Name/Student Registration Number:

1. (10 points) Please select which statement is correct or wrong:

MLR requires collinear variables	O correct	O wrong
ANOVA tests for equal variances	O correct	O wrong
Please assign the dendrograms A, B, C, and D to the datasets 1, 2, 3, and 4:	O correct	O wrong
Forward selection can be used to find outliers	O correct	O wrong
The VIF of a particular variable has nothing to do with PRESS	O correct	O wrong
The confusion matrix can be used to calculate the probabilities of type 1 and type two error	O correct	O wrong
Confounding cannot be reduced by randomisation of the observations	O correct	O wrong
The order of a model is equal to the number of non-zero eigenvalues	O correct	O wrong
The kMeans algorithm can be used to estimate the means of k neighbors	O correct	O wrong
PLS/DA stands for "PLS-based Data Analysis"	O correct	O wrong
The F value of an MLR model can be calculated from the coefficient of determination if the number of variables and objects are known	O correct	O wrong
MLR can be used to determine parameter b in the model y = a*sin(b*x)	O correct	O wrong
The model error can be obtained by stepwise regression	O correct	O wrong
The reliability of an MLR model can be checked by cross validation	O correct	O wrong
Blocking can be used to control confounding variables	O correct	O wrong
The ratio of the number of variables to the number of object influences overfitting in MLR models	O correct	O wrong
Majority voting is used in connection with dendrograms	O correct	O wrong
ANOVA can be used to detect effects of factors on the independent variable	O correct	O wrong
Random generators have to be used for randomizations	O correct	O wrong
LDA is based on multilinear regression	O correct	O wrong
Variables which exert an undesired influence on the dependent variable are confining variables	O correct	O wrong
Treatments are conditions for a given experiment	O correct	O wrong
The order of a model can be recognized in a scree plot	O correct	O wrong
PLS can be used for problems with more variables than observations	O correct	O wrong
Fractional factorial designs reduce the number of required experiments	O correct	O wrong
The significance of an MLR coefficient can be calculated from the ratio of its standard error to its value	O correct	O wrong
Linear discriminant analysis maximizes the t-value between two groups	O correct	O wrong
ANOVA is used to check the overall reliability of an MLR model	O correct	O wrong
PRESS is used to detect multi-collinearity	O correct	O wrong
Balanced experiments have the same number of cases per treatment	O correct	O wrong
Multi-collinearity means that the multi-collinear variables can be related to each other by linear equations	O correct	O wrong
Overfitting occurs if there are more objects than variables	O correct	O wrong
The inverse ratio of a regression parameter to its standard error is t-distributed	O correct	O wrong
The general Minkowski distance includes the Mahalanobis distance	O correct	O wrong
PLS1 calculates a model having only one input variable	O correct	O wrong
Heteroscedasticity implies that the variance of repeat measurements depends on y-hat	O correct	O wrong
Homoscedasticity means equal means of repeat measurements	O correct	O wrong
PCR stands for "Pricipal Component Reduction"	O correct	O wrong
MLR applied to principal component scores results in PCR	O correct	O wrong
Majority voting requires an odd number of nearest neighbors to be unambiguous	O correct	O wrong
For the analysis of variances the variances of the compared factor levels have to be equal	O correct	O wrong
Variable selection can be used to fight the curse of dimensionality	O correct	O wrong
The Mahalanobis distance takes correlations into account	O correct	O wrong
PRESS is the square root of RMSEP	O correct	O wrong
The Lance-Williams equation controls the type of hierarchical clustering	O correct	O wrong
Confounding variables are unimportant in experimental designs	O correct	O wrong
A Latin square can be used to set up an experimental design	O correct	O wrong
Cluster analysis is based on distances in the p-dimensional space	O correct	O wrong
Stepwise regression always results in the best set of variables	O correct	O wrong
Fractional factorial designs require less variables than full experimental designs	O correct	O wrong

2. (3 P) Explain the advantages and drawbacks of kNN-based models

3. (3 P) What are the main assumptions of multiple linear regression? How can you check each of these assumptions?

4. (2P) What is a confusion matrix? Draw an example and explain the particular cells of the matrix.

5. (2 P) Draw the dendrogram of the following data: