Breast Cancer Classification Report

Jessica Montealvo

Introduction

The human body is a unique and resilient organism. It is constantly battling and seeking balance within its systems. Much of the burden on our bodies comes from our environments, where conditions we are exposed to, such as sunlight, pollution, certain foods, and specific behaviors we maintain trigger mutations and changes in our cells. Usually, our body can handle a set number of repairs, but the longer we live and the more noxious environments and behaviors we frequent, mutations start to accumulate. Cancer happens when the accumulation of mutations exceeds the body's ability to repair and control the cell cycle, meaning that cells do not die but instead continue to divide uncontrollably.

Last year, breast cancer took the lead in the number of new cases reported, with 2.26 million diagnoses. Although prominent, breast cancer is not as deadly as other cancers like lung, colon, liver, and stomach, mainly because it benefits from early detection. Early detection strategies have been implemented throughout the years and rely on both clinical and self breast exams as well as mammogram screening. If a lump or abnormality is detected, a biopsy will be taken to assess whether the cells are cancerous or not. The biopsy will result in a diagnosis, and the next steps will be dictated from there. It is important to note that earlier-stage cancers can be treated more effectively, which readily contributes to survival. So the goal is always to identify cancer early.

Problem Statement

In this project, I will use the *Breast Cancer Wisconsin (Diagnostic) Data Set* to train a model to predict whether a tumor is benign or malignant.

The Data

The data set used for this project can be found on the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29). The data describes 569 cases of digitized images of biopsies, in particular a fine needle aspirate (FNA) of breast tissue, when breast cancer was suspected. The features in this data set are measurements derived from the images and describe the cell nuclei in the tissue. The data documentation describes ten measurements gathered, with four significant digits, for each cell nucleus:

a) radius (mean of distances from the 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)

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were also computed for each image. The resulting data set contains 30 features per case and a diagnosis label (malignant = cancer, benign = not cancer).

Data Wrangling

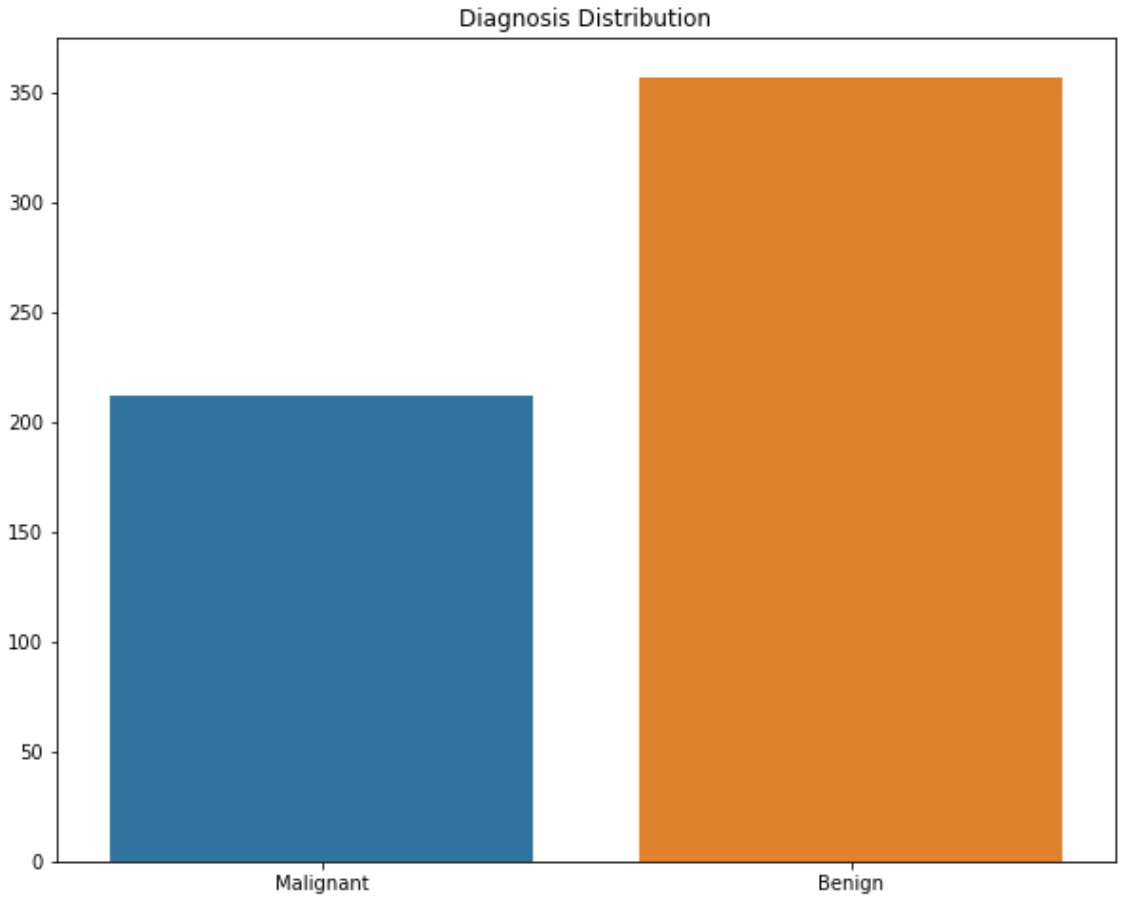
This particular data set did not require much manipulation. There were no missing values, but depending on the source, the ID column can be dropped, and the diagnosis column can be translated to (malignant = 1, benign = 0) in preparation for the EDA.

Exploration Data Analysis

The first thing to explore is the distribution of our label. The following graph shows that our data set has two classes (cancer/ not cancer). The classes are not balanced. For this reason, I will upsample in my training data set to avoid my models predicting correctly solely based on the fact that there are more *Not Cancer* instances.

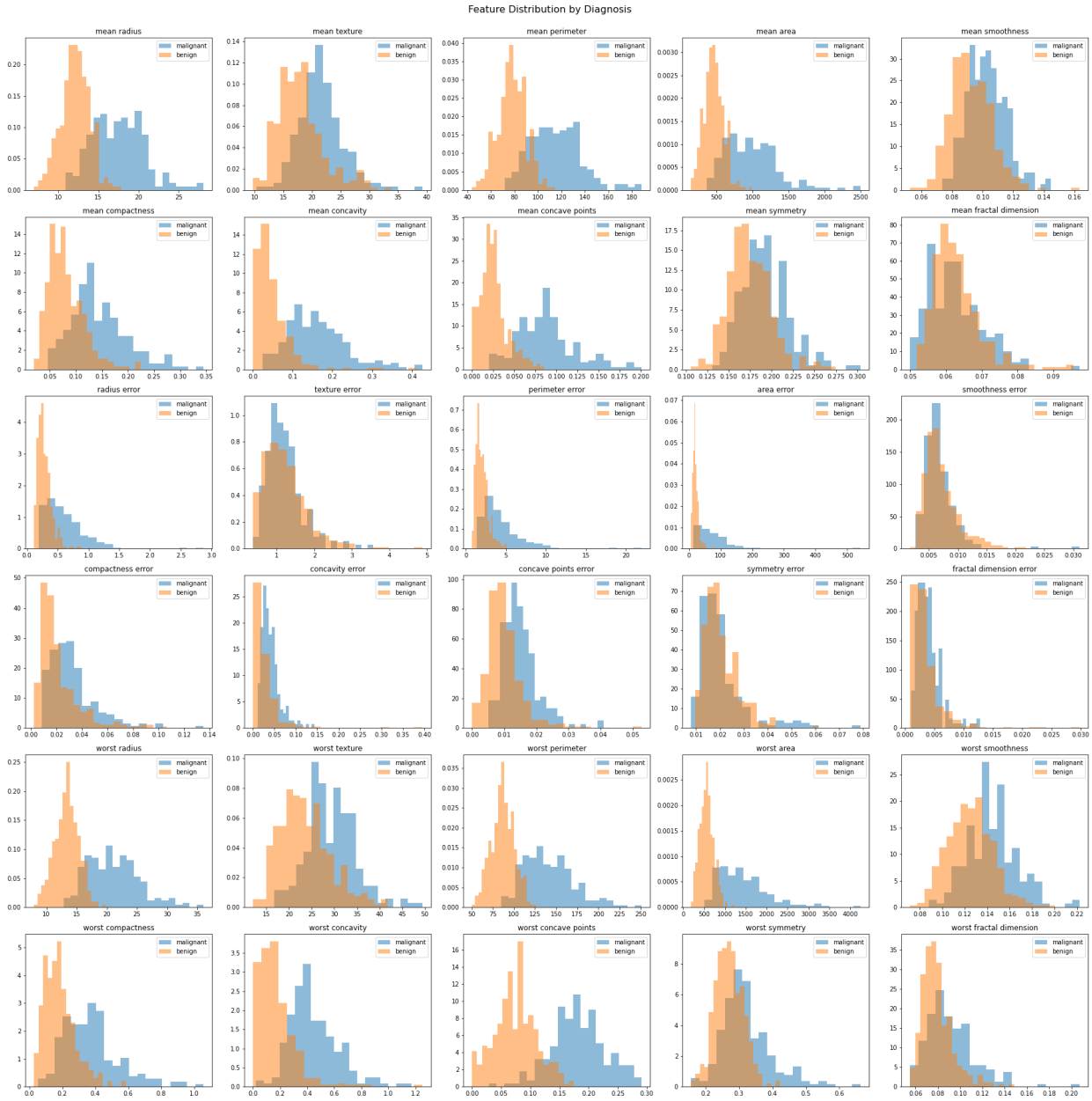
Not Cancer (benign): 62.74%

Cancer (malignant): 37.26%

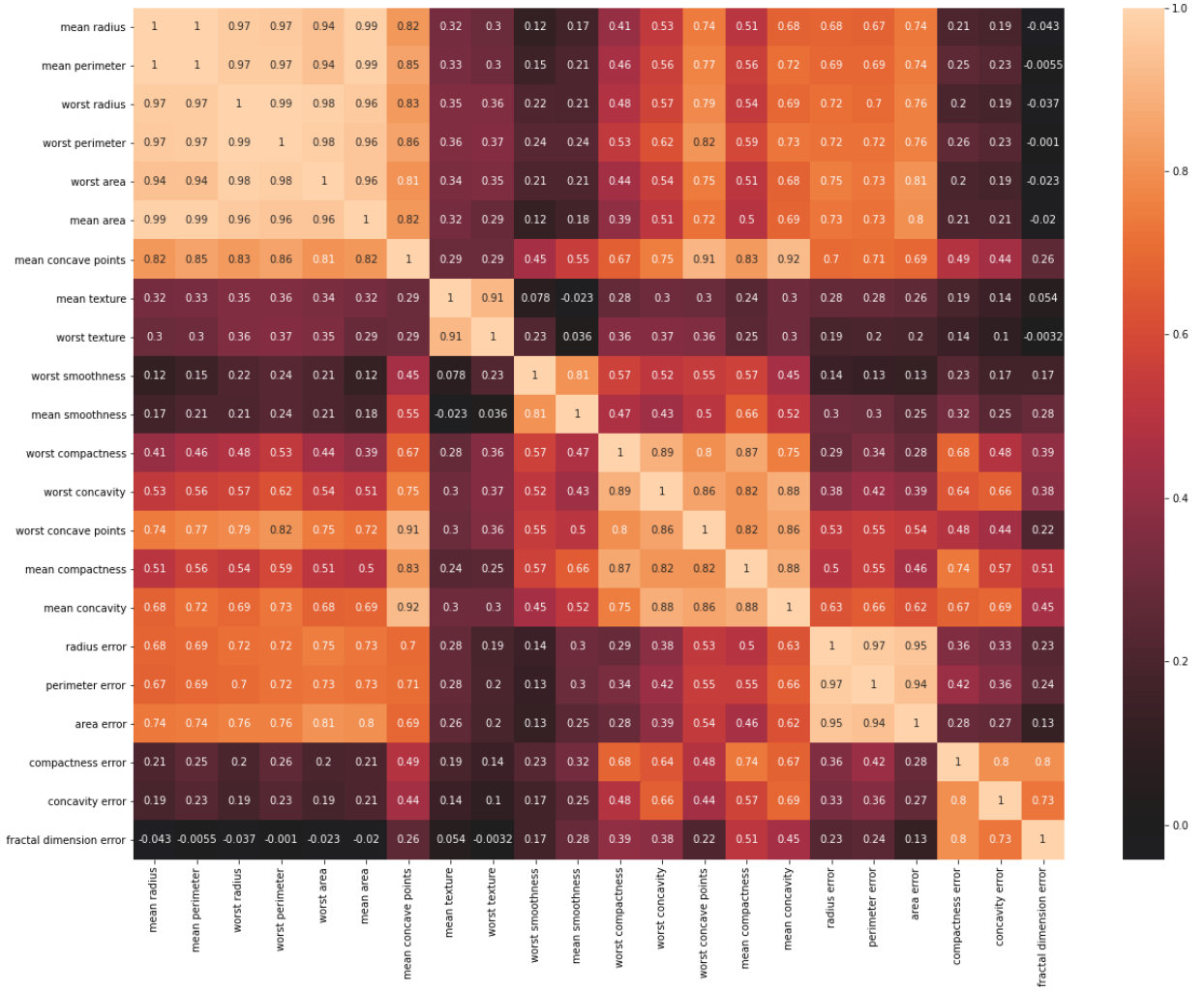


The following image shows the distribution of all 30 features by diagnosis. Here we begin to observe some trends and features characteristics that can be used to describe a diagnosis.

For instance, according to the following image, cancer cells have less compact nuclei; accordingly, they are observed on the larger portion of the spectrum for radius, mean texture, mean perimeter, area, concavity, and concave points.



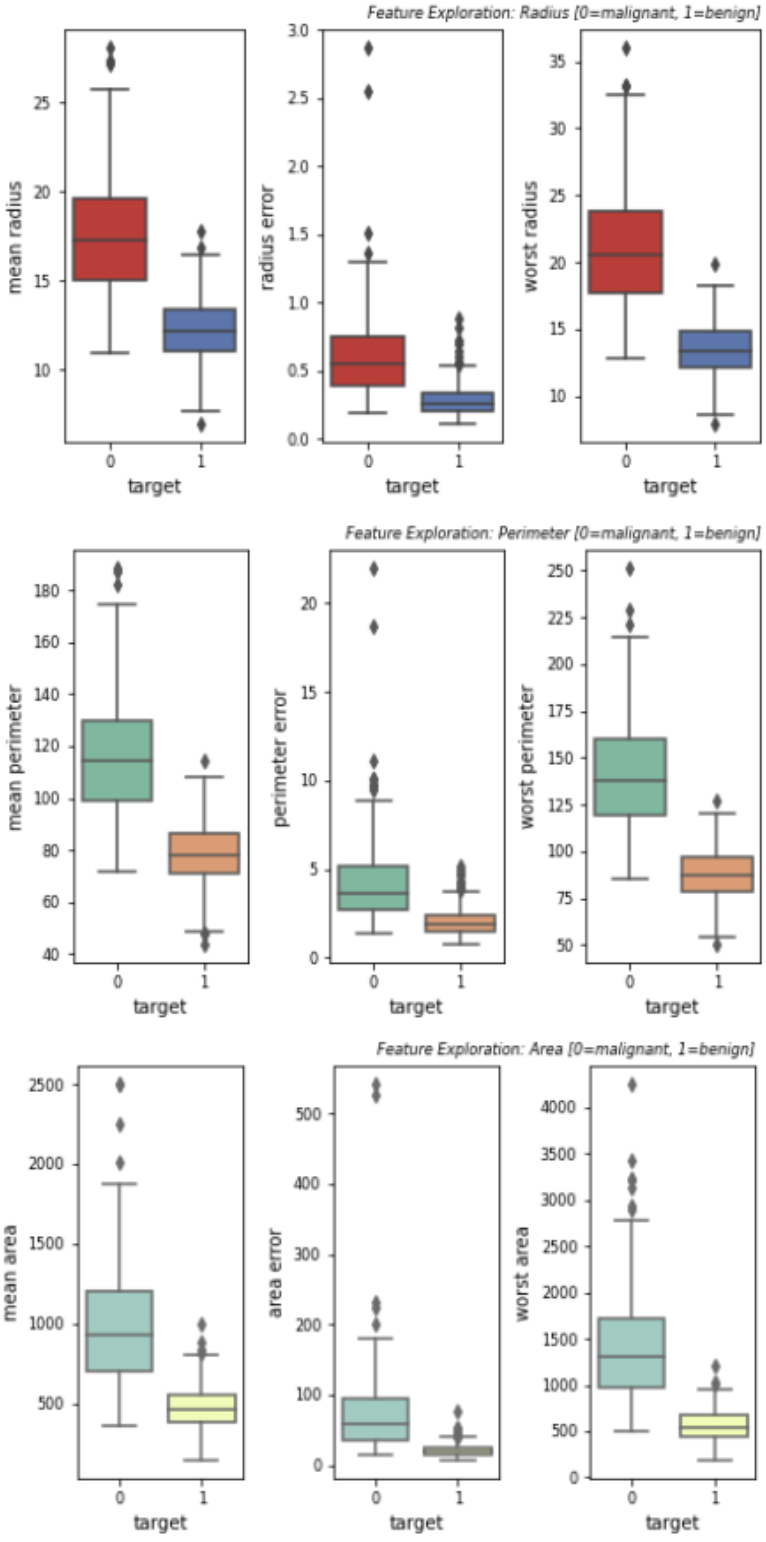
Because there is a notable overlap across the sets of measurements, the next thing to look at is a correlation matrix to visualize any potential correlations or collinearity.



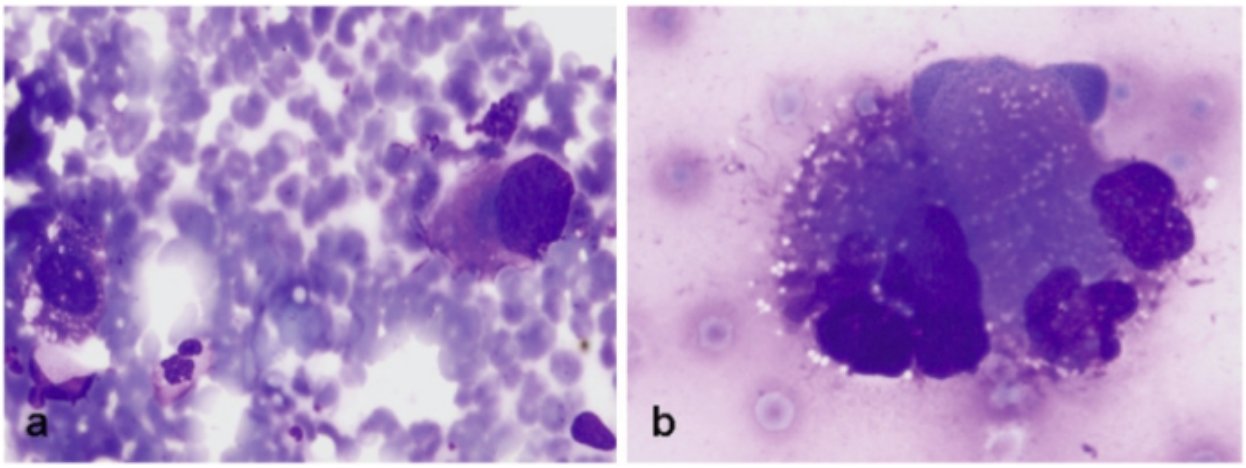
As evidenced by the correlation matrix, namely by the peach-colored squares in the matrix, in many instances, the *mean*, *worst*, and *error* measurements for each category correlate to each other.

Furthermore, we can observe that many of the measurements go hand in hand. To focus on a few correlated features - a larger area would inherently result in a larger perimeter and a proportional radius. Because they correlate to each other and overlap in their information, it is possible that one of these measurements could provide an accurate classification. Features that go hand in hand are known as collinear features. They will

typically be features that describe the same phenomenon, in a way, repetitive, which may confuse or mislead a model. However, this does not mean that they will not contribute to a model in different ways to provide a more accurate prediction.



Looking at the box plots for the distribution of radius, perimeter, and area, although the y axis of each measurement is different, their spread is remarkably consistent. In effect, if we hone in on their interactions in the correlation matrix shown previously, we see that mean radius and mean perimeter are 100 percent correlated. In comparison, the mean area is 99 percent correlated to mean radius and mean perimeter alike. The error measurements fall between 94-100 correlation amongst each other and are correlated 65-81percent with the mean and worst measurements. The extremely high correlations between these nine features cast doubt on whether each of these features will positively impact and inform our predictive model.



For some additional context, the above image is an example of the cytology of a fine needle aspirate. This image was published in a research paper by Muzumder et al. in the Current Oncology journal and is an example of the cell nuclei described in the data set we are exploring. In this case, we can see the contrast between normal cells and malignant, cancerous cells. In line with our data exploration, malignant cells are generally larger, which would indicate a larger radius, perimeter, and area. Figure b represents an example of a zoomed-in cancer cell. These observations do not lose power even though the mean was taken across the sample, presumably, because even in an earlier cancer stage, a significant quantity of cancerous cells will be present in any given biopsy.

Considering the significant potential for collinear features, I used various feature selection tools to create multiple feature datasets, paring down the number of characteristics and testing them to find the best one for breast cancer classification.

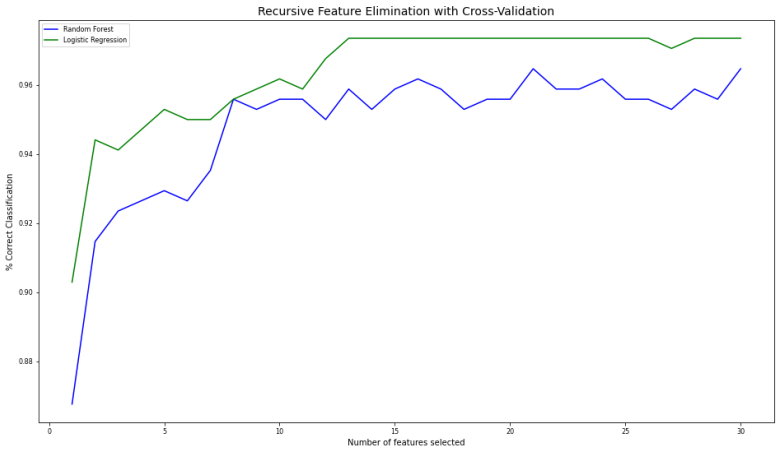
Feature Selection

It is important to note that the thought process of going through an extensive feature selection process was to ensure that we would be able to find a balance between the number of features and the performance of our model. The range in the number of features tested was from 10 to 30 components.

Given the collinearity inherent in this dataset, it seemed imperative to reduce the feature set to see if that could lead to improvements in modeling. To arrive at the best feature set, I used six different methods to arrive at the following feature sets.

1) *Select K Best* is a univariate feature selection tool that selects features based on univariate statistical tests. In this case, we tested f_classif, the computation of the ANOVA F-value for the recorded measurements. SelectKbest will keep any features where the variance observed between the means of two populations are significantly different.

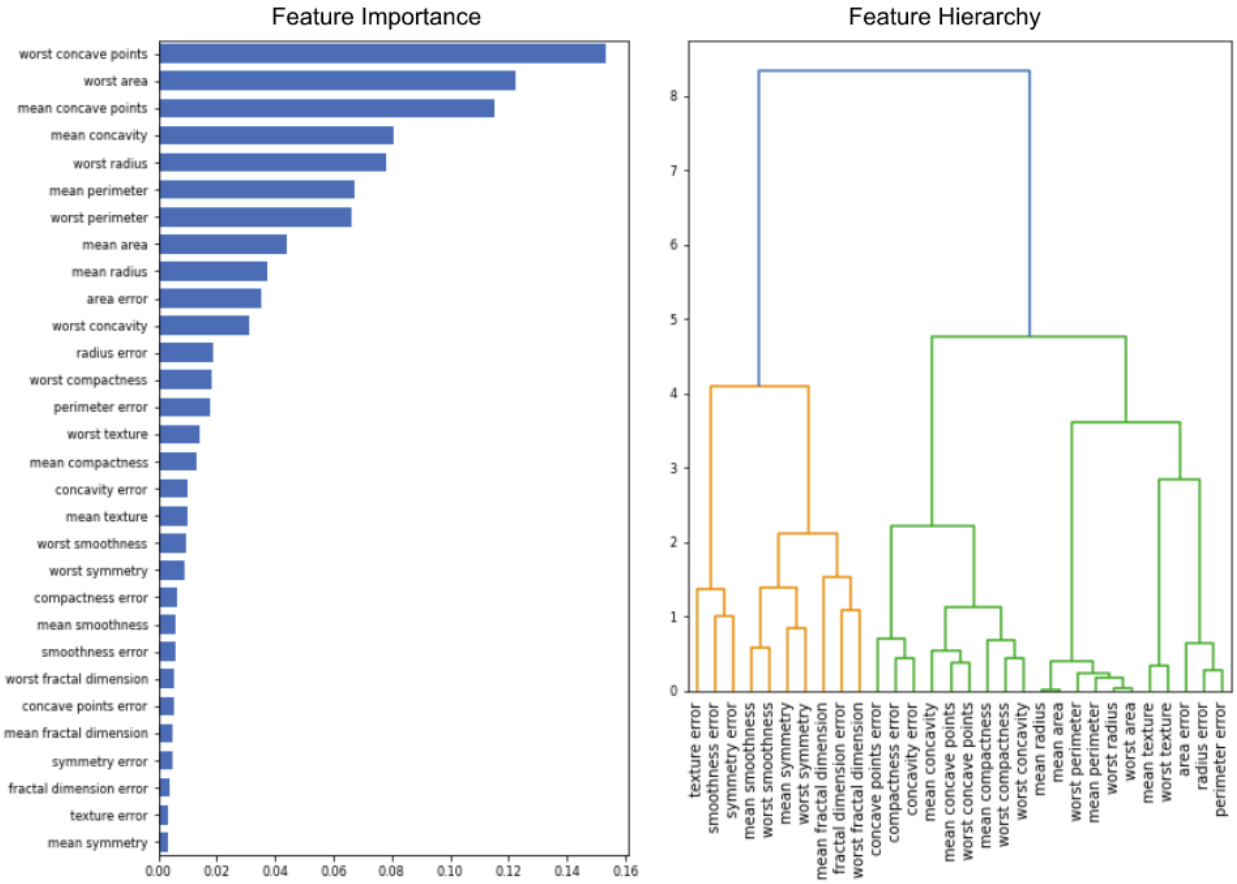
2) *Recursive Feature Elimination* uses a random forest estimator to assign weights while considering different permutations of features. The features that are least important are pruned from the data set. RFECV also performs a cross-validation loop, where the training data is segmented in different batches to arrive at the optional number of features.



3) *LassoCV* creates a linear model and continuously fits the data along a path. This estimator limits the sum of the absolute values of the model parameters. So as to

not exceed the upper limit, this method applies a shrinking or regularization process where it penalizes the coefficients of the regression variables, reducing some of them to zero. The Lasso feature selection also relies on cross-validation to select the most impactful features.

4) The *RF Feature Importance* method used in this project relied on permutation importance to rank the features. Following this, hierarchical clustering of the features and their relationships (Spearman rank-order correlations) subsequently only keeps the features at a threshold that is far enough away from each other to ensure that the weight and importance are not duplicated. Effectively, only selecting one feature per cluster.



5) Interestingly enough, two-thirds of the data (error and worst) are computed from the mean measurements of the nuclei, but they do seem to contribute to a more accurate prediction. The possibility that only ten features (*Mean Features*) could inform a precise model was also explored by testing a feature set that only contained the observed mean measurements.

6) Lastly, a subset of features that were at least 80 percent correlated to each other was dropped from the original data set, leaving behind a somewhat arbitrary feature set to test (*Less Correlated Features*).

Feature Set	Selected based on	Number of Features	Baseline (LogisticRegression) ROC AUC Score
Random Forest Feature Importance	Hierarchy based on Correlation and node distance	14	0.9975
Recursive Feature Elimination with Cross-Validation	RFECV	18	0.9974
Original	Kept all features	30	0.9972
Lasso CV	LassoCV/LogReg	22	0.9961
SelectKBest	SelectKBest	15	0.9906
Less Correlated Features	Removed correlated features (threshold=0.8)	16	0.9896
Mean Features	Removed error and worst feature sets	10	0.9894

Although the ROC AUC score differences seem minimal, they are incredibly impactful since the prediction/classification dictates whether a person is diagnosed with a life-threatening disease that will result in many hours of excruciating therapy, potentially multiple surgeries, and thousands of dollars in medical expenses.

With less than half of the features, the Random Forest Feature Importance data set performed the best before any fine-tuning.

Modeling

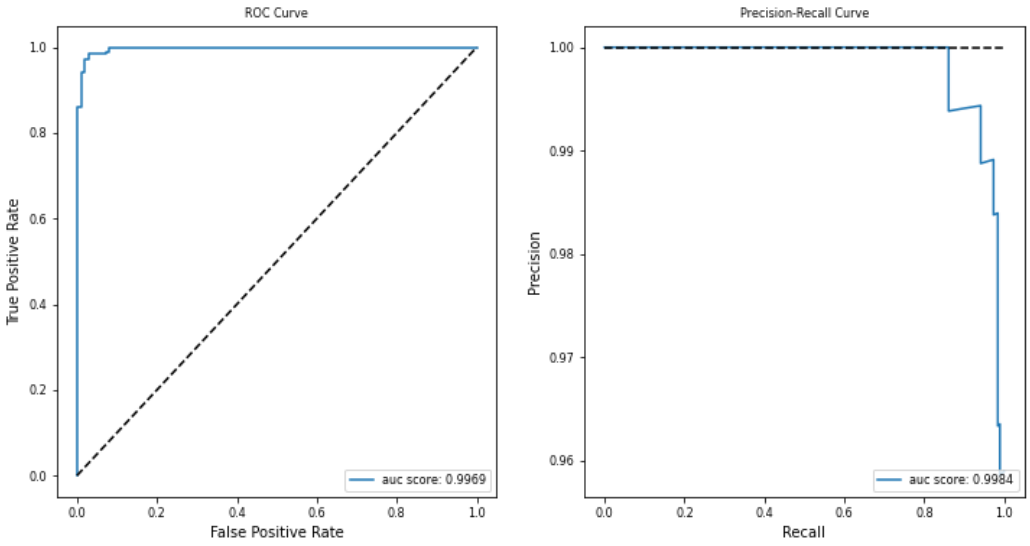
Given the baseline data from the feature selection process, we observe that the most accurate data set is the one selected from the Random Forest Feature Importance Ranking and Selection.

The selected data set was used to build three different models. Grid Search Cross Validation was then applied to select the best parameters for each model.

Model	Best Score	Best Parameters	ROC AUC Score
LogisticRegression	0.9676	{'C': 1.0, 'max_iter': 100, 'penalty': 'l1', 'solver': 'liblinear'}	0.9969
Gradient Boosting	0.9529	{'learning_rate': 0.5, 'max_depth': 2, 'n_estimators': 500}	0.9915
Random Forest	0.9471	{'criterion': 'entropy', 'max_depth': 5, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 15}	0.9906

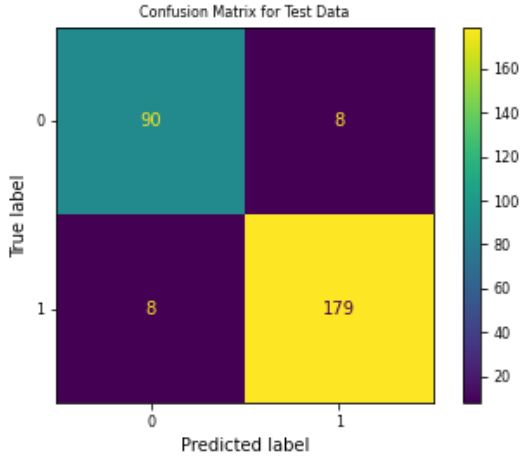
The logistic regression model performed best. The fine-tuned model did exceptionally well on the test data. The ROC AUC score was 0.9969, while the precision/recall score was

calculated at 0.9984. Close to a perfect model.



However, we need this model to be useful, not perfect. Bearing in mind that 0 is the malignant class and 1 is the benign class, the classification report and confusion matrix show that we can further impact the predictions made by this model.

	precision	recall	f1-score	support
0	0.92	0.92	0.92	98
1	0.96	0.96	0.96	187
accuracy			0.94	285
macro avg	0.94	0.94	0.94	285
weighted avg	0.94	0.94	0.94	285



In order to arrive at the best instance of this model, I decided to calculate the Fbeta measure to determine a decision threshold that would avoid any false-negative predictions.

The Fbeta measure is calculated using the *Precision* and the *Recall* of a model. However, it uses a coefficient, *beta*, to add weight to either parameter.

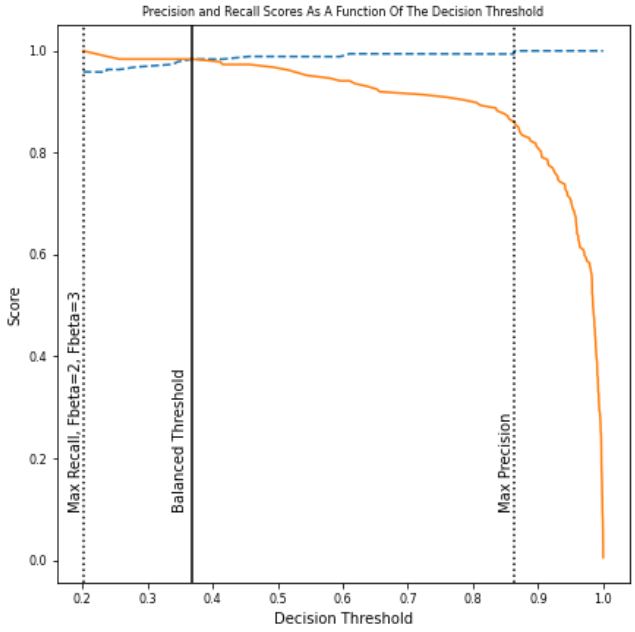
*Fbeta = ((1 + beta^2) * Precision * Recall) / (beta^2 * Precision + Recall)*

Precision summarizes the correct predictions for the positive class, while *Recall* outlines the percentage of correct predictions for the positive class considering all the possible positive predictions. Prioritizing *Precision* would minimize false-positive errors, which in this project are less harmful since diagnosing a human with cancer while not having cancer will not be detrimental. On the other hand, and more importantly, this project must prioritize *Recall* to minimize the false-negative errors. These would be the instances where a person walks away with a “not cancer” diagnosis, all the while harvesting a malignant tumor.

A beta value that is less than one will place more weight on *Precision*. A beta value of one balances the weight between *Precision* and *Recall*. A beta greater than one will lessen the weight on *Precision* and will increase the weight on *Recall*.

The proposed threshold would provide a recall of 1, meaning we would avoid false negatives. Calculating Fbeta with a beta of two and then three, both arrived at the same threshold: 0.201174

Precision	Recall	Threshold	Fbeta = 2	Fbeta = 3
0.958974	1.000000	0.201174	0.991516	0.995740



Our model is a regression model, meaning we don't have clean-cut predictions (yes/no, 0/1). However, the thresholds calculated will help us select the boundary of where it should be considered yes or no.

Below are the classification reports based on two thresholds, the one where we maximize recall and the one where we balance the importance of recall and precision.

```
Classification Report - Max Recall
              precision    recall  f1-score   support

   malignant       1.00      0.92      0.96        98
      benign       0.96      1.00      0.98       187

    accuracy                           0.97       285
   macro avg       0.98      0.96      0.97       285
weighted avg       0.97      0.97      0.97       285

Classification Report - Balanced Recall/Precision
              precision    recall  f1-score   support

   malignant       0.96      0.97      0.96        98
      benign       0.98      0.98      0.98       187

    accuracy                           0.98       285
   macro avg       0.97      0.97      0.97       285
weighted avg       0.98      0.98      0.98       285
```

Conclusion

In conclusion, through this project, we selected the best 14 features out of 30 that would inform a tuned logistic regression model to predict whether the cells are cancerous or not while avoiding any erroneous negative diagnoses.

A thorough feature selection process only kept the most essential and informative features, effectively avoiding collinear features. I assessed Logistic Regression, Random Forest, and Gradient Boosting models for the best ROC AUC. After fine-tuning the most accurate model, Logistic Regression, I selected a threshold (0.201174) that would provide the least number of false negatives with a recall of 1. Any prediction below 0.201174 will be considered cancerous, while anything above will be classified as benign.

A potential expansion to this project would be to analyze the images directly and derive the nuclei measurements via an image processing model to absolve a pathologist from having to read and calculate each specimen.

Appendix A

The following table showcases the presence of a feature within a test feature set.

Feature	Baseline	Select Kbest	RFECV	Lasso CV	RF Feat Importance	Mean Features	Less Correlated Features
mean radius	yes	yes	no	yes	yes	yes	yes
mean texture	yes	no	no	yes	yes	yes	yes
mean perimeter	yes	yes	yes	yes	no	yes	no
mean area	yes	yes	yes	no	no	yes	no
mean smoothness	yes	no	no	no	yes	yes	yes
mean compactness	yes	yes	yes	yes	yes	yes	yes
mean concavity	yes	yes	yes	no	yes	yes	no
mean concave points	yes	yes	yes	yes	no	yes	no
mean symmetry	yes	no	no	yes	yes	yes	yes
mean fractal dimension	yes	no	no	yes	yes	yes	yes
radius error	yes	yes	yes	yes	yes	no	yes
texture error	yes	no	no	yes	yes	no	yes
perimeter error	yes	yes	no	no	no	no	no
area error	yes	yes	yes	yes	no	no	no
smoothness error	yes	no	no	yes	yes	no	yes
compactness error	yes	no	no	yes	yes	no	yes
concavity error	yes	no	no	yes	no	no	yes
concave points error	yes	no	yes	yes	no	no	yes
symmetry error	yes	no	yes	yes	yes	no	yes
fractal dimension error	yes	no	yes	yes	yes	no	yes
worst radius	yes	yes	yes	yes	no	no	no
worst texture	yes	no	yes	yes	no	no	no
worst perimeter	yes	yes	yes	no	no	no	no
worst area	yes	yes	yes	yes	no	no	no
worst smoothness	yes	no	yes	yes	no	no	no
worst compactness	yes	yes	no	yes	no	no	no
worst concavity	yes	yes	yes	yes	no	no	no
worst concave	yes	yes	yes	no	no	no	no

points							
worst symmetry	yes	no	yes	yes	no	no	yes
worst fractal dimension	yes	no	no	yes	yes	no	yes