UNIVERSITY OF LONDON

CM3005

BSc EXAMINATION

COMPUTER SCIENCE

Data Science

Release date: Tuesday 14 March 2023 at 12:00 midday Greenwich Mean Time

Submission date: Wednesday 15 March 2023 by 12:00 midday Greenwich Mean Time

Time allowed: 24 hours to submit

INSTRUCTIONS TO CANDIDATES:

Section A of this assessment paper consists of a set of **TEN** Multiple Choice Questions (MCQs) which you will take separately from this paper. You should attempt to answer **ALL** the questions in Section A. The maximum mark for Section A is **40**.

Section A will be completed online on the VLE. You may choose to access the MCQs at any time following the release of the paper, but once you have accessed the MCQs you must submit your answers before the deadline or within **4 hours** of starting whichever occurs first.

Section B of this assessment paper is an online assessment to be completed within the same 24-hour window as Section A. We anticipate that approximately **1 hour** is sufficient for you to answer Section B. Candidates must answer **TWO** out of the THREE questions in Section B. The maximum mark for Section B is **60**.

Calculators are not permitted in this examination. Credit will only be given if all workings are shown.

You should complete **Section B** of this paper and submit your answers as **one document**, if possible, in Microsoft Word or a PDF to the appropriate area on the VLE. You are permitted to upload 30 documents. However, we advise you to upload as few documents as possible. Each file uploaded must be accompanied by a coversheet containing your **candidate number**. In addition, your answers must have your candidate number written clearly at the top of the page before you upload your work. Do not write your name anywhere in your answers.

UL23/0216

SECTION A

Candidates should answer the **TEN** Multiple Choice Questions (MCQs) quiz, **Question 1** in Section A on the VLE.

SECTION B

Candidates should answer any **TWO** questions from Section B. No handwriting is allowed.

Question 2 – 30 marks

(a) You are trying to apply machine learning algorithms to predict (i) house prices and (ii) measure customer habits. Which algorithms would you use and what type of data should you have to perform these tasks? **[8]**

(b) Morpheus has 2 identical boxes of pills. The first box contains 7 red and 3 blue pills while the second box contains 12 red and 8 blue pills. A pill is drawn at random from one of the boxes and it is found to be red. What is the probability that it was drawn from the first box? **[8]**

(c) Your colleague gathers data on the patient histories and their clinical outcomes. You notice that some records have missing values. What is the preferred method for dealing with missing data? When would you perform ffil over a bfill and why these techniques are limited? In SKlearn library there are several imputation methods. Describe 2 of them. **[6]**

(d) List the challenges you might encounter in tokenizing and identifying named entities in the following sentences: **[8]**

```
Mr. Martin O'Neill was Nottingham Forest's manager for the period
January-June in the 2018-19 season. Result: it didn't work out...
```

Question 3 – 30 marks

(a) According to Bayes Theorem, what is prior probability and what is likelihood? **[4]**

(b) A classifier that predicts if an image contains a cat, a dog, or a rabbit produces the following confusion matrix:

		True values		
		Dog	Cat	Rabbit
Predicted values	Dog	20	3	1
	Cat	6	16	2
	Rabbit	4	1	7

Calculate the following:

- Precision and recall for each class **[6]**
- Average precision and average recall **[4]**
- F measure for the each class **[6]**

(c) Give an example where a false positive may be more costly than a false negative. **[5]**

(d) Give an example where a false negative may be more costly than a false positive. **[5]**

Question 4 – 30 marks

(a) Briefly explain the steps you might take to build a bag-of-words model for the following sentence: **[4]**

```
Noshin likes to watch football, especially Nottingham Forest.
Essi likes football too.
```

(b) You are given a dataset on detection of a rare medical condition and have built a classifier which returns an accuracy of 95 percent. Why might this figure be misleading? Suggest 3 strategies for dealing with this situation. **[4]**

(c) A diagnostic test has a true positive rate of 100% and a false positive rate of 3%. The condition it detects exists in the population with a rate of one in 1,000. Given that your test is positive, what is the probability of you having the condition? **[4]**

(d) Suppose that you are building a model to identify topics in a corpus of 1 million news stories. What techniques could you use to reduce the dimensionality of the data? How might you use TF.IDF to represent the documents? What kind of visualization techniques might you use to illustrate the output? **[8]**

(e) Describe the processes of stemming and lemmatization, giving examples of each. In what ways do stemming and lemmatization differ in their use of linguistic context? Discuss the pros and cons of each approach. **[10]**

END OF PAPER