Problem

1) (**10 marks**) Consider the dataset "datacl.csv". Divide randomly the data in training and test set with the test set containing one-third of the observations. Plot the training set data (x1, x2 as the x-axis and y-axis, respectively and class label of y by different colours). Fit logistic regression model: y on x1, x2 using the training set. Report the estimates, test the hypothesis for significance of each variable as well as the whole model, confidence intervals, p-values. Compute the predicted probabilities based on the estimates for both the classes on the test set and assign observations with probability greater than or equal to 0.5 to the class "1". Show the 2×2 confusion matrix with observed and predicted class labels for the test set.

Use leave-one out cross validation with squared error loss function with $L(y_i, p_i) = (y_i - p_i)^2$, where y_i is the observed class of ith response and p_i is the predicted probability of the response being 1, on the whole dataset(training+test) to choose among the following models:
a. $y \sim x1$
b. $y \sim x2$
c. $y \sim x1 + x2$
Which model is chosen ?

2) (**4 marks**) In the dataset "datacl.csv", fit LDA and QDA on the training set considered in Problem 1. Show the confusion matrix for the training and the test set. Report the estimated parameters. Choose among LDA and QDA based on the misclassification error (average zero-one loss) on the test set.

3)(**5 marks**) Consider a dataset to compare the life of cellphones obtained from 2 companies 10 observations were collected : 5 from the cellphone form company A (A1, A2, A3, A4, A5) and 5 from the cellphone of Company B (B1 ,B2, B3, B4, B5).
The following were the observations:
Cellphone A: A1, A2, A3 stopped working after 1.1, 2.2 and 3.3 years, respectively. A4 and A5 were working for at least 4 years.
Cellphone B: B2, B3 stopped working after 1.5 and 3.5 years, respectively. B1 was working for at least one year. B4 and B5 were working for at least 4 years.

a. Which group has the higher median survival time.
b.) Consider the dataset from a company Cellphone C: C2, C3 stopped working after 1.5 and 3.5 years, respectively. C1 was working for at least one year. C4 was working for at least 2 years. C5 was working for at least till 5 years. Can you compare the median of groups A and C? If yes, how and if not, why ?

c.) Combine the dataset of Cellphone A and cellphone B. What is the survival probability at 2.25 years? What is the survival probability at 3.9 years?

4. (**3 marks**) For each example, state whether or not the censoring mechanism is independent. Justify your answer.
(a) In a study of disease relapse, due to a careless research scientist, all patients whose phone numbers begin with the number "2 are lost to follow up.
(b) In a study of longevity, a formatting error causes all patient ages that exceed 99 years to be lost (i.e. we know that those patients are more than 99 years old, but we do not know their exact ages).
(c) Hospital A conducts a study of longevity. However, very sick patients tend to be transferred to Hospital B, and are lost to follow up.

5. (**8 marks**) In the dataset "datasur.csv", the data for the survival of cancer patients are provided. Plot the Km curve showing the survival probabilities for male and female patients. Using log-rank test, check whether there is a significant difference between male and female patients at level of significance 0.05. Show the summary using cox-proportional hazard model. If the instantaneous probability of dying for a female patient at age 40 is 0.000001, can you provide an estimate of the instantaneous probability of dying for a male patient at age 50?