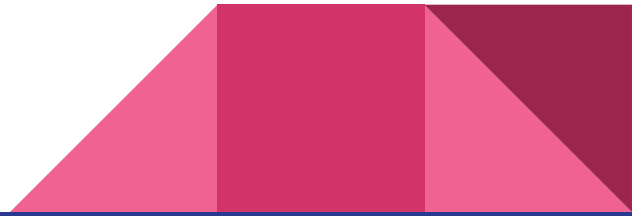
Breast Cancer Wisconsin (Diagnostic) Using Classification Models

Anthony Castillo, Patricia Kurniawan

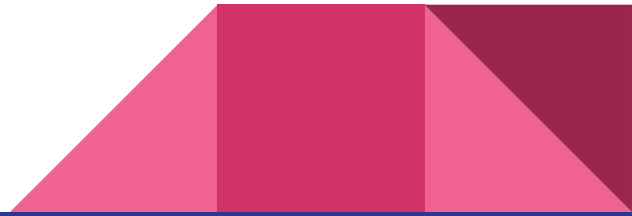
Introduction

The data chosen for this project was the Breast Cancer Wisconsin (Diagnostic) Data Set. The dataset contains 569 patients and 32 attributes that could lead a patient in being malignant (cancerous) or benign such factors as the radius, texture, perimeter , area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension of the breast cancer we have the mean, standard error, and worst from the listed factors above.



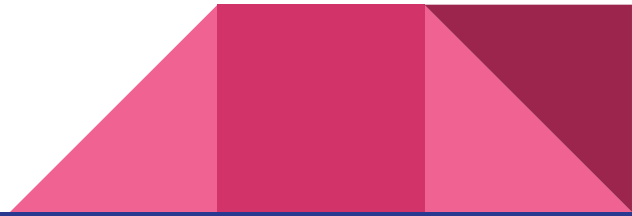
Problem Statement

The data set include many features that could decide the diagnosis of breast cancer. The data scientists goal is to create a predictive model that helps classifying if the diagnosis of the breast mass is malignant (cancerous) or benign using classification models.



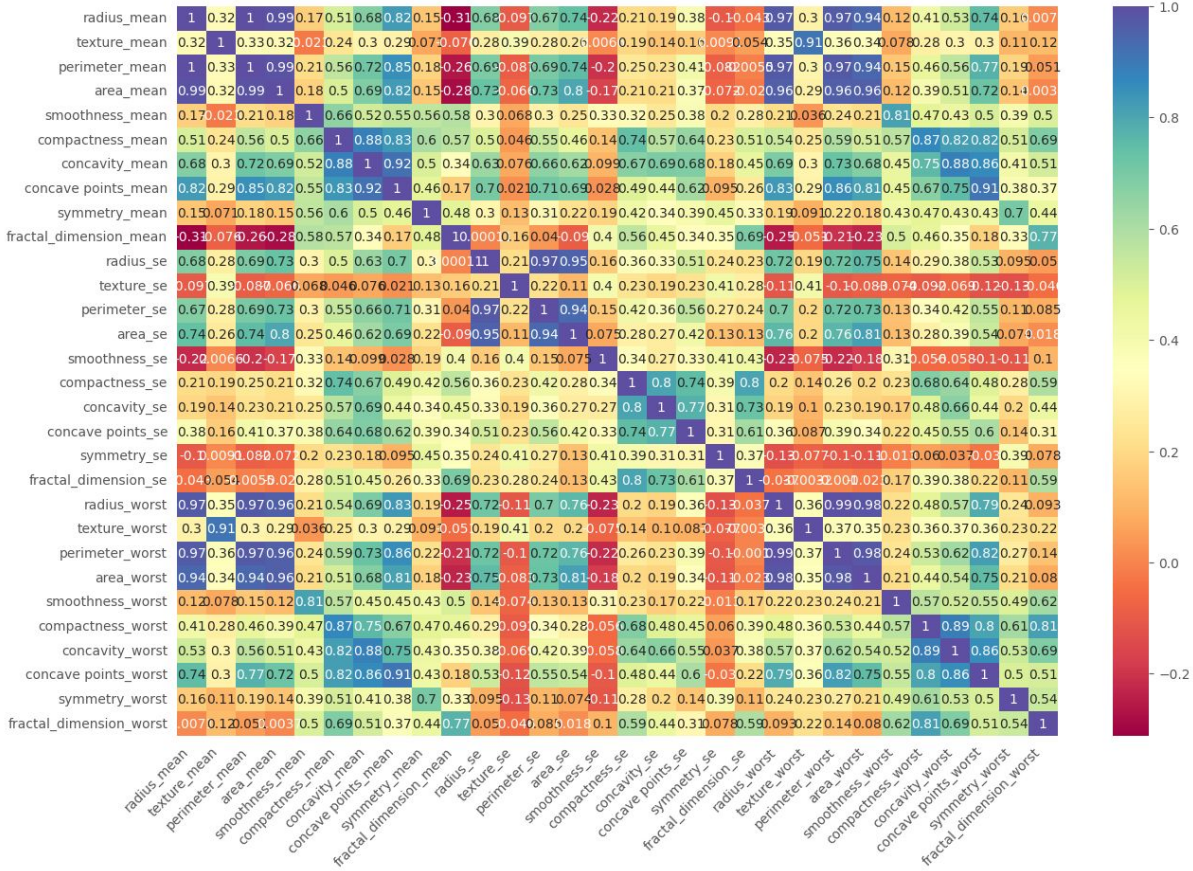
Methods / Implementation

- Logistic Regression
- Decision Tree Classification
- Random Forest Classification
- Naive Bayes Classification
- KNN Classification



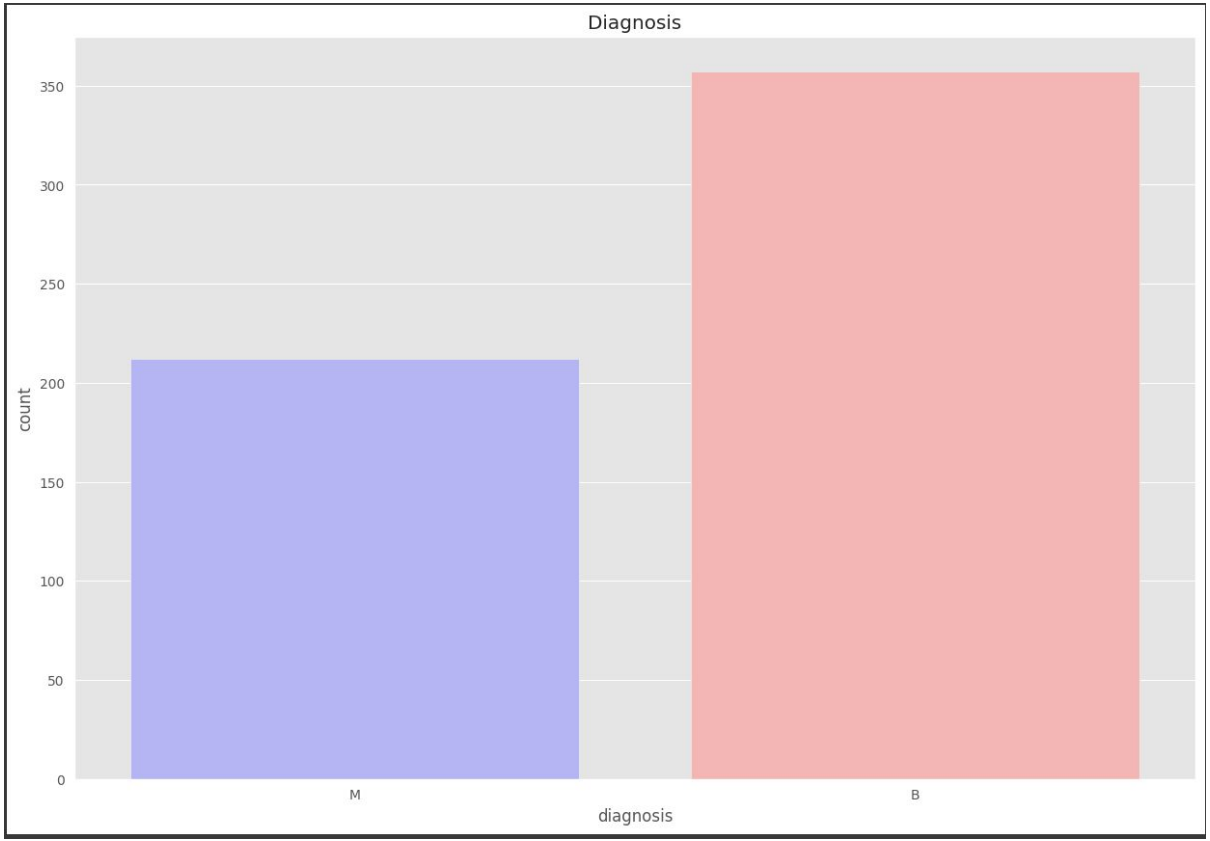
Data Exploration and pre-processing

The first thing done with dataset is drop the features “id” and “Unnamed: 32” since they do not have any effect on the diagnosis. We then look at the correlation matrix to find out which features are highly correlated which seem to be area_mean, radius_worst, perimeter_worst, and area_worst which means the features have strong relationship with each other.



Data Exploration and pre-processing

This plot show the amount of patients that were diagnosed with malignant and benign. It looks like the chances of a patient not having breast cancer are much more higher than being malignant.



Data Exploration and pre-processing

The first thing done is to split the features and the target feature.We then scale the features so that one feature does not affect training models and leads to loss in accuracy. After we split the data into training and test sets with 70%/30% split. The training sets are used to train the model and the test sets are to test accuracy on the trained model since the test set is untouched by the classification models and will not overfit.

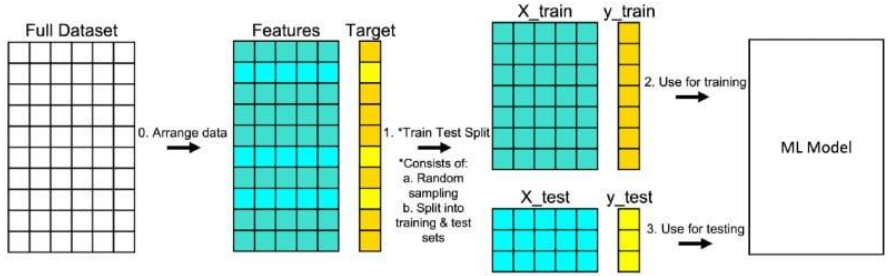
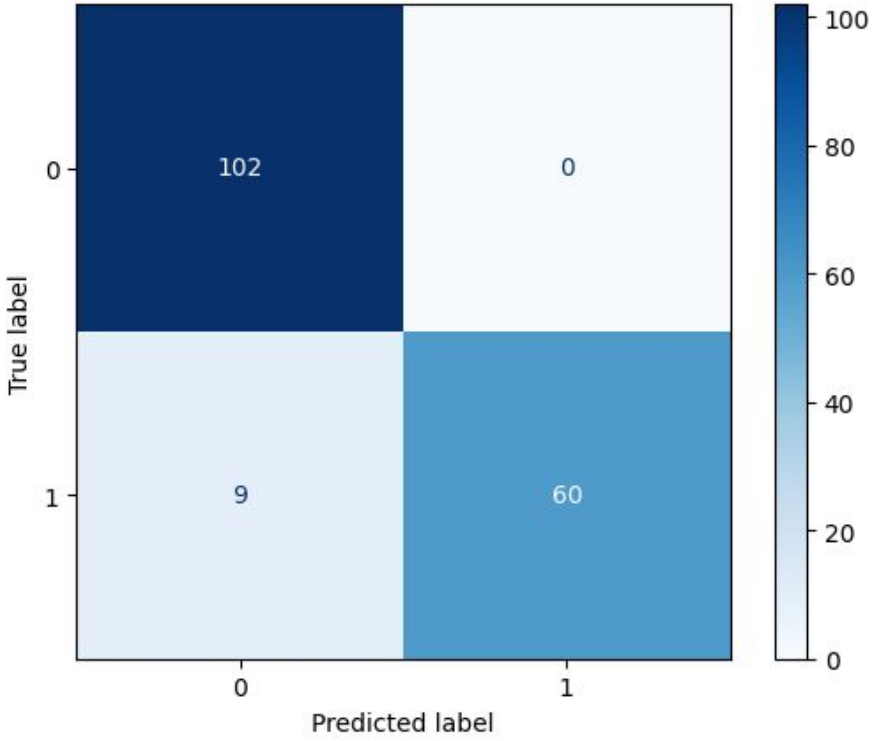


Image credit: https://builtin.com/data-science/train-test-split

Logistic Regression

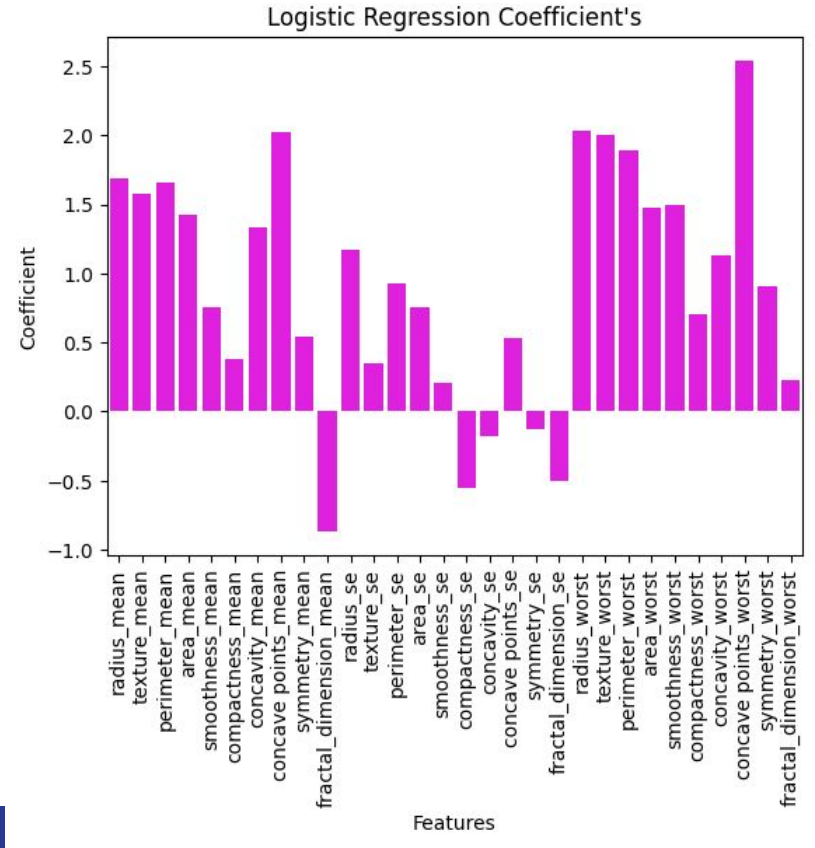
The first model we used to train the data is Logistic Regression. The benefits of using this model is that it's simple to implement and easy to understand the outcome of the target feature. The confusion matrix seen on the right gives us an accuracy of 94% which tells us the amount of correctly classified True Positives and True Negatives.

We can also use a different measuring metric in recall score which shows the correctly classified positives out of all the actual positives. The recall score for the confusion matrix is 87% we want a higher percentage since we want to decrease the amount of patients who classified with breast cancer.



Logistic Regression (continued)

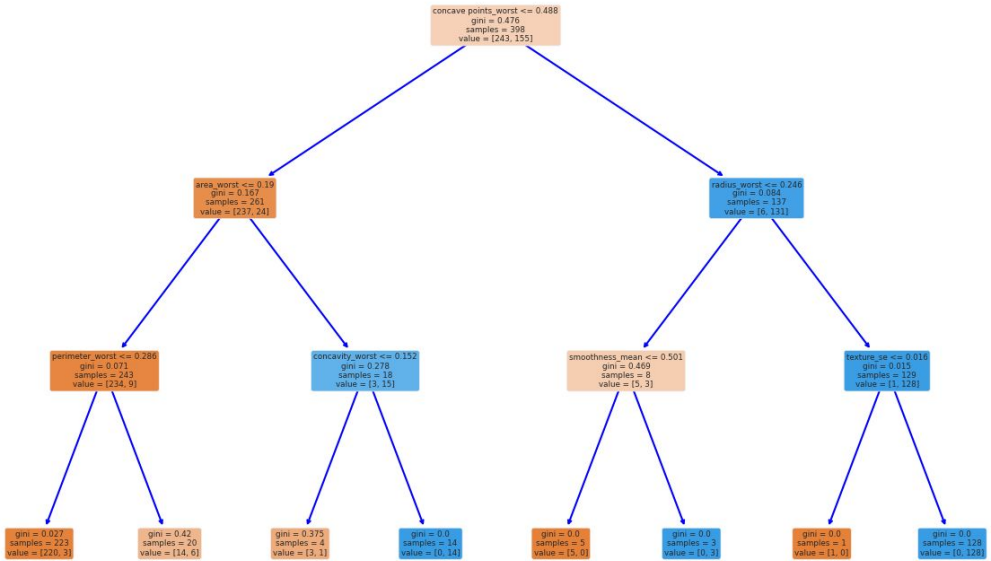
Here you can see the coefficients from training the logistic regression model. The negative coefficients in the barplot indicate the log odds probability has decreased and also decrease the chances of classifying correct. While the positive coefficients mean that the log odds probability increase which will give a higher chance of the model classifying correct.



Decision Tree

We created a decision tree of max depth 3 and created a visualization seen on the right. Through this decision tree we can see that one of the primary variables is the mean concave points of the mass. This visual representation gives us a general idea of how the model classifies whether or not a breast mass is deemed malignant or benign.

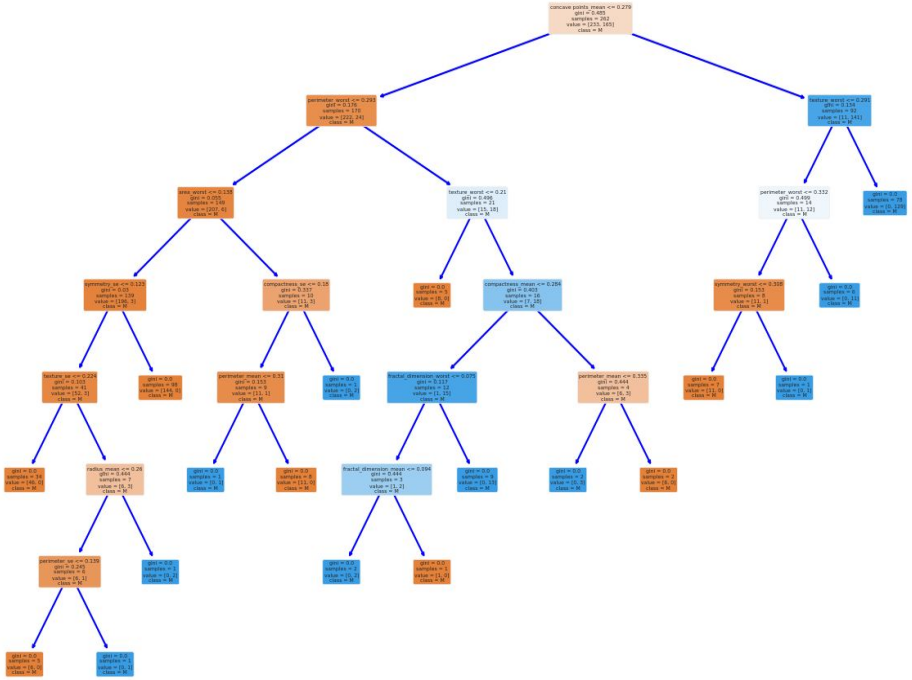
The accuracy of this model is quite high, at 0.9474, meaning that the model is accurate about 95% of the time given a testing set. The recall is also high with 0.8986, though not amazingly high.



Random Forest

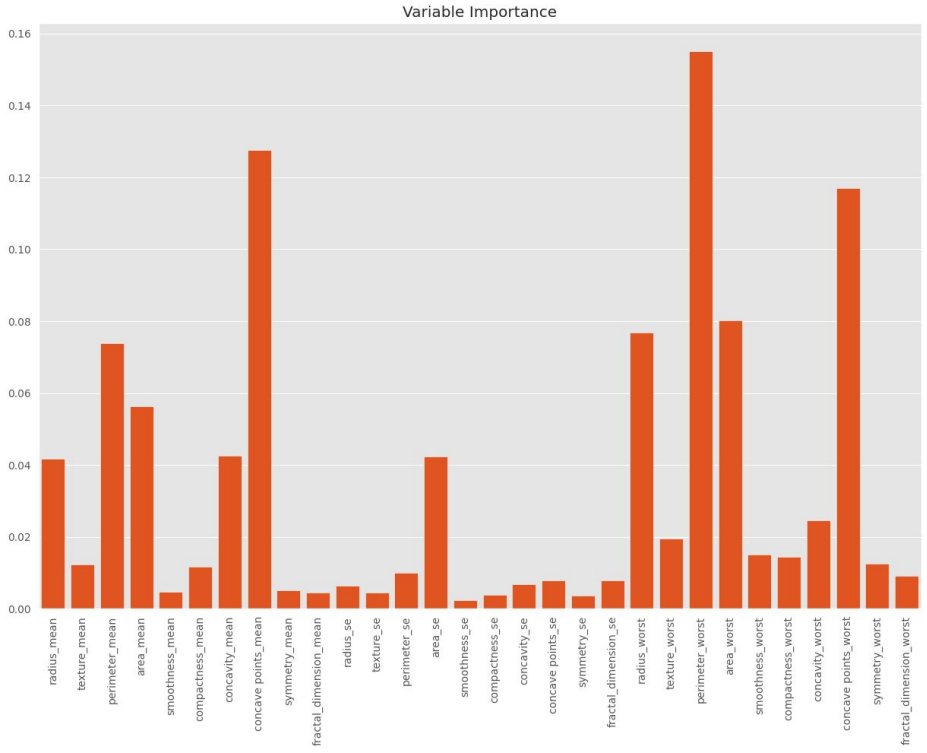
Using Random Forest can lead to better results compared to decision tree but the drawback of random forest's it creates a group of decision trees with each one using different features called bagging so that leads this algorithm to be slower than the rest. But some advantages to random forest are that there are less chance of overfitting compared to decision tree.

This algorithm performed the best out of all the models with an accuracy of 95% meaning that we are 95% confident that the patient will be classified benign.The recall score which is a better indicator on the performance on the model and tells us the amount of correctly classified positives which was 91%.



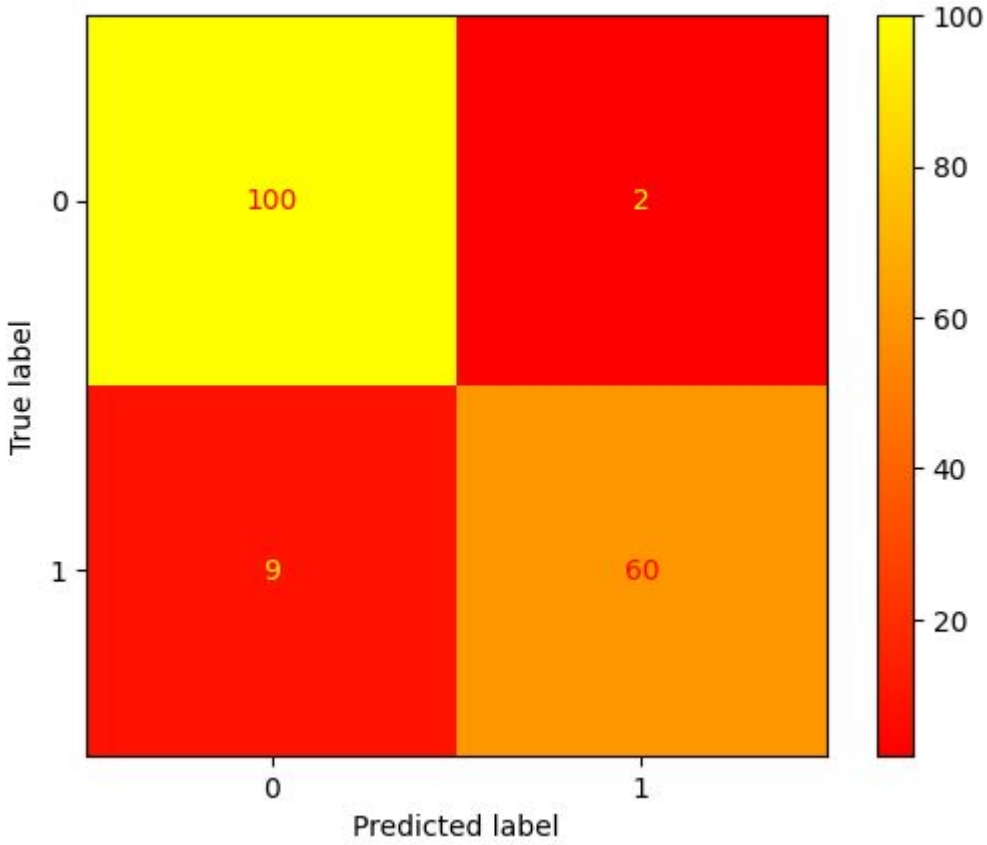
Random Forest (continued)

Here are the most important features that the random forest algorithm concluded that were the most important in classifying whether a patient was malignant or benign. It looks like the most important were the perimeter_worst, concave points_mean, and concave points_worst. These features can be used to run the random forest model again and improve the models accuracy.



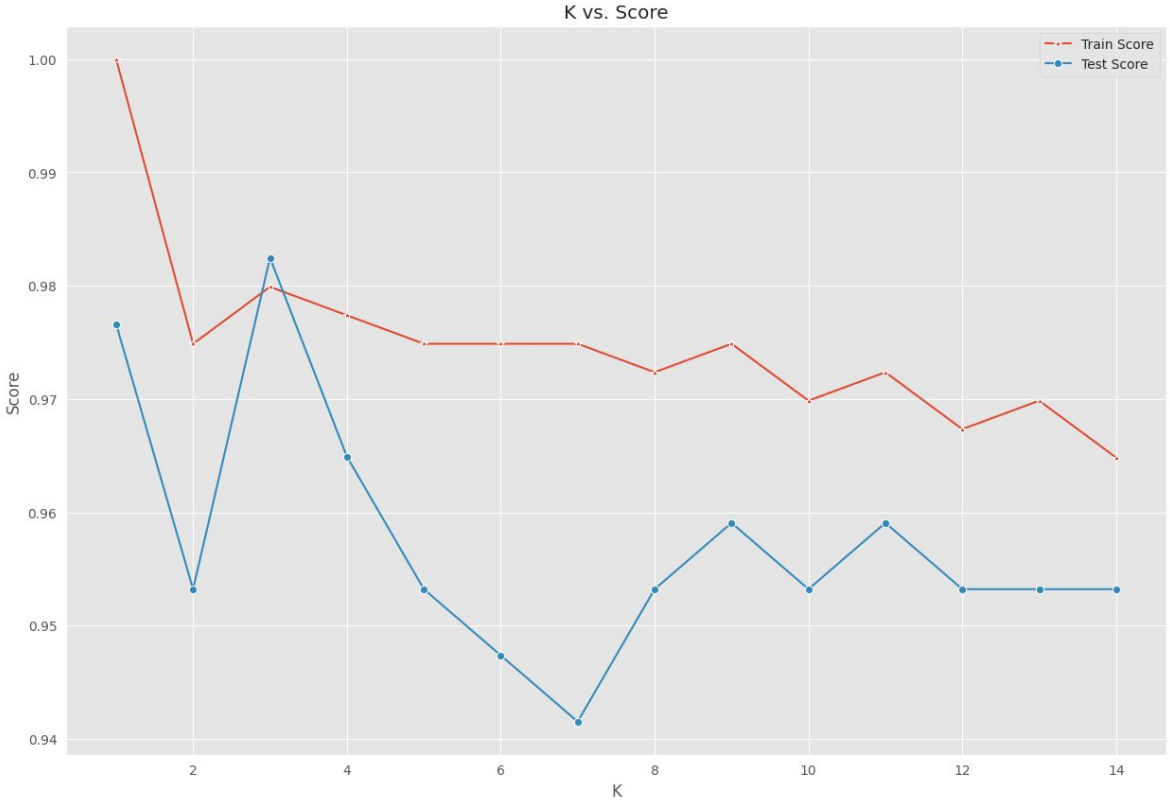
Naive Bayes

The Naive Bayes algorithm was implemented since it is computationally efficient and easy to understand the outcome but the drawbacks of this algorithm is that it assumes the features to be independent. After training the model, our accuracy was high at 94%, but our recall was low with it being 87% compared to the other algorithms with high accuracy.



K-Nearest Neighbors

For the KNN model the first thing to do is to find the optimal k value for the model.If we use a k-value too high it will overfit the data and is also computationally heavy. We used the elbow method to find the best k and the best k seems to at value = 2. We then train the model with k value of 2 and the model had an accuracy of 95% and recall score of 88%.



Results

Even with the low accuracy, higher false positives would be beneficial as having preventative care is better than having missing someone who needs care.

The model that performed was Random Forest with accuracy of 95% and recall score of 91% but it comes at the cost of it being slower compared to that of Logistic Regression or Naive Bayes. The other models also performed well so it would be beneficial to use the simpler and easier to understand algorithms such as Logistic Regression, Naive Bayes, or KNN.