Data Mining in R
Assignment 2

a) Short answer questions

Q.1 - In the context of the diagnostic tests associated with linear regression models what is the meaning of a value of R^2 near 1? **[2 points]**

Q.2 - What is a dummy variable? **[2 points]**

Q.3 - What does it mean to say that a regression tree is over-fitting a data set? **[2 points]**

Q.4 - What is the goal of pruning a regression tree? **[2 points]**

Q.5 - What is the main reason why calculating the predictive performance of a model on the same data used to obtain it is not a good idea? **[2 points]**

Q.6 - What is the key issue that makes an estimate of predictive performance obtained by cross validation more reliable than another estimate calculated on the training data (the data used to obtain the model)? **[2 points]**

Q.7 - What is the conclusion we should reach when observing that some models obtained an estimated value of NMSE higher than 1? **[2 points]**

b) Assignments [16 points]

i) Use the cross validation experiments infra-structure provided by function "experimentalComparison()" of the book package to carry out a comparison of 10 different regression trees (choose the parameter settings that you want for these variants), on the 7 algae prediction problems, using the Mean Absolute Error evaluation statistic. Estimate the MAE of the 10 variants using 5 repetitions of a 2-fold cross validation process. Explore the results of this experiment using the facilities of the book package. Produce a graph showing the CV results of the model variants that were the best at any of the 7 algae.