Data Mining in R
Assignment 4

a) Short answer questions

Q.1 - On which situations the median is different from the shorth? **[2 points]**

Q.2 - What is the problem addressed in feature selection and why it is necessary? **[2 points]**

Q.3 - What is the basic principle behind the Anova filter selection method? **[2 points]**

Q.4 - What are misclassification costs? **[2 points]**

Q.5 - How does Leave One Out Cross Validation works? **[2 points]**

Q.6 - Someone tells you that have a model that obtained 90% classification accuracy on a separate test set. Is this enough information for you to conclude that this is a good model? **[2 points]**

Q.7 - Comment on the following sentence: "the presence of irrelevant variables may have a negative impact on the performance of k-nearest neighbors". **[2 points]**

b) Assignments [16 points]

Use the data of the case study used in this lesson for a small experiment of comparing several variants of kNN and random forests, using leave one out cross validation. Regards the problem of having too many features on the original data, simply choose 30 genes randomly. Having this filtered data set with 30 random genes, compare a few variants of random forests (varying for instance the parameter that controls how many trees make up the ensemble), with a few other variants of k-Nearest Neighbours (varying the number of neighbours). For the comparison use the LOOCV routines available in the book R package and explained in the chapter supporting this lesson. Check and comment the obtained results, in particular in comparison with the results shown in the book, to observe the impact of the random feature selection you have used.