Breast Cancer

Jasiel Feliberty
Dr. Chantal D. Larose

How to classify a mass as benign or malignant

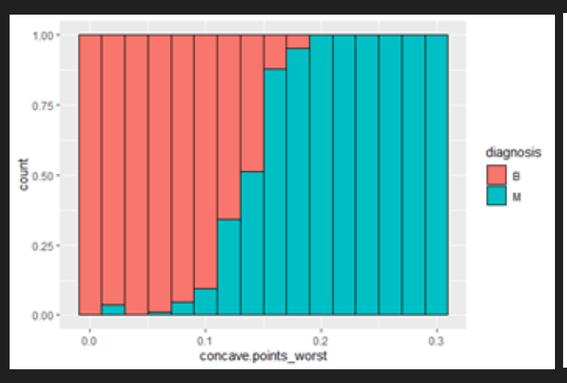
Abtract

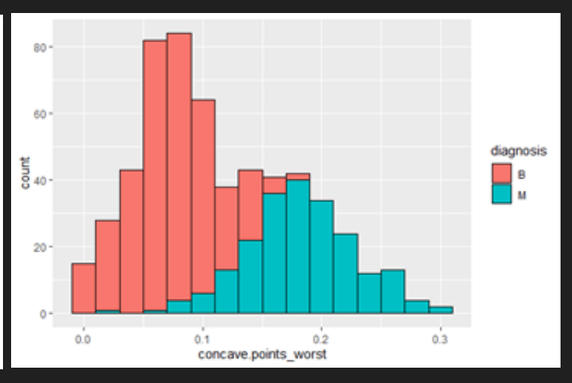
Many women in the United States develop breast cancer. In fact, about 1 in 8 U.S. women (about 12%) will develop invasive breast cancer over the course of her lifetime. There is a big difference between being diagnosed with a benign breast mass or a malignant breast mass. If it is benign it is non-cancerous. If it is malignant it is cancerous. This is why its very important to understand how a mass gets classified as benign or malignant.

Introduction to Research

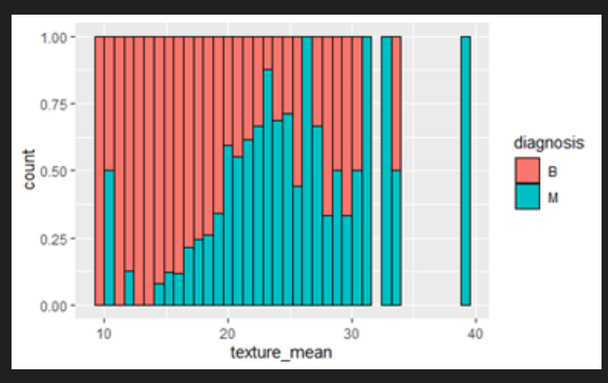
Data has been collected with all of the characteristics of a breast mass. In our reaserch, we want to clearly identify what characteristics determine the diagnosis of the mass. Finding this answer can help properly and easily diagnose if a mass is dangerous or treatable.

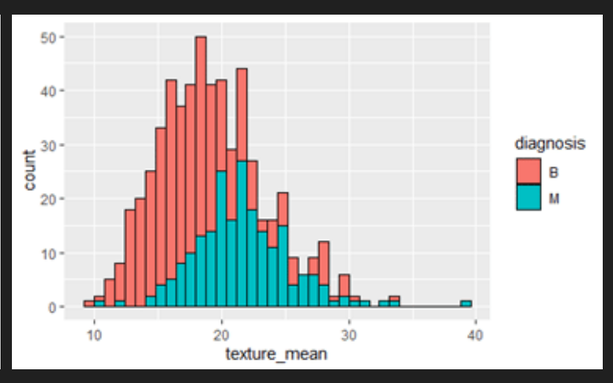
Exploratory Data Analysis





In the display to the right, we can see that the data is slightly bell-shaped for a mass being malignant. The left display shows that the proportion of a mass being malignant increases after 0.1 and continues to increase.

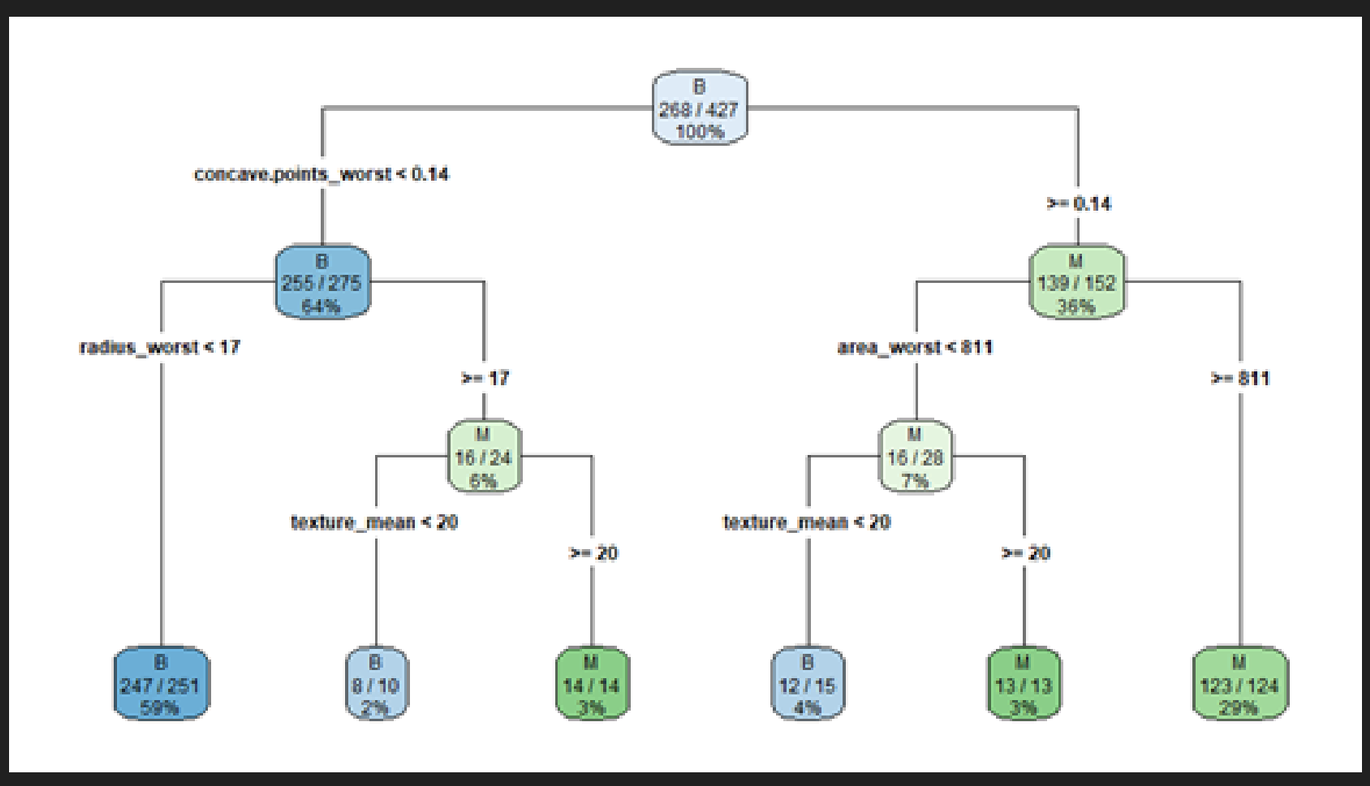




On the display to the right we can see that the data is slightly normal with two high peaks on 20 and around 22. On the display on the left, we can see that the proportion of texture mean being malignant is has several peaks and some missing values after 30.

Methodology

We want to learn about the breast mass itself. We want to identify any characteristics that the mass might have that would lead to it being malignant or benign. We also want to develop a method to identify which masses are malignant. The method we will be using is a CART model, which stands for the Classification and Regression Tree. A CART model is a binary decision tree that splits nodes into child nodes. It uses Gini index to create splits. Each of these splits will separate the data into pure child nodes in terms of the target variable.



Results

A CART model was created to classify the different characteristics of the masses. The decision tree used 4 variable (concave points worst, area worst, radius worst, and texture mean). The decision tree had 5 splits. Most of the breast masses were benign.

Train Data
Accuracy= 97.66%
Sensitivity=94.34%
Specificity=99.63%

Test Data
Accuracy= 92.25%
Sensitivity=84.91%
Specificity=96.63%

Conclusions

The CART model helped identify from the data set important characteristics that helped classify a breast mass as benign or malignant. Based on our CART model, we can see that our first split is concave points mean. Most of the concave points mean is less than 0.14 and benign, 64%. The ones that are greater than or equal to 0.14 are malignant, 36%. Both splits will continue down the tree. The 64% benign splits into radius worst. Radius worst less than 17 is benign, 59%, and ends the tree on that side. Radius worst greater than or equal to 17 is malignant, 6%, and continues down the tree. The next split is texture mean, with less than 20 benign, 2%. Texture mean greater than or equal to 20 is malignant, 3%, this ends the tree on this side. Going back to concave mean that are greater than or equal to 0.14 malignant, it splits into area worst. Area worst that is greater than or equal to 811 is malignant, 29%. Area worst that is less than 811 is malignant, 7%. This splits into texture mean, with less than 20 being benign, 4%, and greater than or equal to 20 being malignant, 3%. This concludes the tree.

Reference

Learning, UCI Machine. “Breast Cancer Wisconsin (Diagnostic) Data Set.” Kaggle, 25 Sept. 2016, www.kaggle.com/uciml/breast-cancer-wisconsin-data/discussion.

Larose, Daniel T., and Chantal D. Larose. Data Mining and Predictive Analytics. Wiley, 2015.