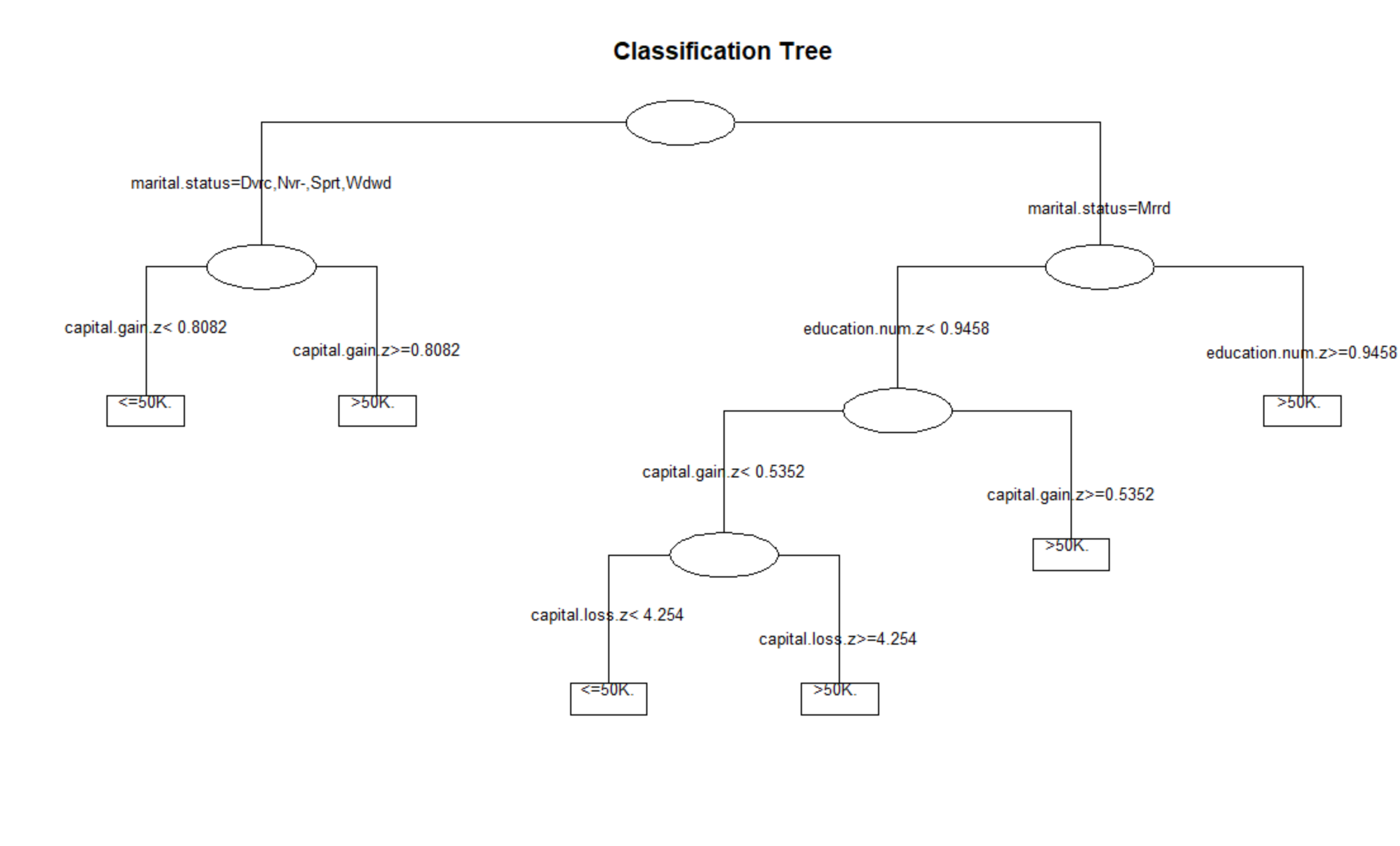
DAT 520 Problem Set 4

Michael Surdek

Southern New Hampshire University



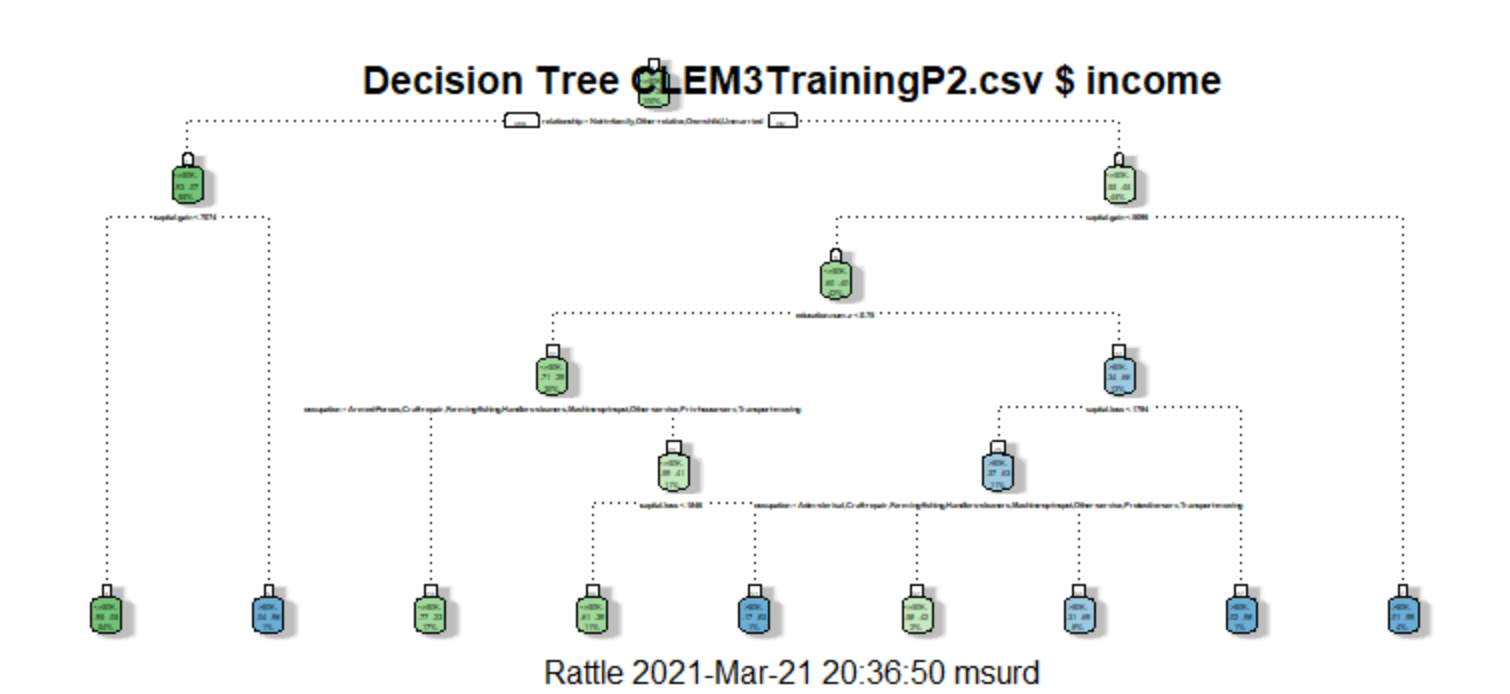
The dataset contains 25,000 observations of individuals and includes details such as age, education, marital status, income, race, native country, capital gains and losses, and weekly hours worked. The decision tree above shows us a method to predict whether an individual’s income is less or greater than $50,000 based on some of this information.

In order to formulate a prediction when looking at an individual observation, the first step in this decision tree is to determine whether they are currently married or not. Once this has been established, you would move down the tree to the right if they are married or to the left if they are not, where you will find decision nodes that are split based on the z-score of other variables. The z-score is another way of displaying the percentile where the individual falls among the population based on a numerical variable. A z-score of 0 represents the 50th percentile, and positive and negative z-scores represent percentiles above and below the 50th percentile, respectively.

The next node if the individual is not married is based on whether their capital gains are above or below a z-score of 0.8082, or about the 79th percentile. If the individual’s capital gains are below the 79th percentile, you can predict that their income is less than $50,000, and vice versa if their capital gains are above the 79th percentile.

The next node on the other side of the tree if the individual is marries is based on whether their education level is above or below a z-score of 0.9458, or about the 83rd percentile. If their education level is above the 83rd percentile, you can predict that their income is greater than $50,000. If a married person has an education level below the 83rd percentile, there is not yet enough information to make a prediction, and you must continue to the next node down to the left. At this node, if a married and less educated individual showed capital gains above a z-score of 0.5352, or about the 70th percentile, you can predict that their income is greater than $50,000.

Finally, if the individual is married, less educated than the 83rd percentile, and showed capital gains below the 70th percentile, you can proceed to the last decision node which is based on whether their capital loss was above or below a z-score of 4.254, or about the 100th percentile. If this individual’s capital loss is in the 100th percentile, you can predict that their income is above $50,000, and vice versa if their capital loss is not in the 100th percentile. This decision nodes makes logical sense, because if an individual had capital gains in the 100th percentile of such a such a large sample, they likely have an income that provided that amount of money to lose. However, this might be an example of overfitting in this decision, and this node might not be the best predictor when applied to new data.



This decision tree also formulates predictions about whether an individual in the dataset has an income greater or less than $50,000, but it uses different combinations of variables than the Cartfit tree explained previously. This tree makes distinctions between the individuals by analyzing numerous variables based on the raw values, the z-scores, or whether or not they are in a list of certain categories. The rules of this decision tree are more complex which leads to more leaves on this tree than the previous one. The first node in this decision tree is whether the relationship attribute is one of four values. If it is, a prediction can be made based on the individual’s capital gain. If it is not, variables such as capital gain, capital loss, education, and occupation must be analyzed based on how the individual fits into the decision tree.