CART ASSINGMENT

Name: Anahi Quezada
Data: heart disease
Source: UC Irvine - Machine Learning Repository
link: "https://archive.ics.uci.edu/dataset/45/heart+disease"

RESULTS

This database contains 76 attributes, however, one of the text files in the repository has already been processed and initially cleaned, leaving 14 columns as a result.

By using this heart disease data with the predictive attribute (num) that refers to the type of chest pain (values from 1 to 4, where 1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic), I can fit a model that can determine the nature of my data, work with categorical and continuous variables and help me to build a conclusion about decision flow for patients.

The model I obtained that best fits the data and has the lowest cross-validation error is "Pruned cart" with a value of 0.3693676. However, when presenting this model with the d_test, 25% of my original data, it presents a cross-validation error of 0.5526316. With this result, I can make two observations, first is that more work is needed with the variables "ca" (number of main vessels (0-3) colored by fluoroscopy) and "thal" (3 = normal; 6 = fixed defect; 7 = reversible defect), since these were the ones that presented the most problems throughout my code due to their mix of "character", "missing value", "factor" and "number" format. Given this problem, the following models are shown to be incomplete (values such as 1, 0 or NA). The second observation is that the "Pruned cart" model does not satisfy the correct definition of a "Classification and Regression Tree". Since, both cross validation values are not similar and close enough. With these results, I can assume my models have overfitting being that "Pruned cart" do not work well for the d_test as it worked for the d_val.