Jonathan Daughtry

Final Report:
Analysis of Breast Cancer Tumors

Problem Statement

The problem that is being addressed in this project is the potential to recognize and predict the likelihood of a tumor being malignant in patients with breast tumors. Thankfully, research and treatments have greatly improved for breast cancer, but one of the key indicators of success is early detection. Cancer has affected almost everyone in this world, whether directly or through a relative or friend, and breast cancer is one of the most common types of cancer in women, so this is an issue that everyone can agree is very important. The stakeholders that would find this the most useful were doctors that were trying to predict the likelihood of a tumor being malignant and the patients involved.

The dataset that was used for this project is the Breast Cancer Wisconsin dataset from kaggle. This dataset has characteristics of 357 benign and 212 malignant tumors. The dataset focuses on the characteristics of the nucleus of the tumor cells, including the radius, texture, perimeter, smoothness, etc.

The goal of this project was to determine which of the characteristics that are given are the most important in determining if a tumor is benign. I used four factors to determine how well the models performed:

1. High accuracy - A model with high accuracy correctly predicts what tumors are benign and what tumors are malignant.
2. Low false negative percentage - The worst case scenario would be that someone's tumor is malignant and it was incorrectly diagnosed as benign. This type of mistake could be life threatening
3. High recall for positive predictions - Recall is the number of tumors diagnosed as malignant divided by the actual number of malignant tumors. We want this to be as close to 1 as possible in order to make sure all of the malignant tumors were correctly diagnosed.
4. High area under the ROC curve (AUC) - AUC provides an aggregate measure of performance across all possible classification thresholds. I wanted to see if doctors adjusted their threshold for what is considered a malignant diagnosis, how well the models would perform.

I used multiple classification models, including logistic regression, which gave me my best model score of:

1. 99.5% accuracy
2. 2 out of 114 (1.75%) false negatives
3. A recall of 95%
4. An AUC of 99.5%

Data Wrangling

The raw dataset had 469 rows, each a different sample of a breast tumor, and 32 columns of features. The dataset was already considered very clean before processing, containing no missing values or duplicate data. The features were all continuous variables that described the nucleus of the cancer cells, and they were divided into three major categories. They were the mean, the standard error, and the worst (largest) of each feature.
The ten real-valued features are 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)

No values had to be initially removed due to the dataset being clean initially.

Exploratory Data Analysis of Continuous Variables

As I began exploring the data, several questions that I tried to answer included: How strongly correlated are the variables? Particularly, how strongly correlated are the mean, standard error, and "worst" of each variable? Are they all needed? Do we want to keep all of the variables for the final modeling? Are there any variables that are more strongly correlated to the diagnosis?

My main focus in the beginning was comparing the mean values to whether they were malignant or benign to see if there was a noticeable difference. I eliminated perimeter and area mean from consideration because they had a direct correlation with radius. I checked to see the count of malignant cases versus benign cases for each variable.

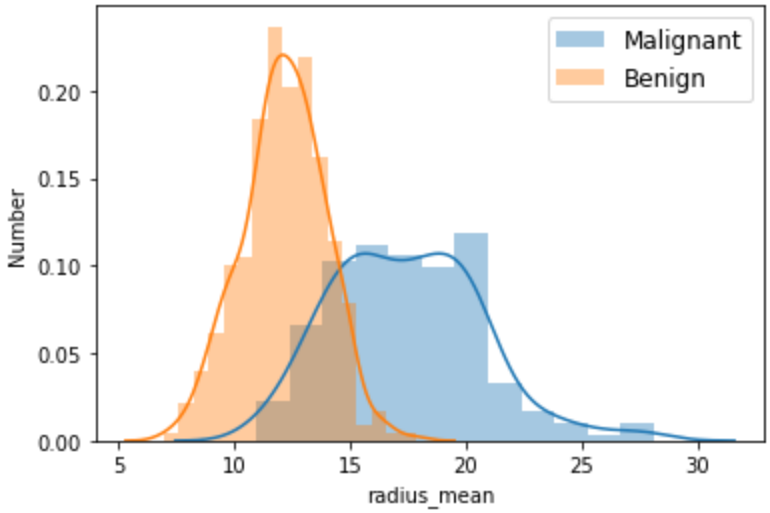


Figure 1: Count of benign vs. malignant tumors with different radius means

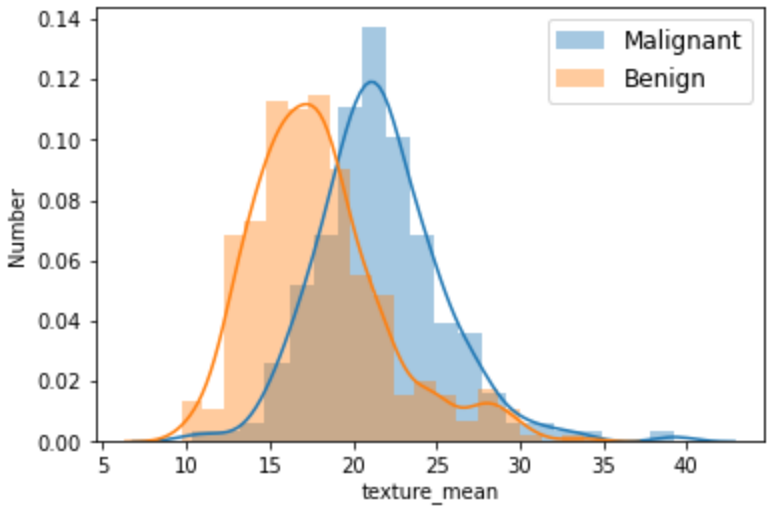


Figure 2: Count of benign vs. malignant tumors with different texture means

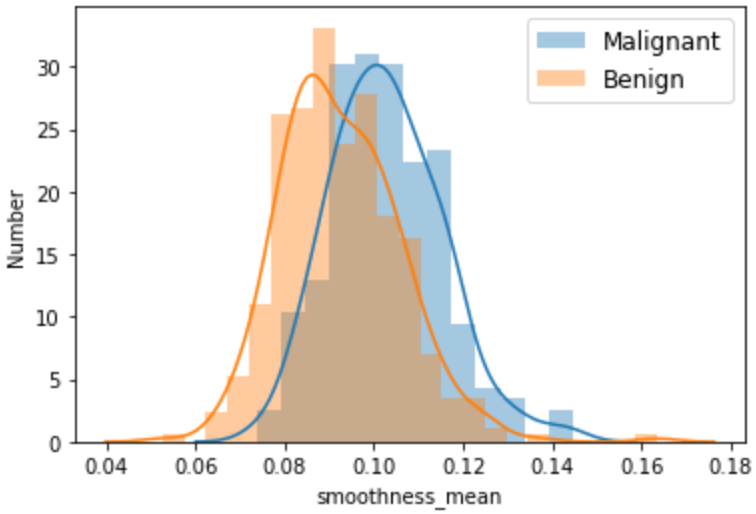


Figure 3: Count of benign vs. malignant tumors with different smoothness means

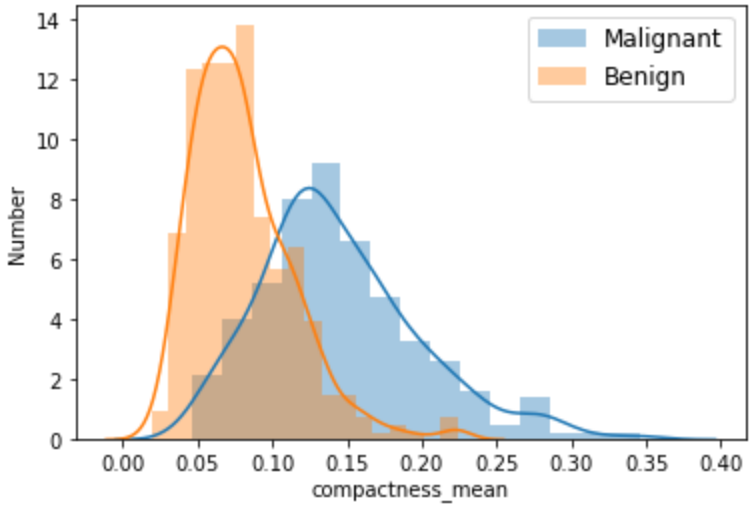


Figure 4: Count of benign vs. malignant tumors with different compactness means

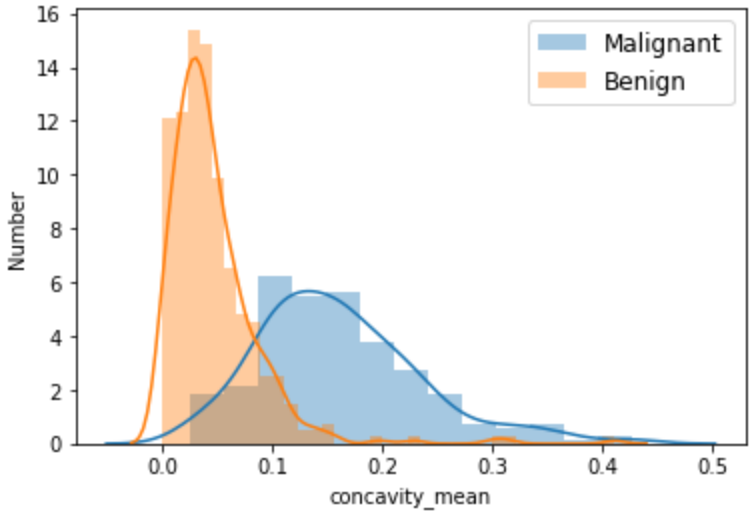


Figure 5: Count of benign vs. malignant tumors with different concavity means

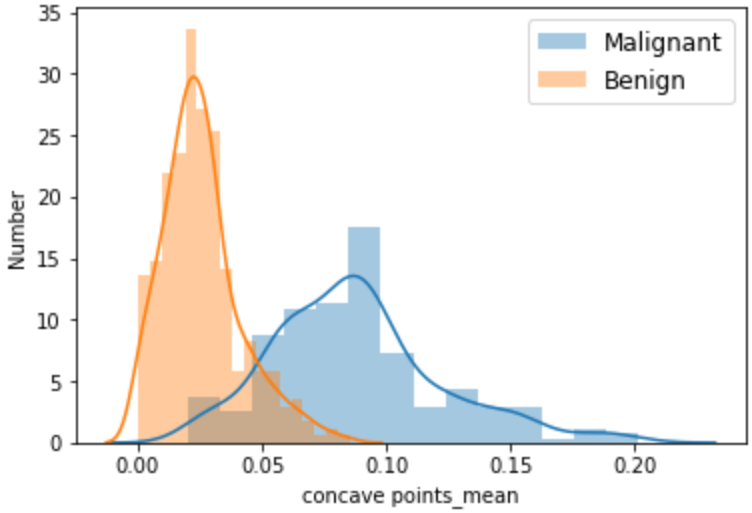


Figure 6: Count of benign vs. malignant tumors with different concave points means

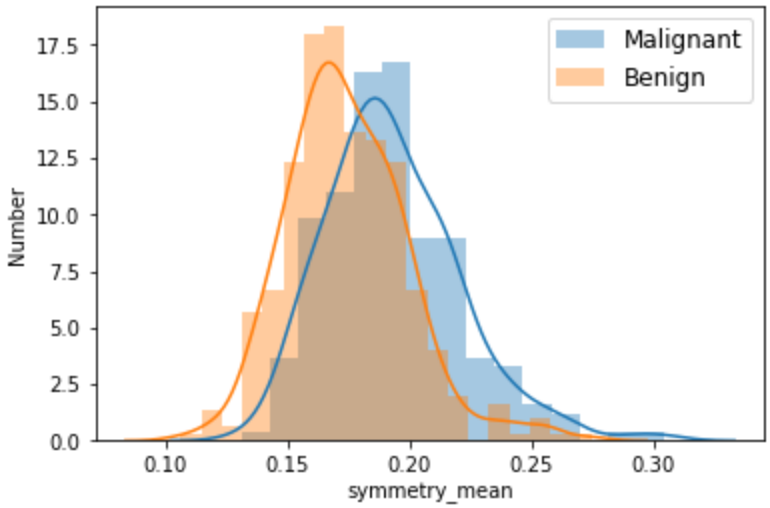


Figure 7: Count of benign vs. malignant tumors with different symmetry means

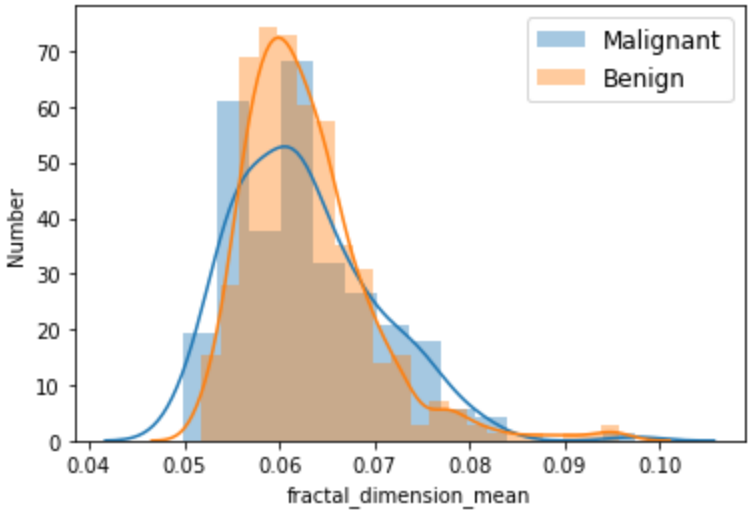


Figure 8: Count of benign vs. malignant tumors with different symmetry means

For most of the graphs, there is a noticeable difference between malignant and benign tumor sizes. For radius, texture, compactness, concavity, and concave points, malignant is noticeably larger than benign. Smoothness and symmetry are closer but there is a slight difference. Fractal dimension is too hard to tell.

I also performed hypothesis testing to compare the malignant values and the benign values for each mean. The null hypothesis is that the factors do not determine whether the tumor is benign or malignant. If the p-value is less than 0.05, I then reject the null hypothesis and say these factors are significant. The results are listed in Table 1:

Variable mean	p-value
Radius	8.47×10^{-96}
Texture	4.06×10^{-25}
Smoothness	1.05×10^{-18}
Compactness	3.94×10^{-56}
Concavity	9.97×10^{-84}
Concave Points	7.10×10^{-116}
Symmetry	5.73×10^{-16}
Fractal Dimensions	0.76

Table 1: P-values of Malignant vs Benign Tumors for Each Variable Mean

All of the p-values except for fractal dimensions were very low with the exception of fractal dimensions, and so the null hypothesis was rejected for each, and the differences between malignant and benign for these means is significant. I safely eliminated fractal dimensions as a significant variable.

The next step that I wanted to perform in the exploratory data analysis was to compare the correlation between the different variable means. I was able to quickly visualize this comparison by looking at a heatmap, which is show in Figure 9.

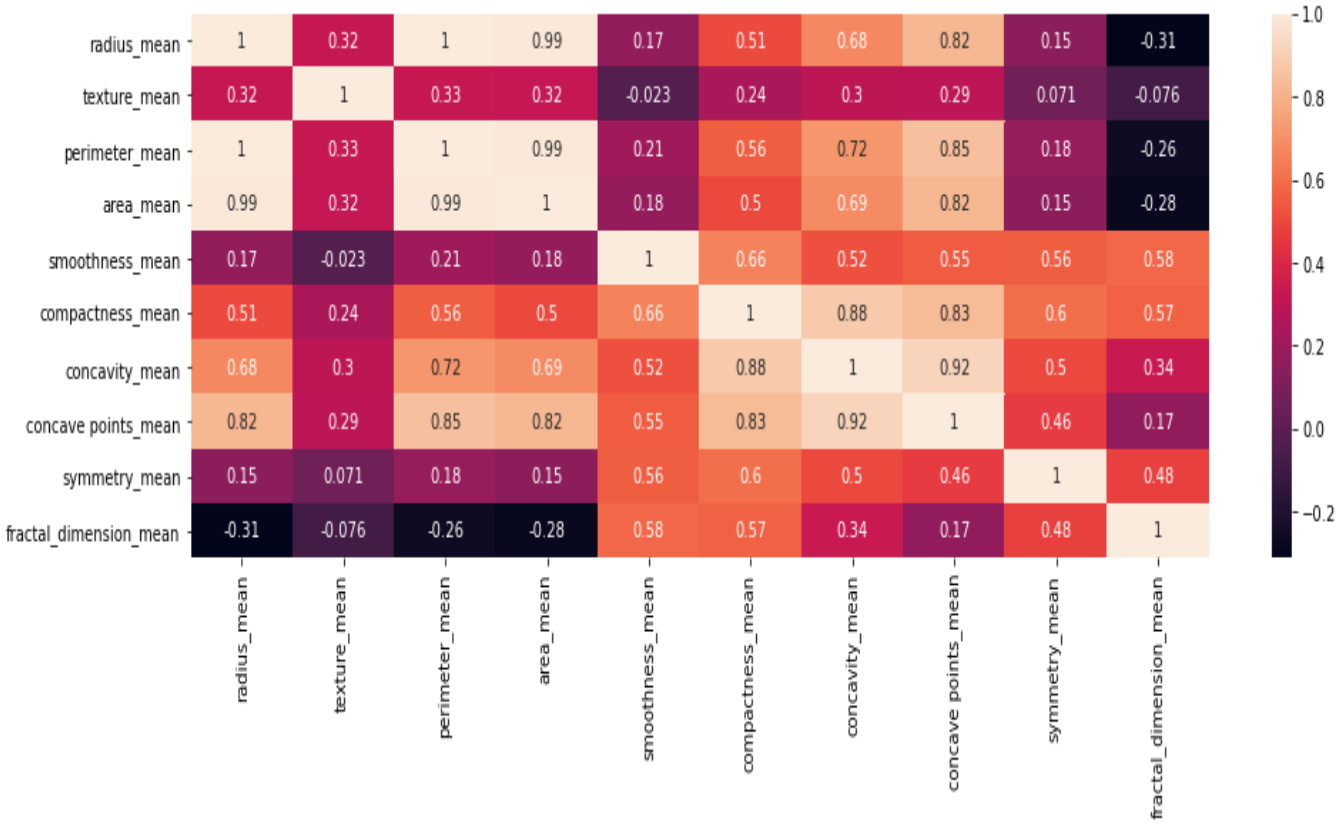


Figure 9: Heatmap of Correlation of Means of Variables

The heatmap shows some high correlations, other than the obvious radius, perimeter, and area. Interesting high correlations that I noticed were radius with compactness, concavity, and concave points, smoothness with compactness, concavity, concave points, symmetry, and fractal dimensions, and concavity mean to compactness, concave points, and symmetry. I decided to find the variance inflation factor with these, and if there are values greater than 5, then we will consider that a severe correlation and eliminate some determining variables. Using this, the compactness mean, concavity mean, and concave points mean are too strongly correlated to radius mean and were removed from modeling. Also, symmetry was highly correlated with smoothness and was removed from modeling.

The next step was to compare the correlation between mean, standard error, and worst to see if there was a high correlation between each. Heatmaps were used to view the correlations. The results are listed in figures 10-17..

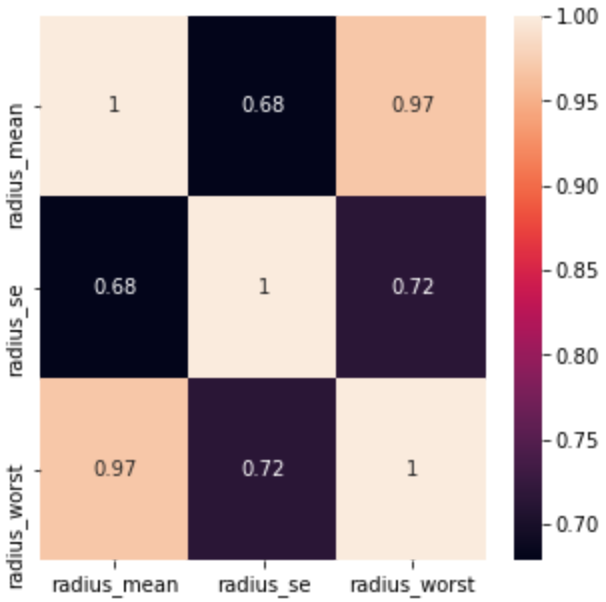


Figure 10: Heatmap of Correlation of Mean, Standard Error, and Worst of Radius

Figure 11: Heatmap of Correlation of Mean, Standard Error, and Worst of Texture

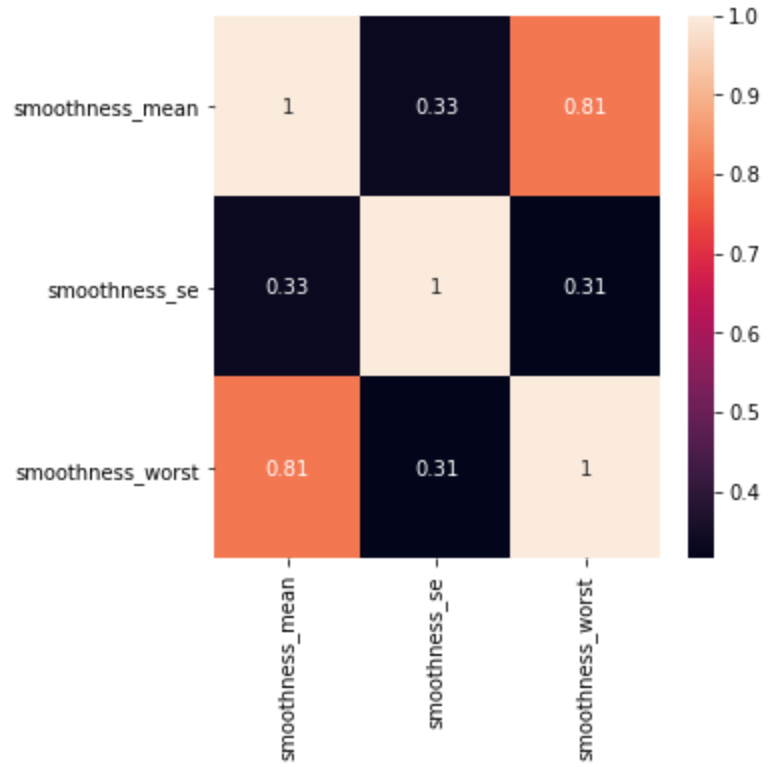


Figure 12: Heatmap of Correlation of Mean, Standard Error, and Worst of Smoothness

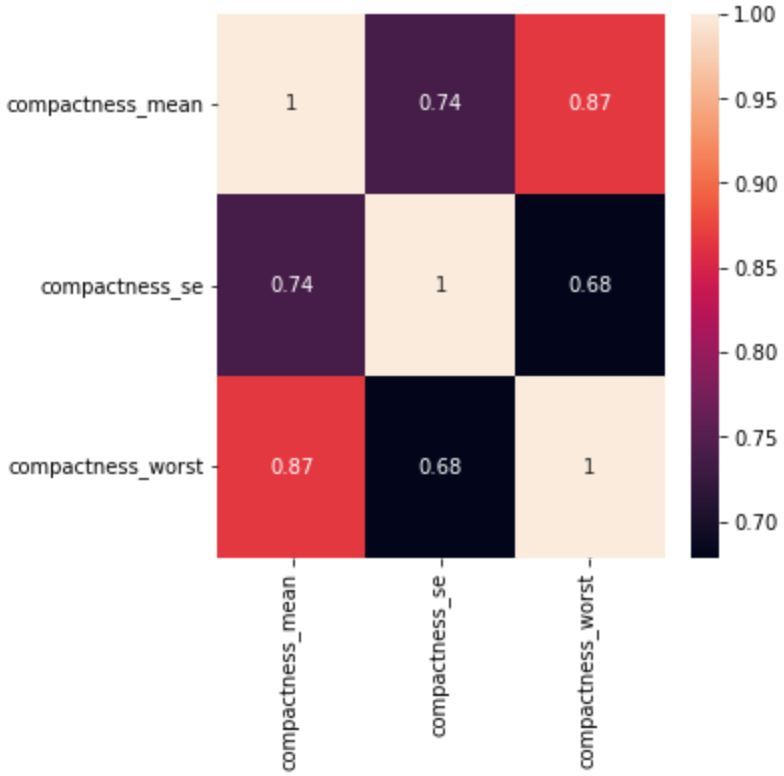


Figure 13: Heatmap of Correlation of Mean, Standard Error, and Worst of Compactness

Figure 14: Heatmap of Correlation of Mean, Standard Error, and Worst of Concavity

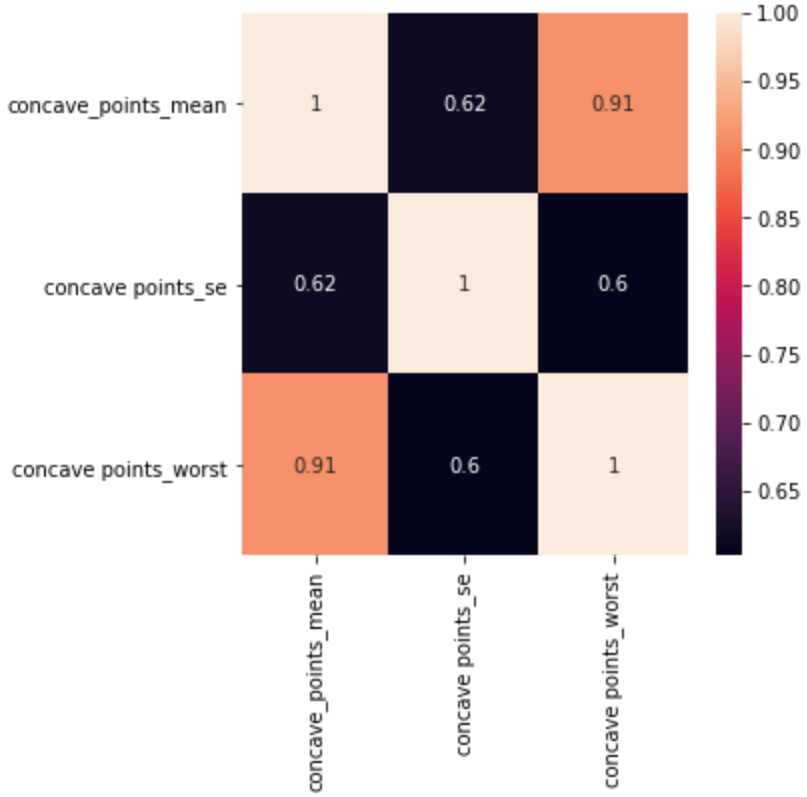


Figure 15: Heatmap of Correlation of Mean, Standard Error, and Worst of Concave Points

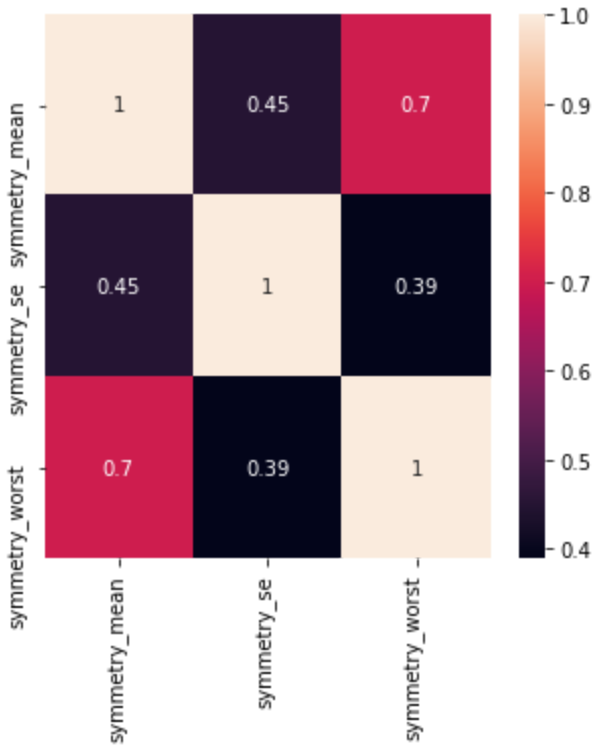


Figure 16: Heatmap of Correlation of Mean, Standard Error, and Worst of Symmetry

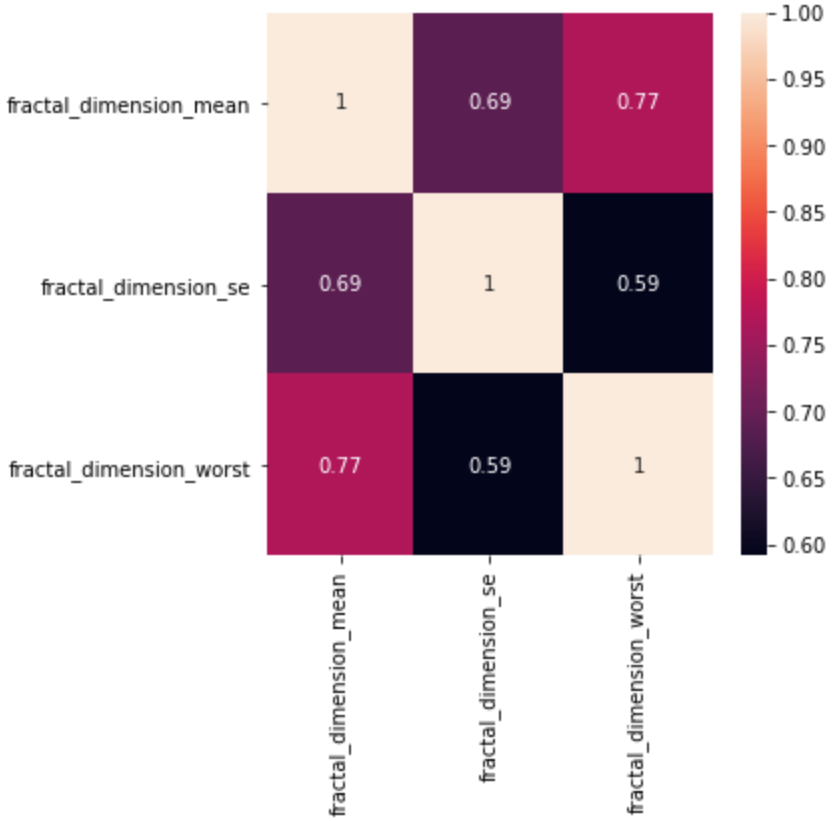


Figure 17: Heatmap of Correlation of Mean, Standard Error, and Worst of Fractal Dimension

For most of these, the correlation was not as high as I predicted. While the correlation of worst to mean was pretty high, the standard error was much lower. I removed from consideration the worst variables because of this high correlation, but I decided to run hypothesis testing comparing the standard error of malignant tumors to benign tumors. The null hypothesis is that the factors do not determine whether the tumor is benign or malignant. If the p-value is less than 0.05, I then reject the null hypothesis and say these factors are significant. The results are listed below:

Variable mean	**p-value**
Radius	9.74×10^{-50}
Texture	0.84
Smoothness	0.11
Compactness	9.98×10^{-13}
Concavity	8.26 x 10-10
Concave Points	3.07 x 10-24
Symmetry	0.87
Fractal Dimensions	0.06

Table 2: P-values of Malignant vs Benign Tumors for Each Variable Standard Error

The standard error for radius, perimeter, area, compactness, concavity, and concave points all rejected the null hypothesis and seemed to be significant in determining whether a tumor is benign or malignant. For each one of these, I ran the VIF to see if they were too closely correlated to the mean. If it is a value greater than 5, I removed them from modeling. Radius, perimeter, compactness, and concave points all had a high VIF and were removed from modeling. The standard error of area and concavity were lower than 5 and kept in modeling.

Almost all of this data was significant and useful in predicting whether a tumor was benign or malignant. However, the correlation was extremely high, so most of the determining variables were removed for modeling to reduce multicollinearity. The determining variables that were retained are:

radius mean
texture mean
smoothness mean
area standard error
concavity standard error

Since a large majority of the variables were removed, I decided in my modeling I would model the data using just these five variables and model keeping all of the variables except for area and perimeter mean. For more visualization of the EDA on tableau, click here.

Modeling

The modeling section of my project involved two parts: creating a baseline model and then creating other models for comparison to the baseline. I performed all modeling for both datasets with all the determining variables included and datasets with only the five determining variables identified after the EDA to see which datasets performed better. The baseline model that I chose to use was logistic regression, as it is often considered the standard model in classification. I decided to split the data into 80% for training and 20% for testing. This resulted in 455 data points for testing and 114 for testing. I ran logistic regression for the partial feature dataset and full feature dataset with both the default logistic regression hyperparameters and best hyperparameters found by using Grid Search Cross Validation. These hyperparameters were an l2 penalty and a C of 10. The full feature dataset provided better scores for logistic regression. The values given of the best logistic regression are listed below:

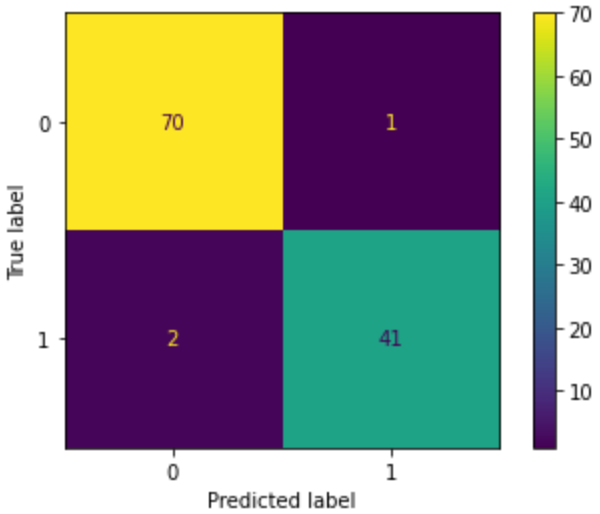


Figure 18: Confusion Matrix of Breast Cancer with Logistic Regression

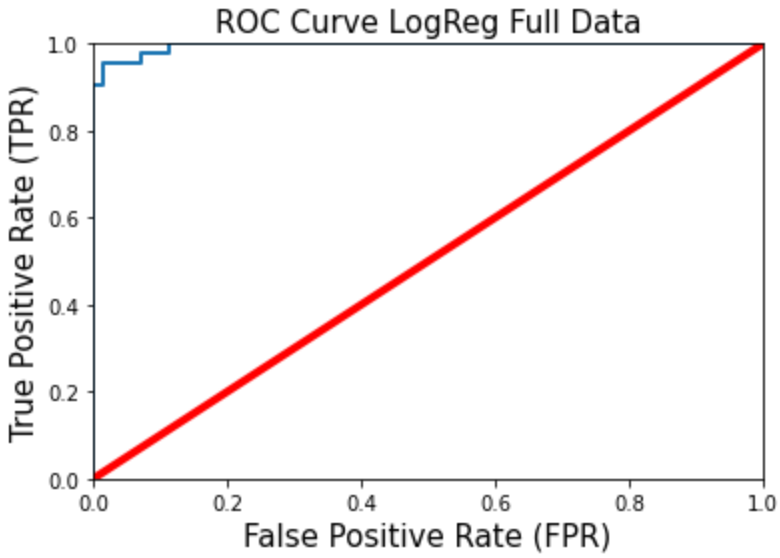


Figure 19: ROC Curve of Breast Cancer with Logistic Regression

Accuracy	False Negatives	Recall	AUC
97.37%	2 out of 114 (1.75%)	95%	99.5%

Table 3: Important effectiveness measurements using Logistic Regression

I also graphed the top features by level of importance. Figure 20 shows that the top four features were the worst values for concavity, compactness, and radius and standard error for texture. After that there was a large dropoff, but it is important to note that none of these were used in the reduced feature dataset.

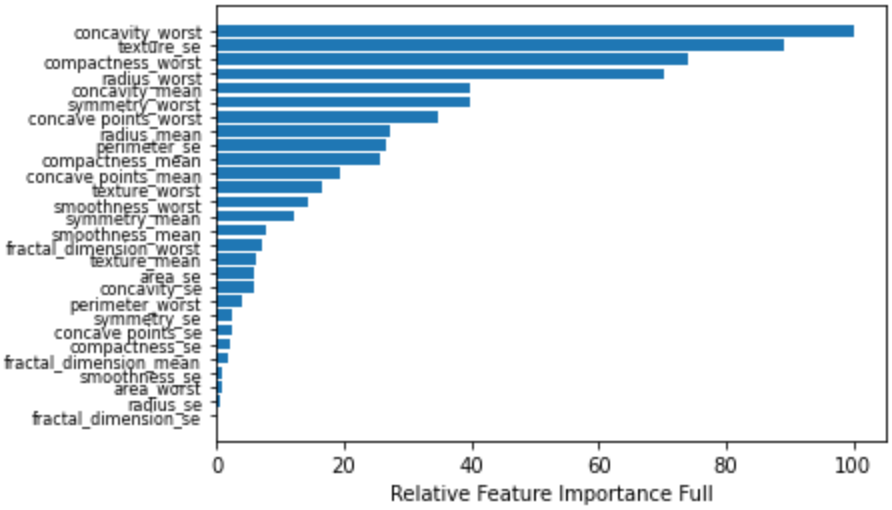


Figure 20: Feature Importance of Logistic Regression

Based on these results, I created the following targeted metrics in my next models to consider the models effective:

1. Accuracy of at least 97%
2. Less than 2 % false negatives
3. A recall of at least 95%
4. AUC of at least 99%

For the modeling after the baseline model, I decided to use three models: K-Nearest Neighbors(KNN), Random Forest Classification, and XGBoost Classification. For each of the models, I also used GridSearch cross validation on three of the hyperparameters available. The best values for KNN were 6 neighbors, a power parameter of Manhattan Distance, and a weight on distance. The best values for Random Forest were a max depth of 30, a log2 of max features, and 300 estimators. The best values for XGBoost were a learning rate of 1, a max depth of 2, and 10 estimators. For all three models, the datasets that had all of the features outperformed the datasets that had the partial features, so these results are the ones that will be listed below.

KNN Model

For the KNN model, the effectiveness models are given below and in table 4.

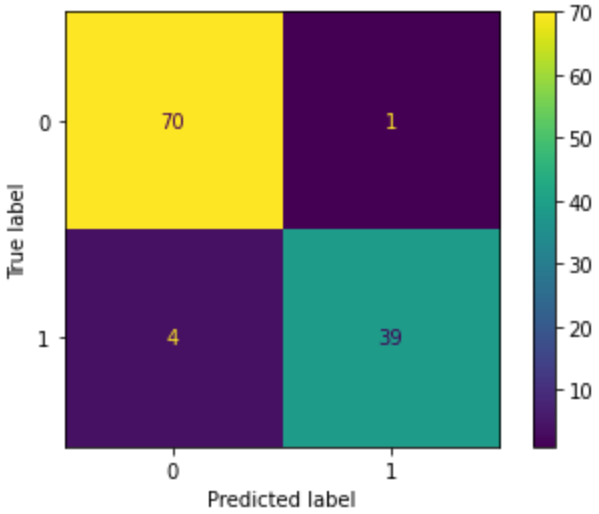


Figure 21: Confusion Matrix of Breast Cancer with KNN

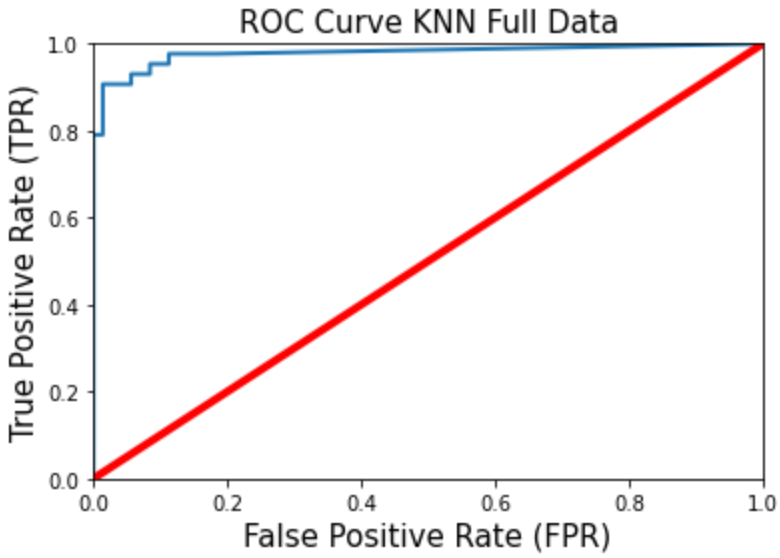


Figure 22: ROC Curve of Breast Cancer with KNN

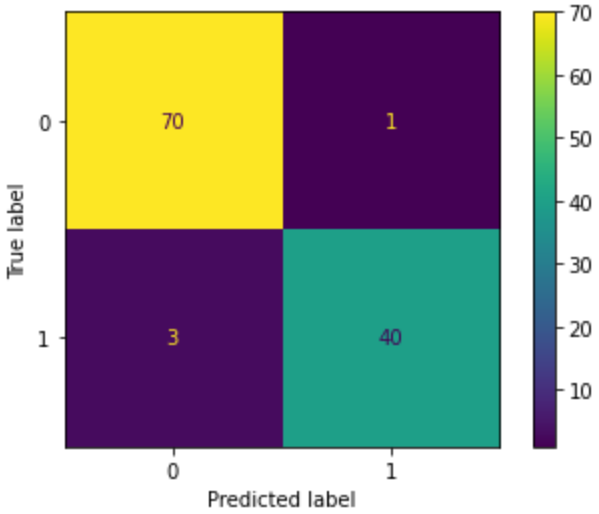
Accuracy	False Negatives	Recall	AUC
95.6%	4 out of 114 (3.5%)	91%	97.8%

Table 4: Important effectiveness measurements using KNN

KNN does not take into account feature importance, so it was not taken into account for this model.

Random Forest Model

For the Random Forest model, the effectiveness models are given below and in table 5



.Figure 23: Confusion Matrix of Breast Cancer with Random Forest

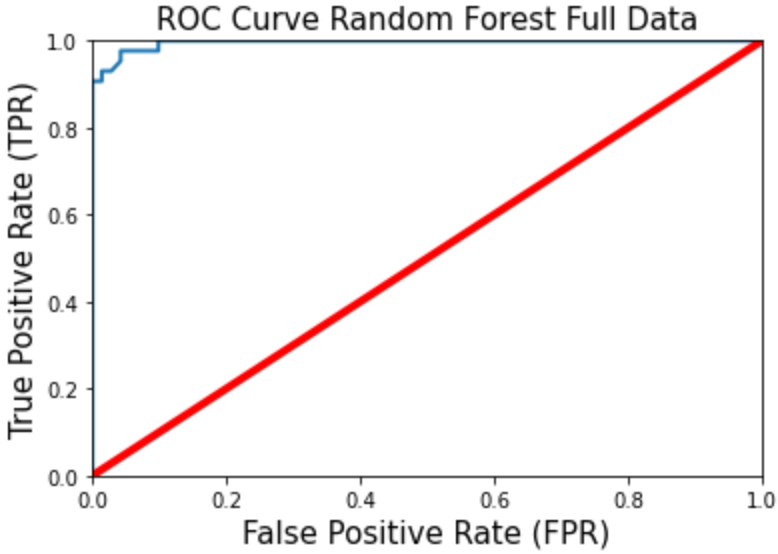


Figure 24: ROC Curve of Breast Cancer with Random Forest

Accuracy	False Negatives	Recall	AUC
96.5%	3 out of 114 (2.6%)	93%	99.6%

Table 5: Important effectiveness measurements using KNN

I also graphed the top features by level of importance. Figure 23 shows that the most important features were the mean and worst values of concave points, followed by the worst perimeter. There is a large drop off before the worst of area and radius are taken into account.

The worst of radius, area, and perimeter are all closely correlated, so this would be predictable. The radius mean is the next most important after another noticeable dropoff, which is the first feature that is included in the reduced feature dataset.

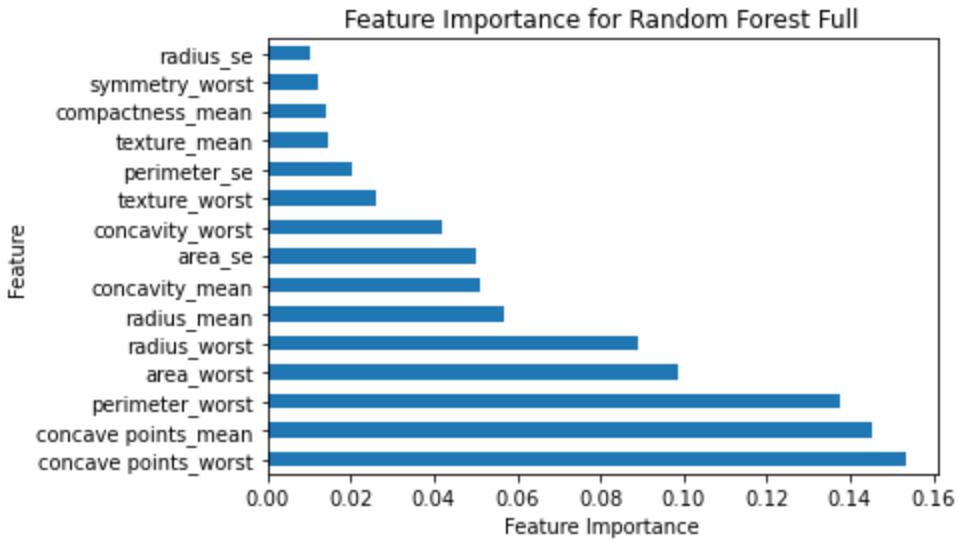
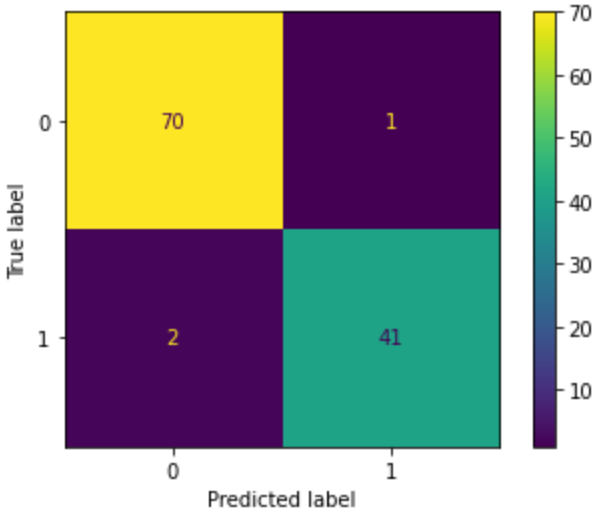


Figure 25: Feature Importance of Random Forest

XGBoost Model

For the Random Forest model, the effectiveness models are given below and in table 6



.Figure 26: Confusion Matrix of Breast Cancer with XGBoost

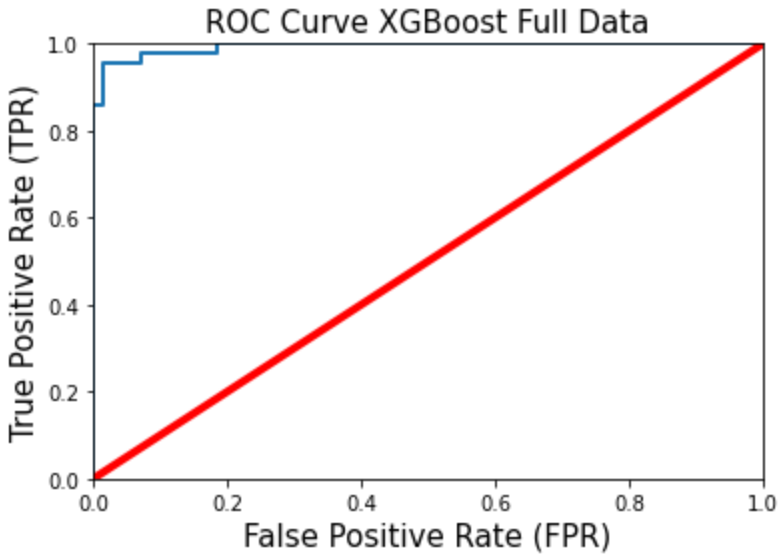


Figure 27: ROC Curve of Breast Cancer with XGBoost

Accuracy	False Negatives	Recall	AUC
97.37%	2 out of 114 (1.75%)	95%	99.3%

Table 6: Important effectiveness measurements using XGBoost

I also graphed the top features by level of importance. Figure 28 shows that the most important feature was the mean of the concave points by a substantial margin. This was the second most important for Random Forest. The second most important feature, which was about half of the mean of concave points, was the worst for perimeter. This was the third most

important for Random Forest. After that was a very large drop off where the features took less importance.

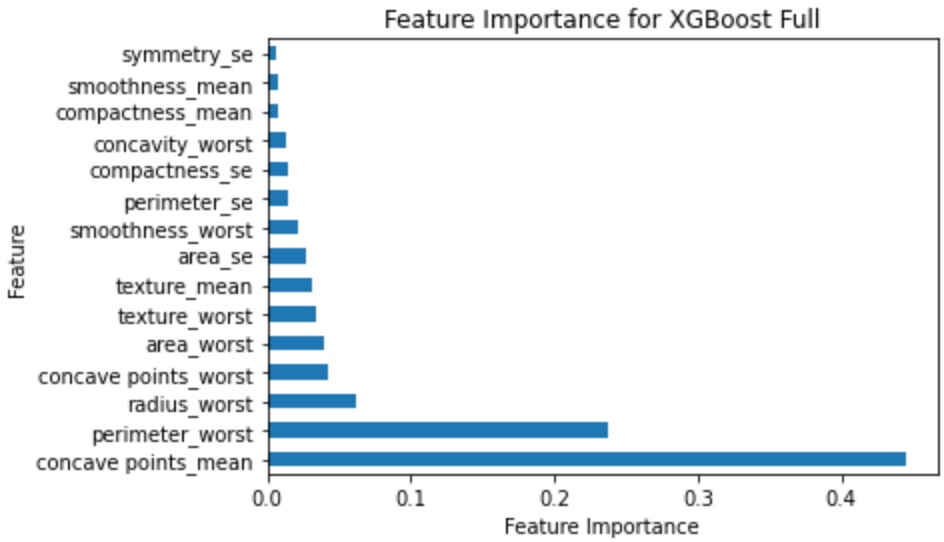


Figure 29: Feature Importance of XGBoost

Comparison of Models

I graphed the results against each other in order to get a visual comparison of how they compared. I am focusing on the four key measurements that I have outlined before. This includes the logistic regression baseline model and all of the results that were found when running the models using the datasets that were reduced to five features.

Accuracy

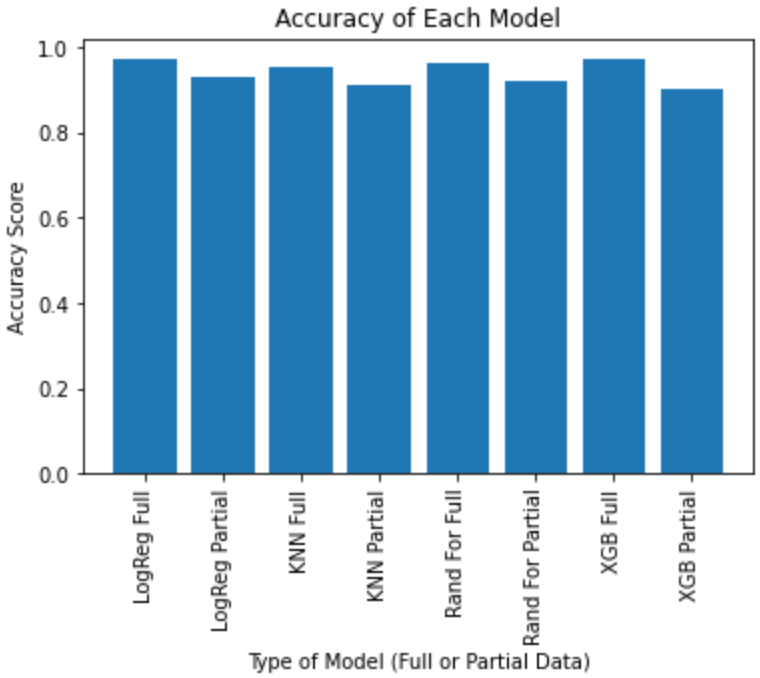
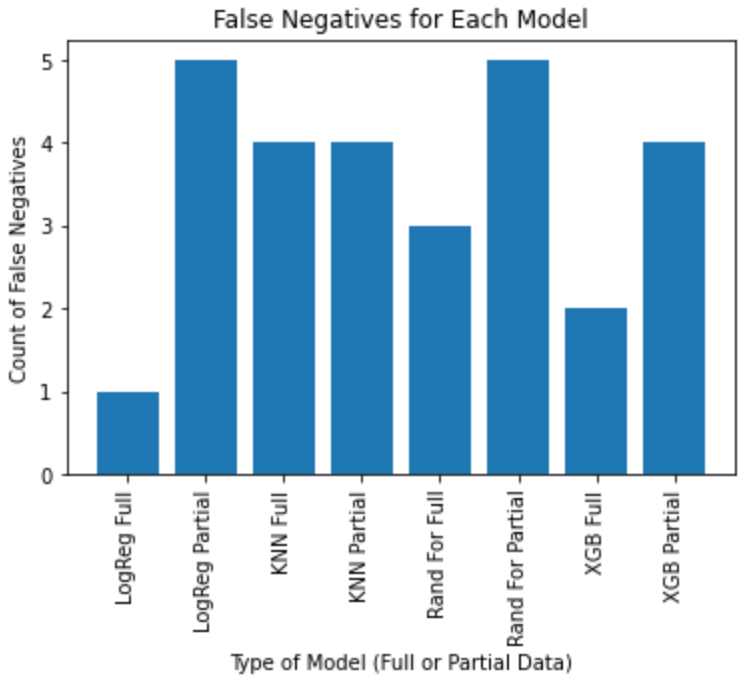


Figure 30:The Comparison of Accuracies Between Models

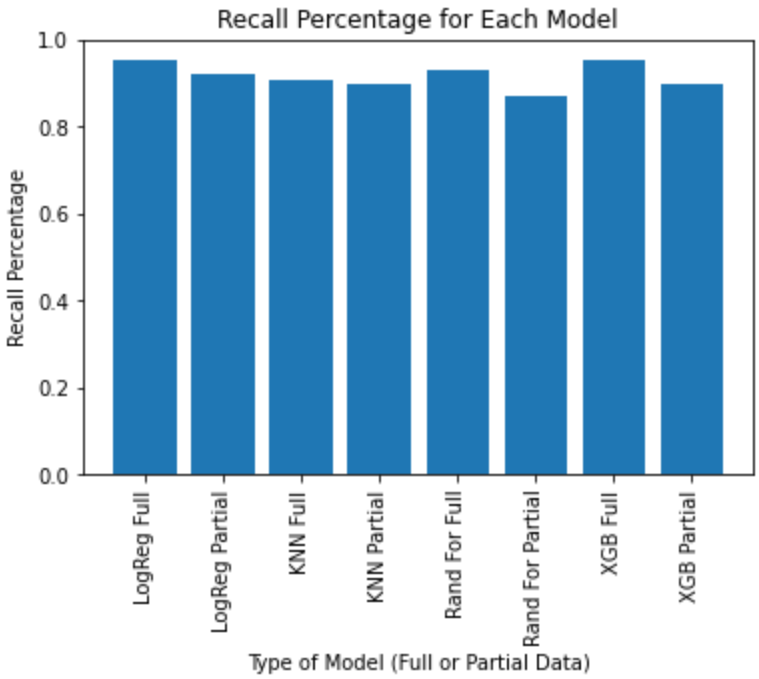
The most obvious first thing noticed is that the models with the full data consistently did better than the models with partial data. In fact, the model that did best with partial data still was not as accurate as the model that was the worst with the full data. Also when looking at the threshold, logistic regression and XGBoost were the only ones that met that 97% threshold, both having the exact same accuracy. So far, XGBoost has been the only model that is comparable to the baseline model, but all of the accuracies were high with a small difference.

Number of False Negatives



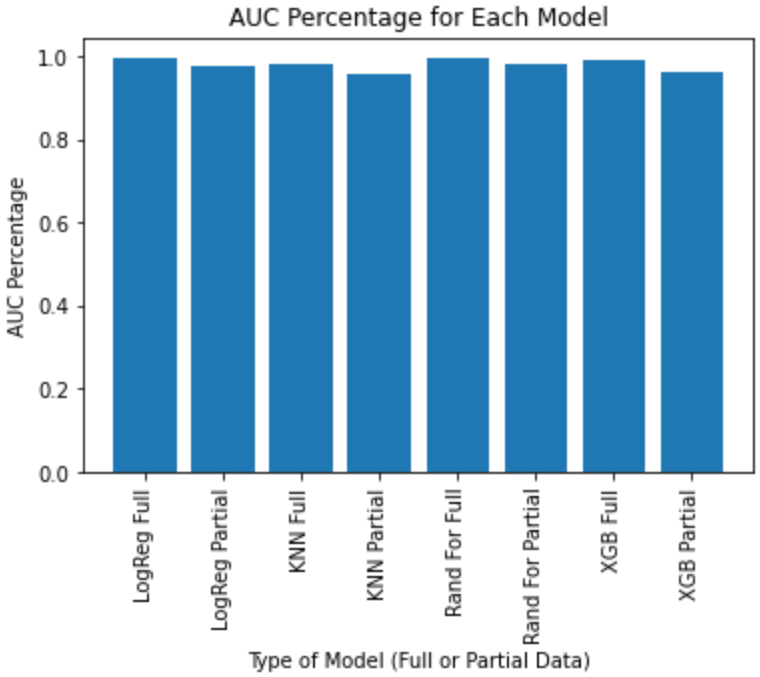
This data is a little easier to differentiate. In this case, lower numbers are better, so the logistic regression with the full data performed best, followed by the XGB Full data. The full datasets once again performed better overall. Logistic Regression and XGBoost were once again the only models that reached the threshold of less than 2% false negatives

Recall Percentage



A pattern is starting to emerge, with logistic regression and XGB full being the best, followed by Random Forest. Like accuracy, these values were all very close so harder to differentiate, but the percentages with the full features all had better values than the partial. The recall percentage of XGBoost and Logistic Regression were the exact same, and they were the only ones that met the threshold of at least 95%

AUC Percentage



Once again, this one is a little harder to tell by just looking at the graph, but I can tell that the full models have been consistently higher. Random Forest was the highest with a 99.56% AUC, followed by logistic regression with 99.51% and XGBoost with 99.3%. All of these models exceeded the threshold of 99%.

Using these results, I developed interactive models on Tableau that could be used to see how changing the percentage threshold for what would be considered a malignant diagnosis changed the number of false negatives, which is the most important feature that is trying to be avoided. Reducing the threshold for considering a tumor malignant is acceptable if it reduces the false negatives, even if the number of false positives increases. This interactive feature has been created for the dataset with reduced features, which can be accessed here, and the dataset with full features, which can be accessed here.

Takeaways

The initial takeaway is that XGBoost was the only model that consistently met all of the established thresholds, along with the baseline Logistic Regression Model. Logistic Regression performed slightly better in AUC, while they performed equally in the other key areas. These models performed close enough that their effectiveness could be affected by further adjustments of the hyperparameters and adding more datapoints in order to test against, since the number of datapoints was relatively low. When looking at the thresholds that I established, both of these models were virtually identical and both acceptable.

I considered the reduction of false negatives the most important factor in modeling, so I wanted to use the interactive Tableau that I created to see what would be the highest threshold for considering a tumor malignant where there would be no false negatives. For Logistic Regression, the highest threshold was 10%, and this created eight false positives, which was a difference of six more than the standard threshold of 50%. The Tableau thresholds were set in multiples of 5%, and at a threshold of 5%, there was still one false positive, with seven false positives. When the threshold was set to 0%, there were zero false positives, but all of the true negatives turned into false positives. I would have to reduce the multiples to a smaller percentage to determine if there was a positive threshold between 0% and 5% that is ideal for XGBoost. However, with the given results, the baseline model of Logistic Regression performed the best.

Future Research

The high correlation between several of the variables made narrowing down the data to avoid multicollinearity very difficult, and the models with all of the features showed main feature importance with the variables that would have been removed in a reduced feature dataset. It would be worth adjusting the features in which to model in different ways to see what would be an ideal set of features to model against that would retain the high scores that are needed and reduce correlation.

The testing set was fairly small, with only 114 sets of data to test. Ideally, these models would be best suited to test new data as it was gathered to see if it continued to meet the threshold that we want. It would help determine if XGBoost or Logistic Regression would be the most useful if there were many more data points.

The models might benefit from improvement by further adjustment and expansion of their hyperparameters. While I did use GridSearchCV, expanding the range of values of the hyperparameters could pinpoint more accurate models since I used a very limited set. I could also add hyperparameters that I did not include initially.

Recommendations

I believe that XGBoost and Logistic Regression are ready to be presented in a business setting. If this was presented, these models provide the structure needed to begin accurate diagnoses of whether a tumor is benign or malignant. As explained above, I would suggest having a low threshold of an initial positive diagnosis between 0-5% to ensure that any potential malignant tumors would not be overlooked, and then do further testing on the malignant diagnoses. Misclassifying some benign tumors as malignant would be a mistake that should be avoided, but would be acceptable if it means reduced risk of the alternative. I would recommend that this model be used as the first predictor of if a tumor is benign or malignant, but in a situation that is as serious as this, continuous monitoring should always be done, and no one method should be considered absolutely correct.