WQD.4 - Applying Tree-Based Methods

Sample R code for Tree-based Models and Random Forest

The response variable quality is assumed to be an ordinal variable, not a continuous variable. It has been noted before that proportions in too low (4 or less) or too high (8 or above) categories are small.

Quality Category	3	4	5	6	7	8	9
Proportion	0.4%	3.3%	29.7%	44.9%	18.0%	3.6%	0.1%

Hence wines are classified into three categories by combining 3, 4, and 5 into one category (Low), 6 (Medium) and 7, 8 and 9 into another (High).

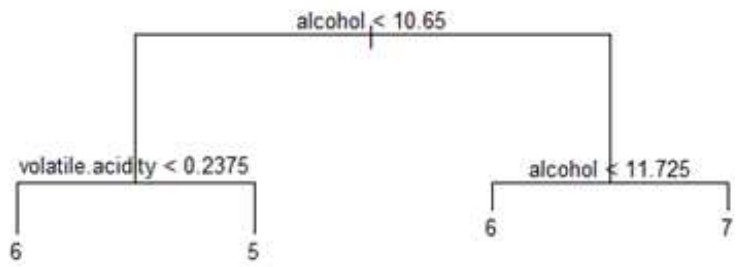
The following regression tree is obtained:

```
n = 2037

Classification tree:
tree(formula = FactQ ~ ., method = "class")
Variables actually used in tree construction:
[1] "alcohol"          "volatile.acidity"
Number of terminal nodes:  4
Residual mean deviance:  1.811 = 3681 / 2033
Misclassification error rate: 0.4502 = 917 / 2037

node), split, n, deviance, yval, (yprob)
      * denotes terminal node

1) root 2037 4301.0 6 ( 0.31026 0.46343 0.22631 )
  2) alcohol < 10.65 1107 2064.0 5 ( 0.46161 0.44896 0.08943 )
    4) volatile.acidity < 0.2375 396  764.0 6 ( 0.24495 0.58081 0.17424 ) *
    5) volatile.acidity > 0.2375 711 1161.0 5 ( 0.58228 0.37553 0.04219 ) *
  3) alcohol > 10.65 930 1832.0 6 ( 0.13011 0.48065 0.38925 )
    6) alcohol < 11.725 511 1044.0 6 ( 0.19765 0.51272 0.28963 ) *
    7) alcohol > 11.725 419  711.7 7 ( 0.04773 0.44153 0.51074 ) *
```



Applying the procedure on Test data, the following mis-classification table is obtained:

	Quality Classification		
Test Data	**Low**	**Medium**	**High**
Low	371	277	38
Medium	214	495	251
High	19	167	205
Accuracy	(371 + 495 + 205) / 2037 = 50%		

Source URL: https://onlinecourses.science.psu.edu/stat857/node/227