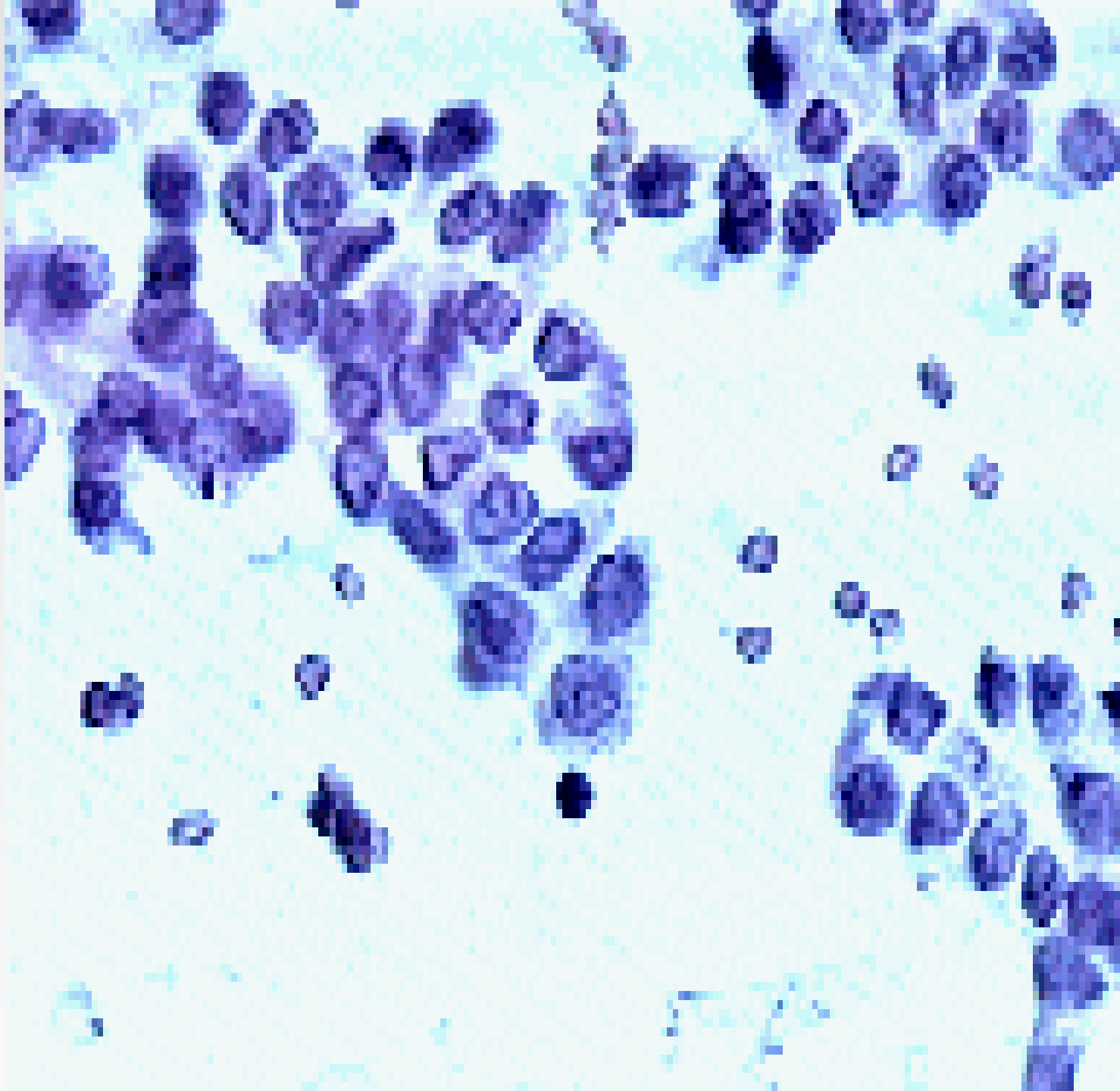
Using Machine Learning to Diagnose Breast Cancer

CHRIS WOODS

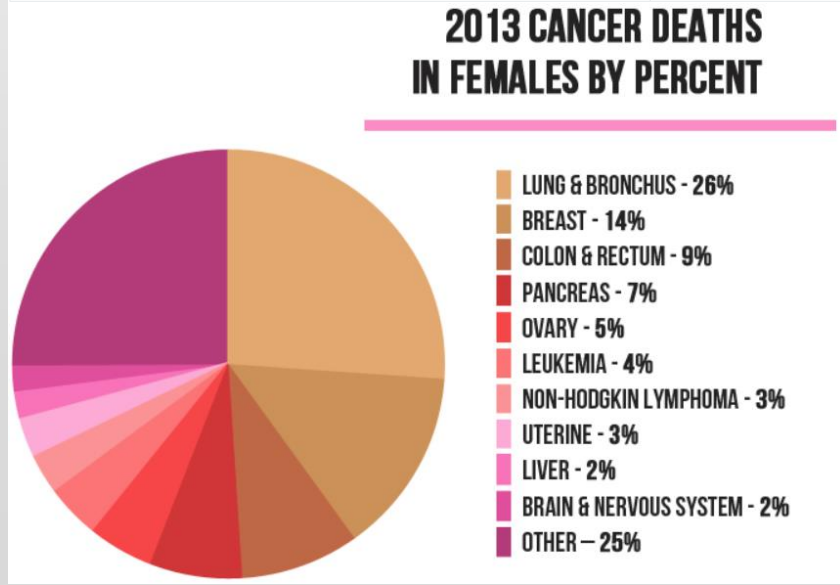


Background

Breast cancer is a serious issue in the US. Here are a few statistic from breastcancer.org:

- About 1 in 8 US women, ~12%, will develop invasive breast cancer over the course of her lifetime.
- Cancer is the #2 leading cause of death of women. Behind lung cancer, breast cancer is the second most fatal type of cancer in women
- In 2019, an estimated 268,600 new cases of invasive breast cancer are expected to be diagnosed in women in the U.S., along with 62,930 new cases of non-invasive (in situ) breast cancer.

All Females, All Ages	Percent*
1) Heart disease	22.3
2) Cancer	21.1
3) Chronic lower respiratory diseases	6.2
4) Stroke	6.1
5) Alzheimer's disease	5.7
6) Unintentional injuries	4.0
7) Diabetes	2.7
8) Influenza and pneumonia	2.3
9) Kidney disease	1.8
10) Septicemia	1.6



Problem

- Fine needle aspiration, FNA, is the most common method to biopsy a suspicious mass
- One issue with this procedure is the high rate of false negative test results. A false negative happens when a test result indicates there is no disease present when there actually is disease.
  - “However, even with needle biopsies, false negative results are not common. One study looking at nearly 1,000 core needle biopsies found a false negative result rate of 2.2%. That’s just over 2 out of 100 biopsies.”*
- This can delay diagnosis and ultimately lead to longer and more extensive treatment

*https://www.healthgrades.com/right-care/cancer/how-common-are-false-negative-biopsies?cid=t12_ccgd

Project Goal

The goal of this project is to increase the overall accuracy of breast tumor diagnosis using machine learning principles

- More specifically the rate of false negative test results.
- A false negative results could result in the treatment of the tumor being delayed six months, or possibly longer, with the cancer getting worse.
- This may result in longer more difficult treatments, or possibly unnecessary death.

Exploratory Data Analysis (EDA)

- For this project the *Breast Cancer Wisconsin (Diagnostic) Data Set*[*] was used.
  - 30 numeric features with 569 observations. 357 observations were classified as benign and 212 classified malignant.
  - Two columns were removed because they served no practical purpose.
  - There were no missing or NaN values. No further data wrangling steps were required.
- For data analysis and visualization purposes, I broke down the features into four groups for later analysis:
  1. All Data Columns
  2. Mean Data Columns
  3. Standard Error Data Columns
  4. Worst, or Largest Data Columns

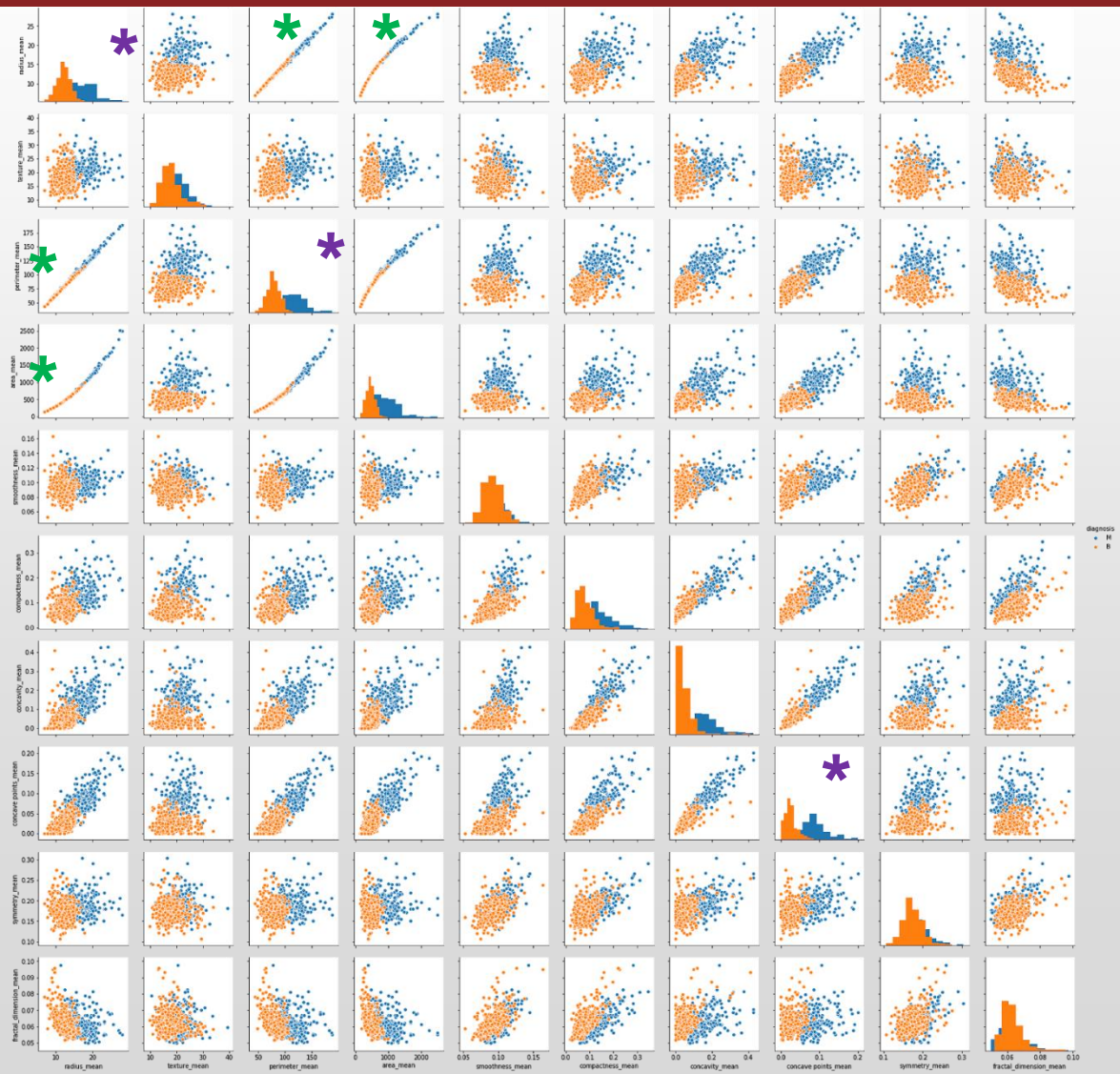
[*]https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

Exploratory Data Analysis (EDA) cont'd

- To get a sense of the basic statistics of the features, I used the **.describe()** function.
  - The major takeaway was the scale of some features are magnitudes of order larger than others.
  - Because of this I had to standardize the data for some visualizations and machine learning algorithms to make sense.
- To get an idea of the distribution of the features I chose histograms, correlation matrices, scatterplots and box plots.
  - Seaborn Pairplot – provided scatterplots and histograms of each feature compared to the others, color coded by 'diagnosis' to see the distribution between benign and malignant diagnosis.
  - Correlation Matrix – used to confirm the collinearity shown in the visual findings of the Seaborn Pairplot.
  - Box Plots – to see if there was a significant difference between the distribution of benign and malignant diagnoses for each feature.

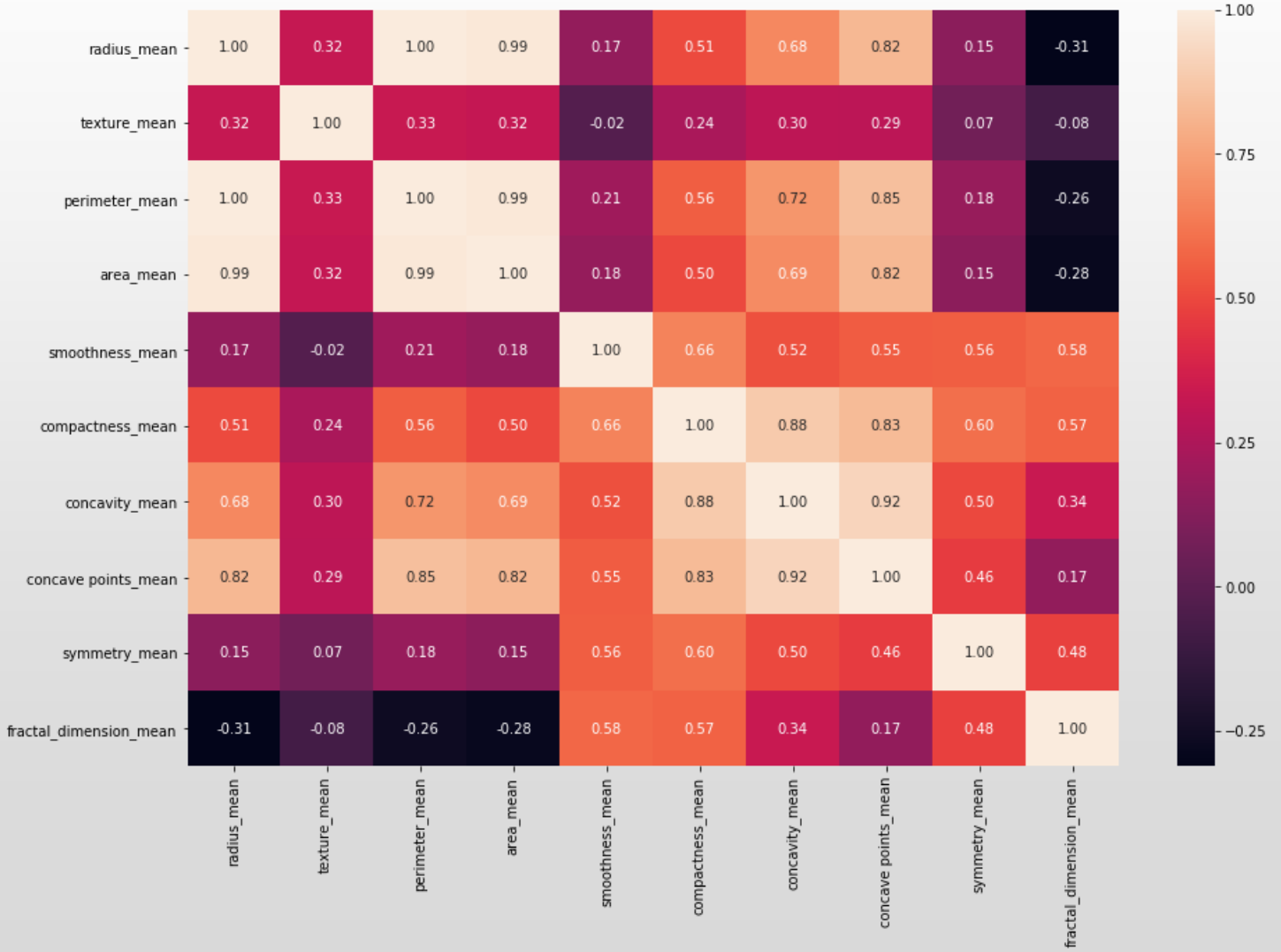
Visual Analysis – Seaborn Pairplot

- From the scatter plots we can clearly see several features are highly collinear. Radius, Perimeter and Area are obvious but others look highly correlated as well. *
- In the individual feature histograms you can start to see which variables would make better predictor features. The more divergence between the Benign and Malignant distributions the better. *



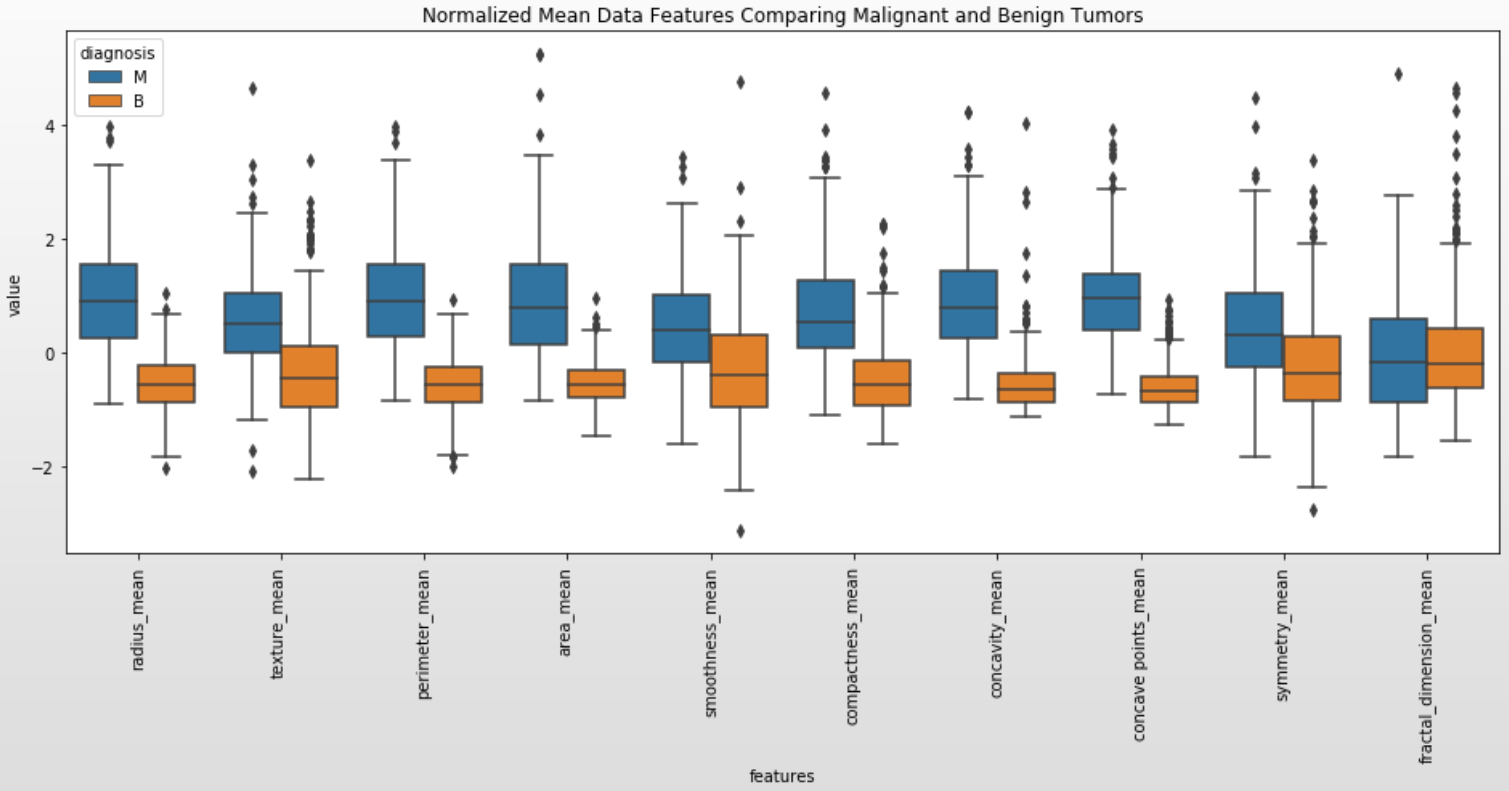
Visual Analysis – Correlation Matrix

- Radius, Perimeter, and Area are almost perfectly correlated!
- Additionally, there are many other pairs of features that are also highly correlated. Concave Points has correlation coefficients > 0.8 with five other features.



Visual Analysis – Box Plots

- Here we can see some of the same patterns as noted before. Notice how the box plots for Radius, Perimeter, and Area are very close to the same. As well as Concavity and Concave points.
- You can also see that Smoothness, Symmetry, and Fractal_Dimension will not make good predictor features because their IQR's (Interquartile Range) boxes overlap. Meaning their p-value is much higher than 0.05.



Machine Learning Analysis

- Analyzing the data we can see that a Supervised Machine Learning model is appropriate because all of the features are labelled.
- Additionally we see that the dependent variable, 'diagnosis' is binary, B (benign) or M (malignant). This makes it a Classification problem.
- For this project I started with a basic k-NN Classifier to get a baseline feel for the accuracy to expect from the model. Then follow-up using Logistic Regression, Random Forests, and Gradient Boosting in order to determine the best model.

Machine Learning Analysis cont'd

Two measures are used to rank or grade the models: Recall and AUC_ROC.

- **Recall** - Recall is the number of True Positives divided by the number of True Positives plus the number of False Negatives. Put another way, it is the number of positive predictions divided by the number of positive class values in the test data. It is also called Sensitivity or the True Positive Rate.
  - Recall was chosen as the primary scoring method because it considers False Negatives.
  - A False Negative in our model would mean that we diagnosed a woman's tumor as being benign when in fact it was malignant. We strive for a Recall as close to 100% as possible with our model.
  - The factor that will ultimately determine the best model is *Recall of the Malignant responses*. In this project we are dealing with human beings where the outcome of our model is Benign/Malignant, or Life/Death. For this reason the 'accuracy' of our prediction is not sufficient.

- **AUC-ROC** - AUC is the area under the curve of Receiver Operator Characteristic (ROC) plot False Positive Rate vs True Positive Rate at different points.
  - AUC represents a model's ability to discriminate between positive and negative classes. An area of 1.0 represents a model that made all predictions perfectly. An area of 0.5 represents a model as good as random.
  - The AUC-ROC is used to get a general sense of the overall accuracy of a model. But

Conclusion

The goal of this project was to increase the overall accuracy of breast tumor diagnosis using machine learning principles. The table summarizes the results.

	Machine Learning Algorithm Performance			
Metric	k-NN	Logistic Recression	Random Forest	Gradient Boosting
Recall	0.8941	*0.9875*	0.9375	0.9500
ROC-AUC	0.9728	*0.9986*	0.9975	0.9972

Using logistic regression, on a relatively small data sample I was to reduce the false negative rate for malignant tumor diagnosis from 2.2% to 1.25%, or a ~ 57% reduction. This a very promising first step with regard to integrating machine learning and medical diagnosis.

The data used in this project was relatively small, 569 observations. It would be interesting to see if the conclusion found here would be confirmed with additional observations, or if a different ML model would perform better than Logistic Regression.