Homework 5

K Nearest Neighbors models and Decision Trees for regression

The aim of the homework is to use kNN and Decision Trees models for predicting the outcome variable `y` in the case of a generated data set. The data for homework is in the file `hw5_studybookno.csv` (where `studybookno` is your study book number), which is attached to the assignment in Moodle.

Problem 1

1) Divide the data into training and model comparison parts (25% for model comparison).
2) Use the training data to find the best Decision Tree model.
3) Produce a graph of the 8 most important splits of the best model
4) Explain by the graph, which decisions are made to and which is the final prediction (based on the graph shown in the previous step) for observation corresponding to the row number xyz (where xyz are the last three digits in your study book number) in the original data set.
5) Copute the performance (rmse) of the best model on the comparison part.

Problem 2

1) Use the training data to find the best kNN model for predicting `y` by considering models with different number of predictors. Take into account the variable importance information from the best tree model to select predictors for the models you try, tune also with respect to the number of neighbors and with respect to the power of the distance metric.
2) Compute the performance of your best model for the comparison data. Which model (from the Decision Tree model and kNN model) would you choose for predicting new data and why?

I remind you that you are expected to solve the homework problems by yourself, sharing solutions (and copying parts of other student's solutions) is not allowed.