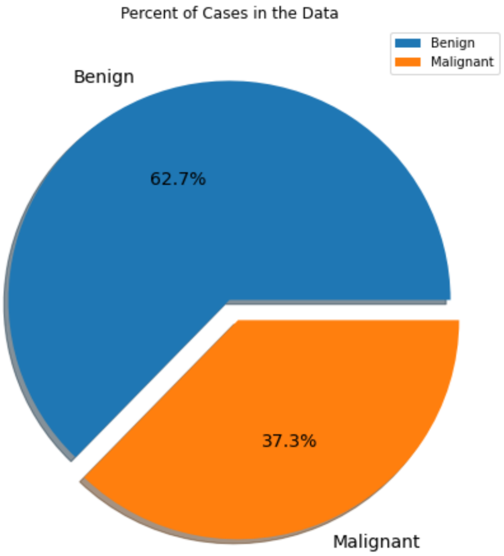


Figure 6: Pie plot of the percentage of Malignant and Benign tumors. The majority are Benign cases with 62,7 percent. Malignant are 37.3 percent.

● Data Visualization

Fig 7 and Fig 8 show a bi-variate analysis box plots of all feature means and standard deviations, comparing malignant to benign tumors. The box plot includes markers indicating the median and the interquartile range, which indicates that malignant tumors have higher mean values and standard deviations for almost all features. The exceptions are listed here:

- smoothness_se,
- texture_se,
- symmetry_se,
- fractal_dimension_mean,
- fractal_dimension_se

(where the median for malignant and benign are very close).

The box plot is also useful for seeing outliers. Even though there are several outliers, I included them in the modeling because there is insufficient information to reject them.
I ran inferential statistics through each chart and verified that certain features were not statistically significant for distinguishing malignant from benign cells. Thus, those features listed above were not included in feature selection and modeling.

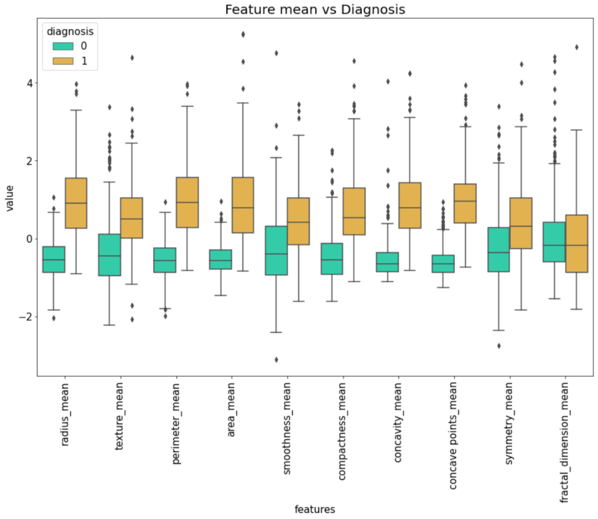


Figure 7: Feature Mean Vs. Diagnosis

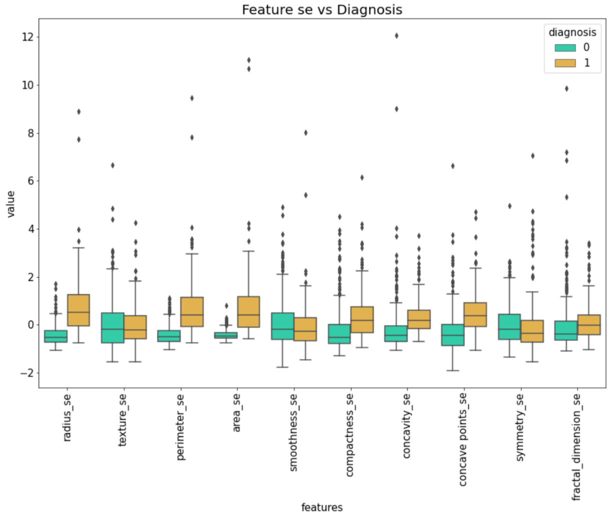


Figure 8: Feature SE Vs. Diagnosis

Below are the p values for the features that were not statistically significant:

Feature	P value
fractal_dimension_mean	p value=0.7599
texture_se	p value=0.8433
smoothness_se	p value=0.1102
symmetry_se	p value=0.8766
fractal_dimension_se	p value=0.0630

- ## Multivariate Analysis

To visualize the correlation between each pair of attributes, a heatmap of Pearson R values between features was created:

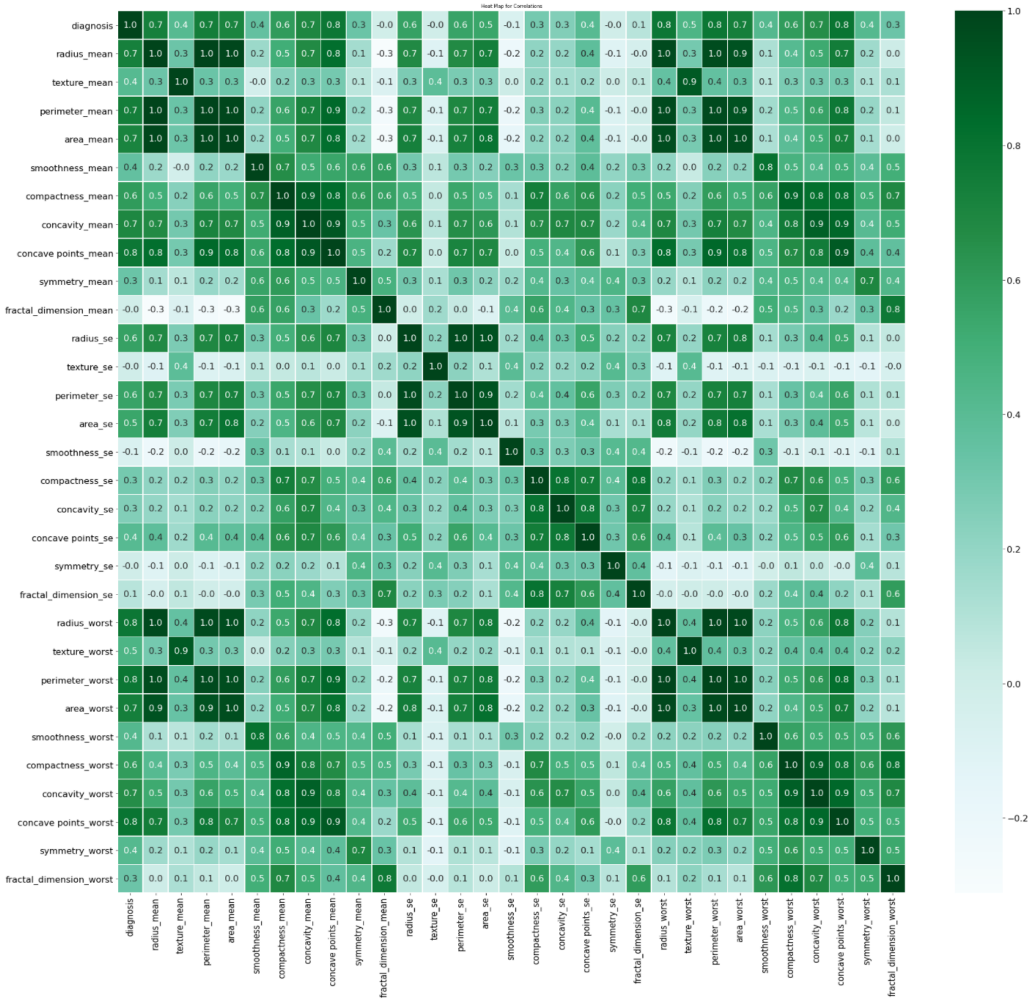


Figure 9. Correlation Heat map

In Fig 9 we can see many indications of strong multicollinearity, especially between features related to nucleus size: radius, perimeter and area. This is not surprising because all of these features are related to nucleus size, and it would make sense that “worst” and “mean” values and standard deviations might vary together. That said, it doesn’t make it any less problematic. Many of these correlations indicate perfect multicollinearity at r=1. As such, feature selection will definitely be needed in the modeling section to account for this multicollinearity and likely only a small portion of the radius, perimeter, and area features will ultimately be included.

Below is a bar graph to help visualize the correlation of the independent features with the target (diagnosis)

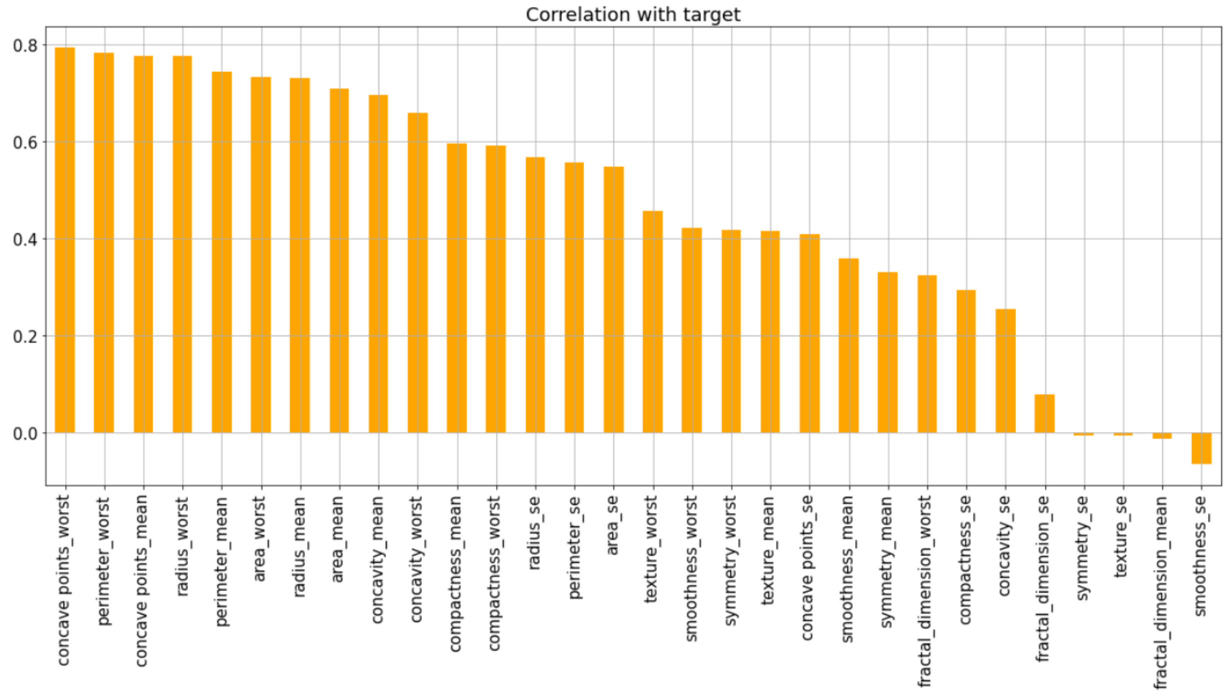


Figure 10. Top correlated features with diagnosis

This analysis shows that size (perimeter, area, radius) and concave points are highly correlated to the categorical target feature. In this display again it is confirmed that features such as smoothness_se, symmetry_se, texture_se, fractal_dimension_mean and fractal_dimension_se are not correlated to the target.

- ## Scaling

Due to the variation in magnitude ranges of the variables, a standardization method was needed prior to regression. For this process, a Robust Scaler algorithm was used. The Robust Scaler uses a similar method to the Min-Max scaler. It uses interquartile ranges rather than the min-max values, so that it is robust to outliers.

- ## Multicollinearity

Multicollinearity occurs when independent variables in a regression model are correlated. This correlation is a problem because independent variables should be independent. If the degree of correlation between variables is high enough, it can cause problems when fitting the model and interpreting the results.

Multicollinearity was detected via Variance Inflation Factors (VIF). The VIF is predicted by taking a variable and regressing it against every other variable. A VIF > 10 is an indicator that the coefficient size of that feature is not reliable for purposes of analysis. As such, we calculated the VIF for the selected features and removed them one by one until our VIF for features we intended to analyze was smaller than 10. The goal of this process is to reduce the variance of the coefficients for these features in a regression analysis so that these coefficients can be more reliably used to determine feature effects on the target and overall feature importance.

Before the feature selection, the highest VIF was over 3000 and perimeter_worst had a VIF of 405.15. After the feature selection, I chose 5 features which all had VIF scores less than 10 which can be seen below in Figure 11:

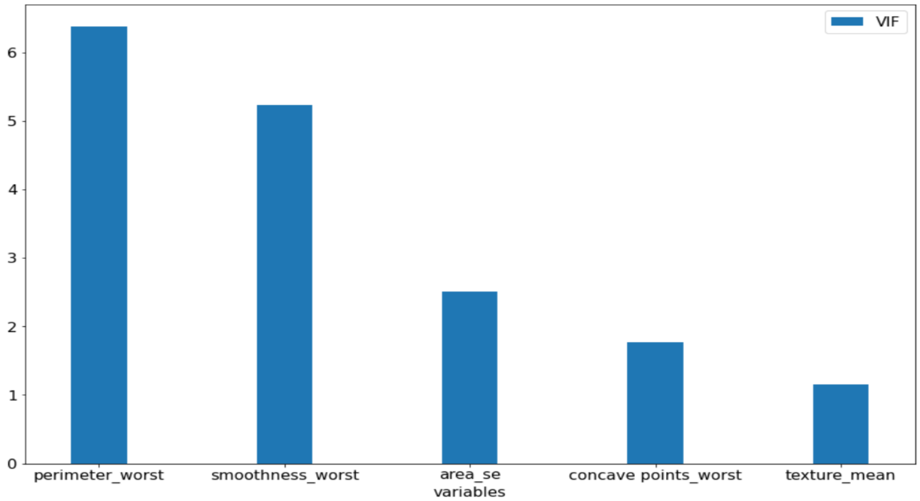
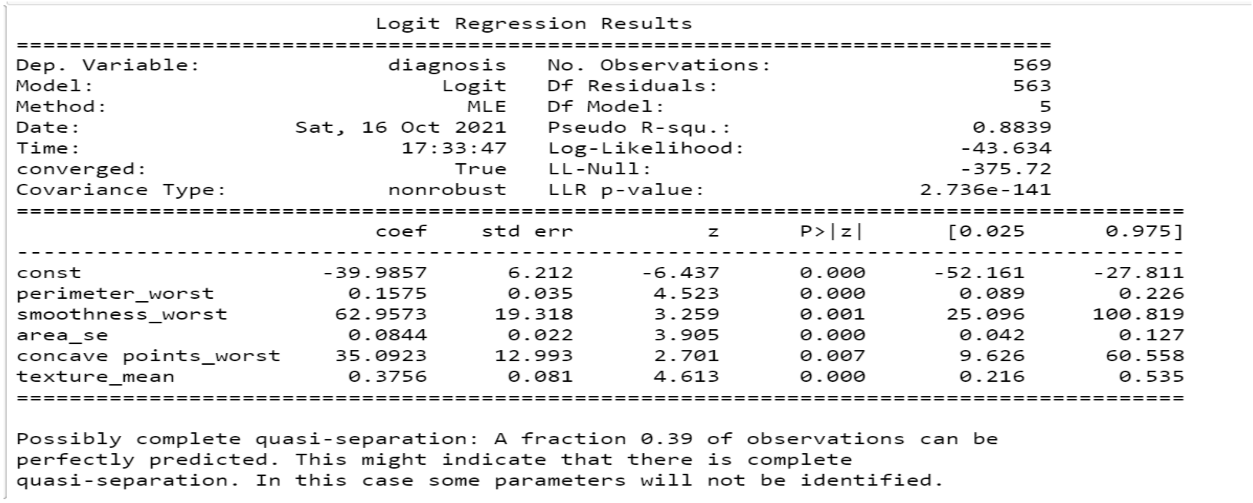


Figure 11. Variance Inflation Factor of the Selected 5 features

3. Regression Analysis of Feature Effects on Target Variable



```
                           Logit Regression Results
==============================================================================
Dep. Variable:              diagnosis   No. Observations:                  569
Model:                          Logit   Df Residuals:                      563
Method:                           MLE   Df Model:                            5
Date:                Sat, 16 Oct 2021   Pseudo R-squ.:                  0.8839
Time:                        17:33:47   Log-Likelihood:                -43.634
converged:                       True   LL-Null:                       -375.72
Covariance Type:            nonrobust   LLR p-value:                2.736e-141
========================================================================================
                           coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                  -39.9857      6.212     -6.437      0.000     -52.161     -27.811
perimeter_worst          0.1575      0.035      4.523      0.000       0.089       0.226
smoothness_worst        62.9573     19.318      3.259      0.001      25.096     100.819
area_se                  0.0844      0.022      3.905      0.000       0.042       0.127
concave points_worst    35.0923     12.993      2.701      0.007       9.626      60.558
texture_mean             0.3756      0.081      4.613      0.000       0.216       0.535
========================================================================================

Possibly complete quasi-separation: A fraction 0.39 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
```

Accounting for the 5 features above we can see some interesting effects of our features on predicting malignancy. Using log-odds, we can calculate the increased probability of malignancy caused by a one unit increase of any of these features while controlling for the other features present in the regression analysis done here.

For example, perimeter_worst has a coefficient of .1575 which corresponds to a 17% increased chance of malignancy with a single unit increase[1] - meaning a single pixel longer perimeter for our longest cell nucleus in the sample.

4. Feature Importance and Feature Selection

For the dimension reduction and feature selection, the Random Forest algorithm was used. The Random Forest algorithm calculates feature importance based on the reduction in impurity that a given feature is responsible for in separating the target variable, in this malignancy, into its appropriate classes. This is visualized below in Fig 13.

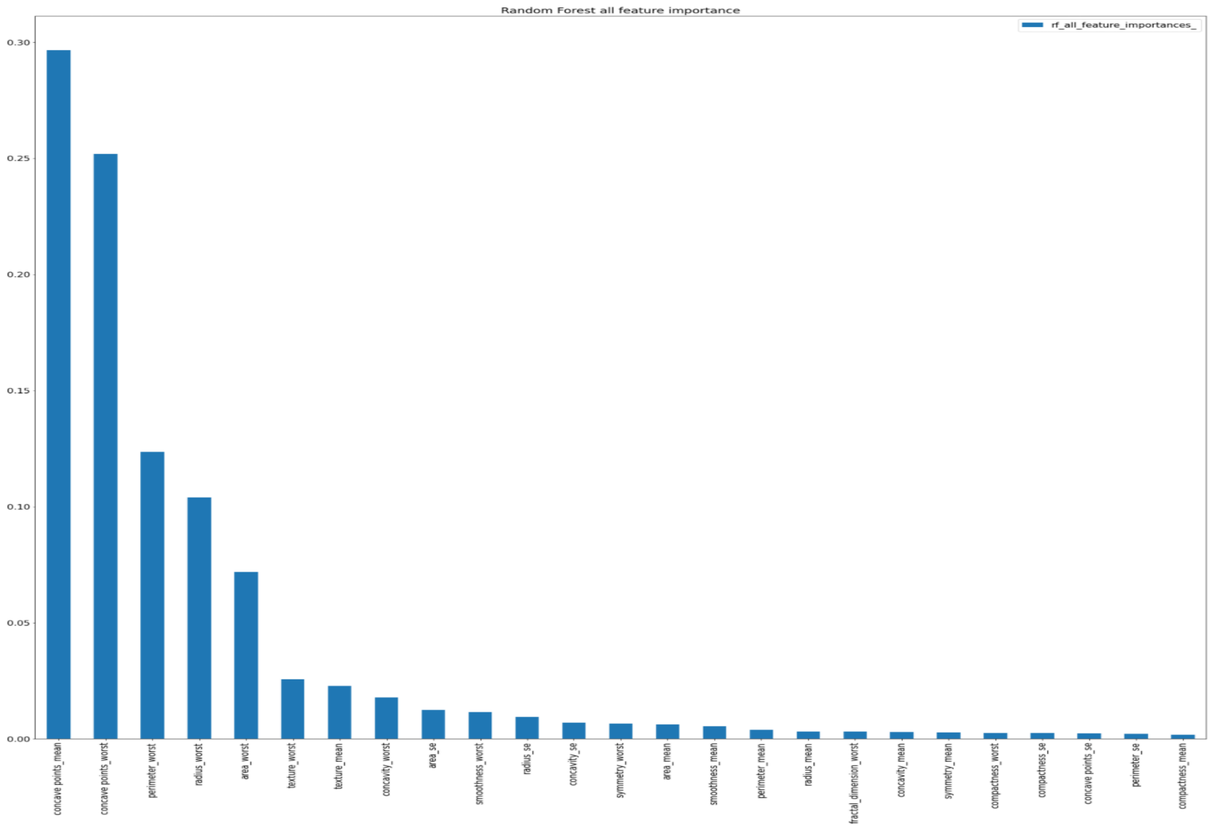


Figure 13. Feature Importance for ALL features

[1] Log odds is calculated by the equation: log-odds = e^(coefficient). Subtracting 1 from the log-odds gives the percentage increased probability.

I used this feature importance to create feature sets that I then tested for model optimization. In all, I tested 5 models that included ALL features, 20, 10 and 5 features where the predictors were selected based on feature importance solely. An additional set of 5 (I called 5v) features was tested. This was a custom set selected based on VIF in conjunction with feature importance and included the following features: perimeter_worst, smoothness_worst, area_se, concave points_worst, texture_mean.

For each set of features (ALL, 20, 10, 5 and 5v), the Random Forest model was run and the ROC_AUC score was calculated to determine which set of features were the best for the final modeling. The following table shows the ROC_AUC results of the Random Forest model for the 5 tested sets

Set	ROC_AUC Score
ALL Features	0.995443857
20 Features from RF Feature Importance	0.995149912
10 Features from RF Feature Importance	0.995149912
5 Features from RF Feature Importance	0.993239271
5v Features from VIF and Feature Importance	0.995443857

The Random Forest model for the set of “ALL features” has the same ROC_AUC score as when the “5v Features” set was used. For this reason, it was decided to continue the model testing with the two sets.

5. Modeling

● Model Selection

Two final feature sets - the ALL features and the top 5 features - were tested on 3 different models to determine the optimal combination for prediction. The 3 chosen models were: Random Forest, Gradient Boosting, and Logistic Regression.

● Tuning for best Parameters

Next, a cross-validated gridsearch was performed to optimize hyperparameters for each combination of models and features. The optimal hyperparameters were determined by the average ROC_AUC score across the cross-validated folds.

● ROC_AUC

The Receiver Operator Characteristic (ROC) curve is an evaluation tool for binary classification problems. It is a probability curve that plots the TPR(True Positive) against FPR (False Positive) at various threshold values and essentially separates the ‘signal’ from the ‘noise’. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. This was the primary metric used in grid searching to determine optimal hyperparameters. This metric was chosen because it is threshold independent and therefore gives the best general idea of how the model performs outside of a particular business scenario.

Below is a table with the results of the Modeling test for both sets of features:

Model	Optimal Features	Best Parameters	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

Below is the visualization of the ROC AUC Curve:

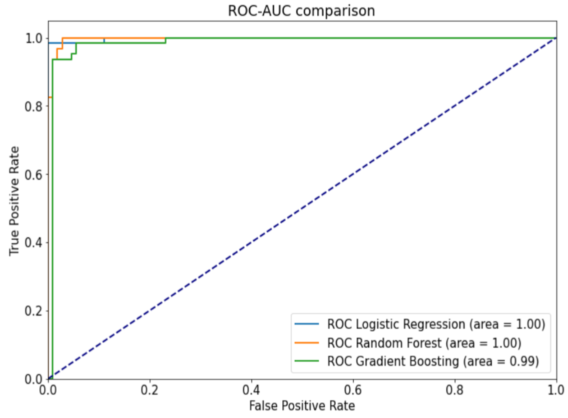


Figure 14. ROC_AUC comparison for ALL features

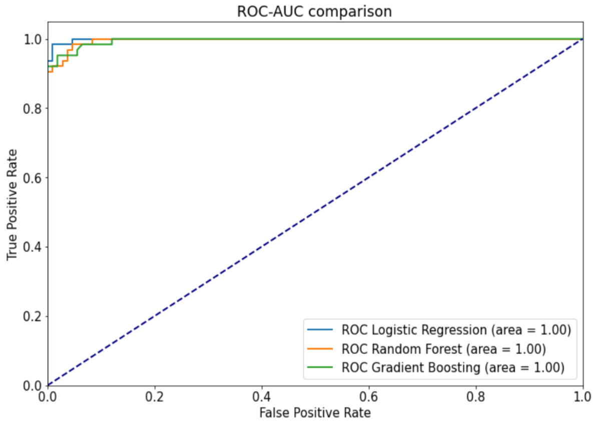


Figure 15. ROC_AUC comparison for the top 5v features

In Figures 14-15 it is clear that the Logistic Regression with the 5v features gives the best result but is only slightly ahead of the Logistic Regression model with ALL features..

Thresholding the Model

As previously stated, the ROC AUC is “threshold independent”. This means that it is determined solely from the predicted probabilities. However, in our case, we need to predict whether or not a tumor is malignant. This requires a 1 or a 0, not a probability, and so we need to choose a threshold probability above which the prediction will be malignant. The default value for the threshold is 0.5 for normalized predicted probabilities or scores in the range between 0 or 1. The problem is that the default threshold may not represent an optimal interpretation of the predicted probabilities and when the cost of one type of misclassification is more important than another type of misclassification selecting an optimal threshold is necessary. Therefore, precision and recall were plotted by threshold in Figs 14 and 15 below so an optimal threshold could be chosen.

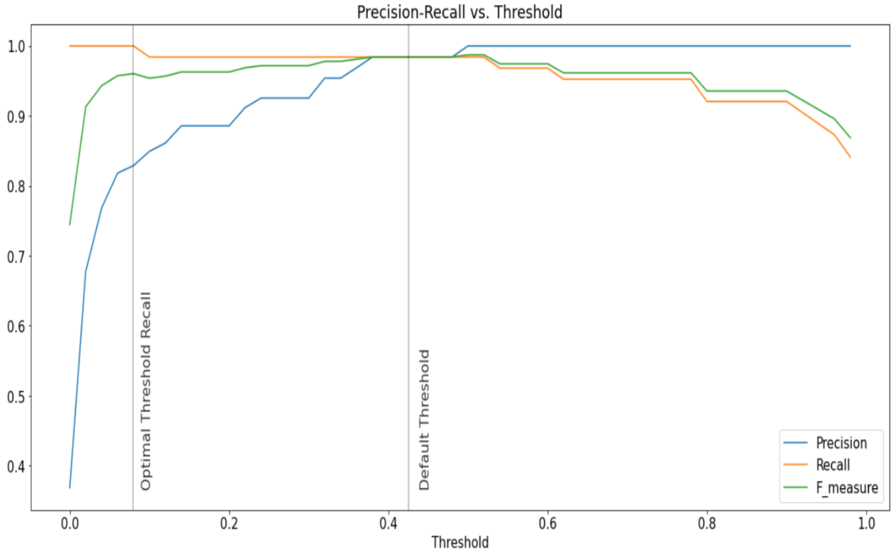


Fig 16. Thresholding for Logistic Regression on all features

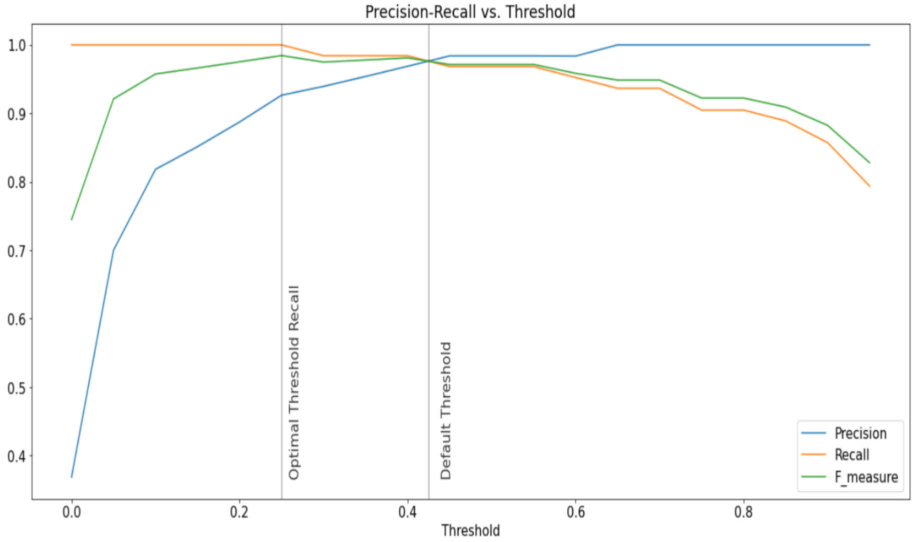


Fig 17. Thresholding for Logistic Regression on Top 5 Features

In this case, given the importance of recall - as a false negative would be far more damaging that false positive for patient outcomes - a threshold was chosen to keep recall at 1. However, this threshold might be adjusted based on further conversation with hospital stakeholders about the importance of reducing false negatives as well. As we can see, the “5v feature” model is able to maintain a recall of 1 at a threshold of .25 thus permitting a higher precision than for the “ALL features” model which has a threshold of 0.08.

This comparison becomes clearer in the classification reports below:

Classification Report

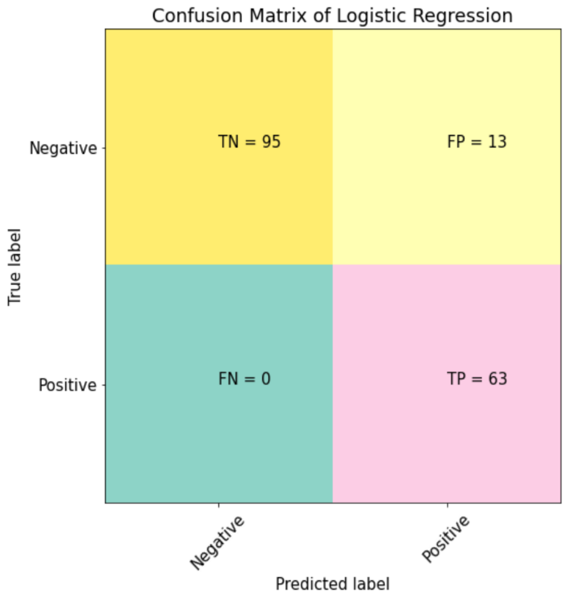
All Features				
	precision	**recall**	**f1-score**	**support**
0.0	1.00	0.88	0.94	108
1.0	**0.83**	1.00	0.91	63
accuracy			0.92	171
macro avg	0.91	0.94	0.92	171
weighted avg	0.94	0.92	0.93	171

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

From the classification report we can see that, while recall for both is 100%, the precision of the 5v features of 93 percent is much higher than the accuracy of ALL features which is 83 percent.

Confusion Matrix

A confusion matrix is a summarized table of the number of correct and incorrect predictions (or actual and predicted values) yielded by a classifier (or classification model) for binary classification tasks. Below are the visualizations of the Confusion Matrix for the 2 final modeling sets where the threshold is adjusted for a recall equal to 1.



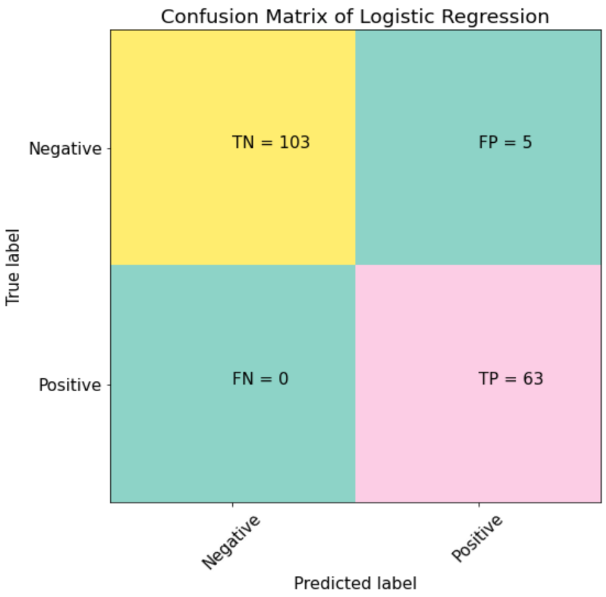


Figure 18. Confusion Matrix for ALL features

Figure 19. Confusion matrix for the top 5v features

From the Confusion Matrix visualization it is clear that when the false negatives are zero (recall=1) the false positive for ALL features model is 13 while for 5v features is 5. This result shows again that the 5v feature logistic model has a better outcome.

6. Conclusion

The Breast Cancer Prediction Wisconsin dataset consists of information about tumor features derived from the 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 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. When predicting breast cancer, minimizing the false negatives are crucial. A false negative would be in our case telling a patient that does not have cancer when in reality they have it.

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. Additionally, 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.

References

1. Lifetime Risk (Percent) of Dying from Cancer by Site and Race/Ethnicity: Females, Total US, 2014-2016 (Table 1.19). https://seer.cancer.gov/csr/1975_2016/results_merged/topic_lifetime_risk.pdf. 2019. Accessed July 31, 2019.
2. CA: A Cancer Journal for Clinicians. (n.d.). American Cancer Society Journals. https://acsjournals.onlinelibrary.wiley.com/journal/15424863
3. B.B. Mandelbrot. The fractal geometry of nature. Freeman and Company, New York, NY, 1977.

Acknowledgements

I would like to thank the following people for their support and assistance in helping with this project.

Benjamin Bell, my mentor, for his guidance.
Steve Higthtower, for all the support.
Eduardo Gutarra, for his invaluable help in Github and IT issues.
Herman Jaramillo, for his help in understanding better some statistics concepts.
Hallie Mccolongue a fellow data science student, for her company in this journey and excellent discussions.