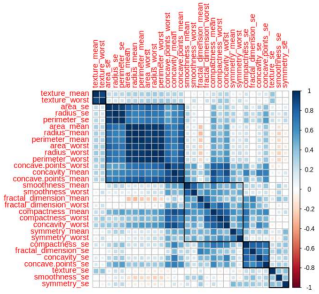
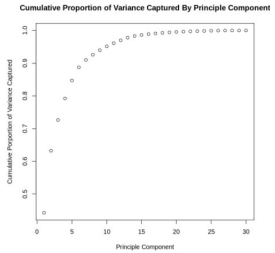
Classification Model Comparison in Breast Cancer Dataset

The Data

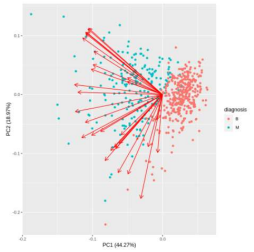


The breast cancer data set being used consists of 31 columns and 569 rows. 30 of which indicating various measurements of the tumor. One column 'diagnosis', indicates if the tumor in question was cancerous (M) or not (B). We see that there is a lot of correlation within the variables of our data set. A lot of this is due to similarity in the measurements (ex – area means vs radius mean). The heavy correlations indicates that we could likely reduce the dimensions of this data set through methods such as PCA.

Principal Component Analysis

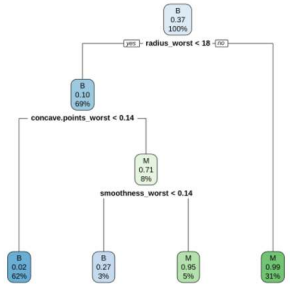


Principle Component Analysis has shown to be an effective way to reduce dimensionality of this dataset. Using the visualization on the left, we can see that our Principle Components can capture 99% of the variance in the data within 17 variables. This would reduce the number of variables by 17, without losing much of the variation. Thus, **principle components of this dataset could likely be utilized by various machine learning models and potentially increasing accuracy and reducing noise**.



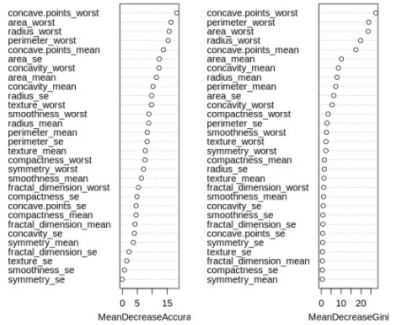
First two principle components combined in a scatter plots with vectors to indicate how the original variables influence the principle components. When adding color to indicate if the tumor was cancerous (M) or not (B), we can see that there appears to be two clusters of data along the x axis. Considering that PCA 1 accounts for 44% of the variation, **there is a reasonable difference between these clusters.**

Model 1 – CART



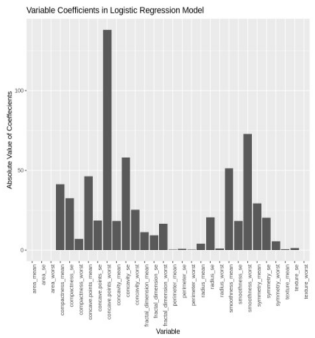
The cart model used for this classification used a binary tree consisting of the variables radius worst, concave points worst, and smoothness worst to classify whether a tumor was cancerous (M) or not (B). **This model had an accuracy of 91.81 on the testing set.**

Model 2 Random Forest



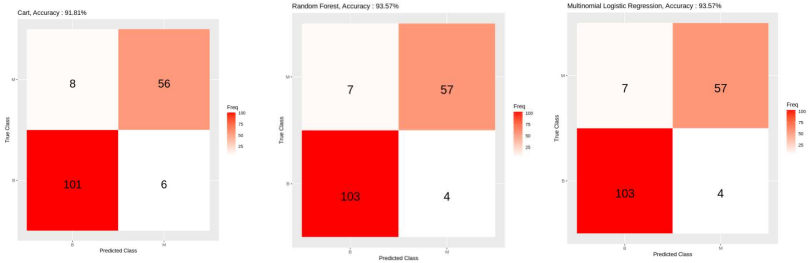
The Random Forest Model I used created 500 decision trees with 5 variables tried at each node to increase accuracy against the testing set. **This model returned an accuracy of 93.57 against the testing set**. The most important variable in this model was concave.points_worst.

Model 3 - Multinomial Logistic Regression



The multinomial regression model was built using every variable. The importance of each variable on the direction of the outcome can be visualized using the bar plot. In an attempt to improve this model, variables presented in the plot that had little impact were removed, but AIC increased, leaving the original model to be the best fit. **The multinomial regression model returned an accuracy of 93.57 against the testing set**. Concave.points_worst was the most important variable of the model

Classification Model Comparison



The best models to accurately predict if a tumor is cancerous or not with the variables given are Random Forest and Multinomial Regression, each returning the same accuracy score of 93.57. The variable Concave Points Worst was highly regarded in each model, implying that it plays an important role when classifying if a tumor is cancerous or not. To further this project, I'd like to fit the regression and random forest models using principle components.