**Section C**

**An analysis aiming to classify two types of lily based on characteristics of stem length, petal width and petal length.**

The file "flowers.Rda" contains an R data frame called flowers. This contains 4 columns:

**petal.length** stores petal length in cm

**petal.width** stores petal length in cm

**stem.length** stores petal length in cm

**type**  stores flower type as an integer: 0 for pink lily and 1 for yellow lily

Our aim is to build a model that can classify the flower type using the predictor variables.

1i) Load in the data to your R environment.

(1 mark)

ii) Add a new column **type.factor** to the data frame that stores the flower type as a factor with appropriate level labels.

(2 marks)

iii) Make three figures that show boxplots to illustrate how the distribution of petal width, petal length and stem length differ for the two types of flowers.

(3 marks)

iv) Comment on which variable(s) seem most suited for discriminating between the two flower types.

(1 mark)

v) Perform a t.test to check if we can reject the null hypothesis that the populations of theyellow and pink lily distributions have the same mean stem length.

Comment on your results.

(2 marks)

vi) Make a scatter plot of petal length vs petal width, using different colours to distinguish the flower types. Label your plot axes appropriately.

(4 marks)

**We will now perform logistic regression to classify the flowers using petal width and petal length as our predictors.**

2i) Divide the data set into two data frames of equal length to create train and test data sets.

Ensure that the split is randomised as the loaded data frame is ordered by flower type.

(2 marks)

ii) Perform a Logistic regression on the **training** data to produce a fitted model that can classify the lily type based on petal width and length.

(3 marks)

iii) Make a new column in the **test** data frame that stores the flower type predicted from the fit.

(2 marks)

iv) Create the confusion matrix, that shows the result of the classification, in terms of the predicted and actual lily types on the **test** data.

(2 marks)

v) Calculate the misclassification rate on the test data and on the training data.

(3 marks)

**We will now use the method of KNN to classify the flowers.**

3i) Create the following data frames in preparation for performing KNN analysis:

1. marks)

- a dataframe containing all training rows, and columns for petal width and petal length only

- a dataframe containing all training rows, and factor column of petal type only

- a dataframe containing all test rows, columns for petal width and petal length only

ii) Carry out KNN analysis using K=10 and store the resulting predictions.

(3 marks)

iii) Calculate the misclassification rate of the resulting predictions.

(2 marks)