Homework 9
Statistical Learning, Spring Term 2019
STAT760

Download the baseball dataset from
https://vincentarelbundock.github.io/Rdatasets/datasets.html

Look for the "Hitters" dataset. The data and explanation are linked.

Download the South African Heart disease dataset from
http://web.stanford.edu/~hastie/ElemStatLearn/

The trees in the two problems below have been computed in Chapter 8 of the ISLR book. You can use the median of the data to split numerical variables instead of using intervals.

Problem 1 (10 points)

Train a regression tree for computing the log of the salary of baseball players as explained in Chapter 8.
It is up to you to decide on the number of regions covered by the leaves of the tree.

Problem 2 (10 points)

Train multiple classification trees (using bagging) to predict heart disease or not, for the second data set. Use entropy as the purity function.
Use OOB estimation to compute the error rate of the final ensemble. Test with 10, 20, 50 trees.