Assignment 5:

Question 1

In this exercise, we will use carseat dataset and seek to predict Sales using regression trees and related approaches, treating the response as a quantitative variable.

(a) Split the data set into a training set and a test set.

(b) Fit a regression tree to the training set. Plot the tree, and interpret the results. What test MSE do you obtain?

(c) Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test MSE?

(d) Use the bagging approach in order to analyze this data. What test MSE do you obtain?

(e) Use random forests to analyze this data. What test MSE do you obtain?

(f) Also, report the important features in your random forest. One can do this by using importance() function in R or feature_importance_ on a fitted model in sklearn in python. Describe the effect of m, the number of variables considered at each split, on the error rate obtained.

Question 2

In this problem, you will use support vector approaches in order to predict whether a given red wine is high-quality or low-quality based on the red-wine data set. You definitely should have some wine after completing the assignment (Please be mindful of the age-limit).

(a) Create a binary variable that takes on a 1 with quality above the mean wine quality, and 0 for quality below the mean.

(b) Fit a support vector classifier to the data with various values of penalty (cost in R, C in python), in order to predict whether a given red wine is high-quality or low-quality. Report the cross-validation errors associated with different values of this parameter. Comment on your findings.

(c) Now repeat (b), this time using SVMs with radial and polynomial basis kernels, with three different values of gamma and degree and penalty each (check cost in R, C in python for penalty). What's happening and what might be causing the change in cross validation error. Summary and comment on your choices of parameters and the corresponding cross validation errors in a table.