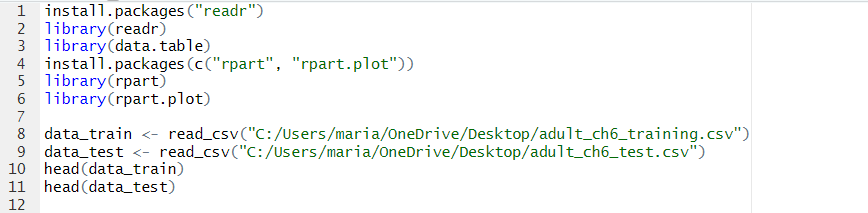
**INFS-417**

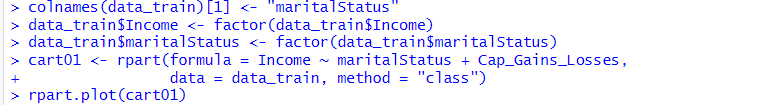
**CHAPTER 6-ASSIGNMENT 3**

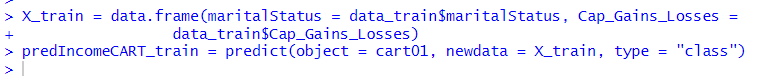
**14. Create a CART model using the training data set that predicts income using marital status and capital gains and losses. Visualize the decision tree (that is, provide the decision tree output). Describe the first few splits in the decision tree.**

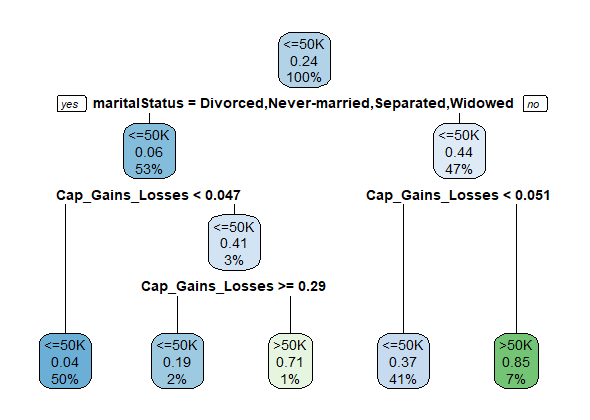
**INSTALLING LIBRARIES & IMPORTING DATA:**



**TRAINING DATASET DECISION TREE:**







**LEFT SIDE: MARITAL STATUS:** DIVORCED, NEVER MARRIED, SEPARATED, WIDOWED.

**RIGHT SIDE: MARITAL STATUS:** MARRIED

***EXPLANATION:***

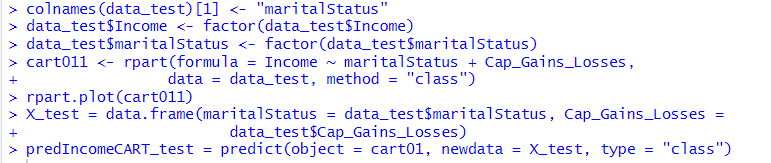
The root node starts with 100% and also tells us that 24% (0.24) of the records in the adult\_ch6\_training data set have low income (<=50 K). Thus, each node tells us the proportion of low‐income records in the node, along with the percentage of the records reaching that node. At the root node split, CART identifies the most efficacious possible binary split as separating the records into two groups, depending on their value for the variable marital status, one group whose marital status was married, and the other group consisting of all the other marital statuses: divorced, never‐married, separated, and widowed. The married group contains 44% low income, while the other group contains only 6% low income. The married group has a sevenfold lower proportion of low‐income records. This striking difference is why this split was chosen as the root node split by the CART algorithm. The root node split partitions the data nearly in half, with 47% married and 53% not married.

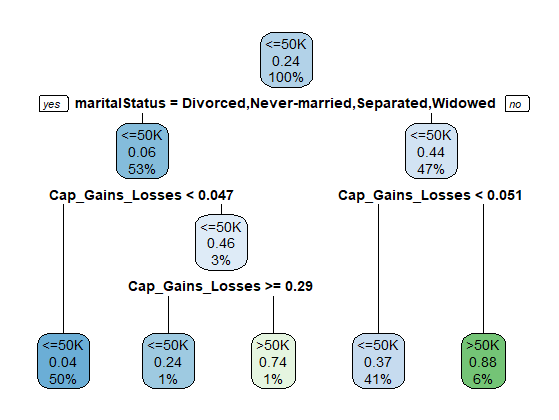
At the married node, CART then makes a second split, based on the variable, Cap\_Gains\_Losses. If the (min–max normalized) capital gains and losses exceed 0.051, then 85% of the records have higher income. However, this group of records representing married people with high capital gains and losses makes up only 7% of the training data set, as shown in the leaf node in the lower right. On the other hand, those who are married but do not have high capital gains and losses represent 41% of the data set, but contain only 37% with low income. The two nodes in the lower right are leaf nodes because no further splits are made on them.

Back at the records from the non‐married marital status node, we see that a split is then made, also based on the normalized capital gains and losses. Individuals who are non‐married and who do not have high capital gains and losses make up a full 50% of the training data set. Of these, only 4% have low income. Non‐married people with high capital gains and losses make up only 3% of the data set, but have a much higher proportion of low income: 41%.

**15. Develop a CART model using the test data set that utilizes the same target and predictor variables. Visualize the decision tree. Compare the decision trees. Does the test data result match the training data result?**

**TEST DATASET DECISION TREE:**





**LEFT SIDE: MARITAL STATUS:** DIVORCED, NEVER MARRIED, SEPARATED, WIDOWED.

**RIGHT SIDE: MARITAL STATUS:** MARRIED

***EXPLANATION:***

Yes, the testing dataset shows almost the same results as the training dataset except for the last layer where there are small differences in the margins which actually doesn’t reflect any big differences.