ST4061 – Computer Intensive Statistical Analytics II
ST6041 – Machine Learning and Statistical Analytics II
2022-23
In-class test 1

NAME AND SURNAME: ……………………………………………………………………………………..

STUDENT NUMBER: ………………………………………………………………………………………..

PROGRAM: …………………………………………………………………………………………………

INTRUCTIONS

- Provide your answers in this document, after each question.
- Paste the R code you used for each question item.
- **Save your files regularly.**

Question 1

Load the following libraries and dataset into your R session as follows:

```
library(MASS)
library(caret)
ca.train = read.csv("ca_train.csv",stringsAsFactors=TRUE)
ca.test = read.csv("ca_test.csv",stringsAsFactors=TRUE)
```

This dataset contains a sample of records from 500 young soccer players after their one-year stay at a soccer academy, with variables:

- y: gets hired by a club on a trial contract as a professional player
- academy: whether the player received academic training previously
- entry: player score sheet after entry test
- exit: player score sheet after exit test

Your goal is to build a prediction model in order to predict variable y.

(1) Fit an LDA model using all predictors to the training data (ca.train). Provide the corresponding confusion matrix obtained for the test data (ca.test).

Your answer:

R code for (1):

(2) Fit a QDA model using all predictors to the training data (ca.train). Provide the corresponding confusion matrix obtained for the test data (ca.test).

Your answer:

R code for (2):

(3) Evaluate and quote the specificities of both models. Comment on the values you obtain.

Your answer:

R code for (3):

(4) Explain the difference in specificities, using relevant R output.

Your answer:

Question 2

No code is required for this question. A student is considering the dataset depicted in Figure 1, comprising of two predictors X_1 and X_2, and one categorical dependent variable Y with 2 categories.

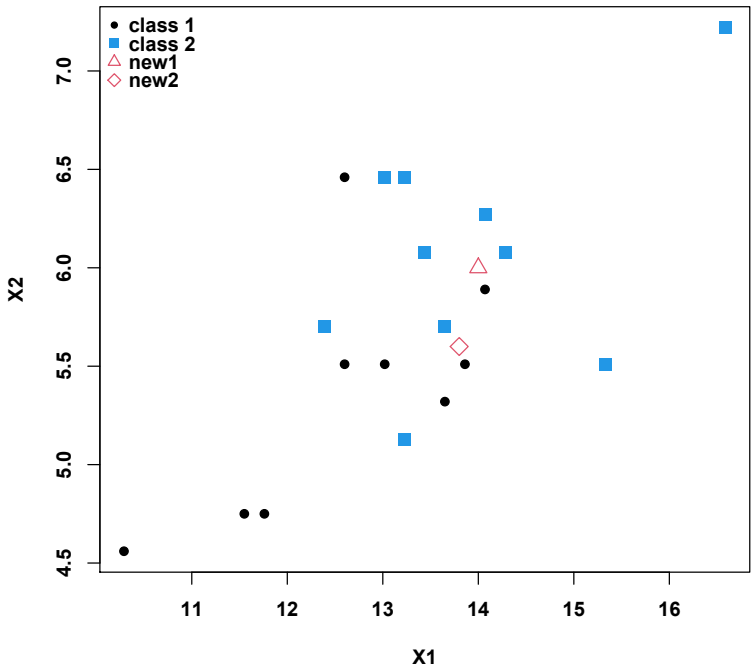

Figure 1 - Dataset for Question 2, with 2 test points new_1 and new_2.

The student is considering using a kNN classifier with k=3 to generate predictions for test points new_1 and new_2 in Figure 1. Indicate in the table below what these predictions would be.

Your answer:

kNN with k=3	**Prediction**
Test point new_1	
Test point new2	

Question 3

For this question you are required to use the following packages:

```
library(ISLR) # for the data
library(gbm)
library(randomForest)
```

Do **not** use the **caret** package, or any other package except for the above packages, for this question.

Start by creating the following dataset:

```
x.train = Khan$xtrain
x.test = Khan$xtest
y.train = as.factor(Khan$ytrain)
y.test = as.factor(Khan$ytest)
```

a) Provide a tabular summary of the distribution of values in the train and test samples of observations **y.train** and **y.test**. Comment briefly on these distributions.

b) Is this a regression or a classification problem? Justify your answer.

c) Fit a random forest to the training data defined above, using default hyperparameter values from the randomForest package. Quote the corresponding confusion matrix. Run the instruction set.seed(4061) before you run the R code for this question.

d) Generate predictions for the test data from the random forest model fit in (c) and quote the test set prediction accuracy.

e) Name the features in this dataset whose variable importance is strictly greater than 0.4, according to the fit from (c).

f) Provide an interpretation for the measure of variable importance used in (e).

g) Fit a gradient boosting model to the training data, using default hyperparameter values from the gbm package. Run the instruction set.seed(4061) before you run the R code for this question. Generate predictions for the test set from the GBM fit from (g) (do not provide these predictions in your answer) and quote the corresponding prediction accuracy.

Your answer:

Question item	Answer
(a)	
(b)	
(c)	
(d)	
(e)	
(f)	
(g)	

R code for Question 3

```
<paste your R code here>
```