UNIVERSITY EXAMINATIONS
2022/2023 ACADEMIC YEAR

SPECIAL/SUPPLEMENTARY EXAMINATIONS
YEAR ONE SEMESTER ONE EXAMINATIONS

FOR
THE DEGREE OF MASTER OF SCIENCE
(COMPUTER SCIENCE)

COURSE CODE : MCS 823

COURSE TITLE : STATISTICAL MODELLING AND COMPUTING

DATE: 15/08/2023 TIME: 2.00 P.M – 5.00 P.M

INSTRUCTIONS TO CANDIDATES:

ATTEMPT ANY THREE (3) QUESTIONS

QUESTION ONE [20 MARKS]

Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n (sample size) and p (predictors).

a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary. [6 marks]

b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables. [7 marks]

c) We are interesting in predicting the % change in the US dollar in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the dollar, the % change in the US market, the % change in the British market, and the % change in the German market. [7 marks]

QUESTION TWO [20 MARKS]

Consider the following lines from the text Romeo and Juliet with punctuations removed. Assume the lines are stored in a text file named `Juliet.txt`

```
But soft what light through yonder window breaks
It is the east and Juliet is the sun
Arise fair sun and kill the envious moon
Who is already sick and pale with grief
```

a) Write Python code to prompt for the file name and open it for reading. Account for the case where the file does not exist. [4 marks]

b) Write a program to count how many times each word appears in the file [6 marks]

c) Write a program to count how many times each character appears in the file. [4 marks]

d) Write a program to write to a file all characters which occur more than 5 times in the text file above [6 marks]

QUESTION THREE [20 MARKS]

a) In a certain community 8% of all adults over 50 years of age have diabetes. Suppose that a health service in this community correctly diagnoses 95% of all persons with diabetes as having the disease. It also incorrectly diagnoses 2% of all persons without diabetes as having the disease. A person over the 50 years of age is selected at random from this community.

i. Find the probability that the community health service will diagnose this person as having diabetes. (5 marks)

ii. Given that this person is diagnosed by the health service as having the diabetes, what is the probability that he really has diabetes? (6 marks)

b) In a random experiment it is given that P(A′UB) = 0.55, P(B) = 0.50 and P(A′UB′) = 0.77. Calculate the following IN THE GIVEN ORDER:

i. $P(A \cap B)$ (3 marks)

ii. P(A) (4 marks)

iii. $P(A' \cap B)$ (2 marks)

QUESTION FOUR [20 MARKS]

a) Distinguish between a flexible statistical learning method and an inflexible method. [4 marks]

b) For each of parts (i) through (iv), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

i). The sample size n is extremely large, and the number of predictors p is small. [4 marks]

ii). The number of predictors p is extremely large, and the number of observations n is small. [4 marks]

iii). The relationship between the predictors and response is highly non-linear. [4 marks]

iv). The variance of the error terms, i.e. $\sigma^2 = \mathrm{Var}(\epsilon)$, is extremely high. [4 marks]

QUESTION FIVE [20 MARKS]

The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	−1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3$ $= 0$. [5 marks]

b) What is our prediction with $K = 1$? Why? [5 marks]

c) What is our prediction with $K = 3$? Why? [5 marks]

d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the *best* value for K to be large or small? Why? [5 marks]