School of Computing and Information Technology

Student to complete:

Family name	
Other names	
Student number	
Table number	

CSCI316
Big Data Mining Techniques and Implementation

Final Examination Paper
Session 1 2022

Exam duration	3 hours
Weighting	50% of the subject assessment
Marks available	50 marks
Directions to students	6 questions to be answered. Each question contains multiple sub-questions. Marks for each sub-question are indicated.

Question 1 (5 marks)

Given a list named X which contains words (as strings), implement a Python function to compute the word(s) with the highest frequency in X. Write down the Python code.

(5 marks)

Question 2 (7 marks)

(2.1) Explain the advantages of stratified sampling over standard random sampling.

(3 marks)

(2.2) Describe three common ways of handling missing values.

(4 marks)

Question 3 (12 marks)

(3.1) Assume that you are given a set of records as shown in the following table, where the last column contains the target variable. Present the procedure of using Gini index to identify which attribute should be split. You need to show all steps of your calculation in detail.

(6 marks)

Record ID	***Size of class?***	***Lecturer experience***	***Programming Subject?***	***Student satisfaction***
1	*Large*	*Strong*	*No*	*Low*
2	*Small*	*Weak*	*No*	*Low*
3	*Average*	*Weak*	*Yes*	*Low*
4	*Small*	*Weak*	*Yes*	*Low*
5	*Average*	*Strong*	*No*	*High*
6	*Small*	*Strong*	*No*	*High*
7	*Small*	*Strong*	*Yes*	*High*
8	*Large*	*Weak*	*Yes*	*High*

(3.2) Present the pseudo-code of a decision tree induction algorithm for a data set with categorical and continuous features. You can also support the pseudo-code with explanations

(6 marks)

Question 4 (8 marks)

(4.1) Explain in which situations sensitivity and specificity are more important than accuracy as performance metrics of a classifier.

(4 marks)

(4.2) Assume that a Bayesian classifier returns the following outcomes for a binary classification problem, which are sorted by decreasing probability values. P (resp., N) refers to a record belonging to a positive (resp., negative) class.

Tuple #	Class	Probability
1	P	0.90
2	P	0.80

3	P	0.70
4	N	0.60
5	P	0.55
6	N	0.54
7	P	0.53
8	N	0.51
9	N	0.50
10	N	0.40

What is the largest true positive rate when the false positive rate equals 0.4, and what is the smallest false positive rate when the true positive rate equals 0.8? Present the process of your calculation.

(4 marks)

Question 5 (8 marks)

(5.1) Why is Apache Spark suitable for data-parallel computation? What is the bottleneck of model-parallel computation for Apache Spark? Also use an example to support your answer.

(4 marks)

(5.2) Assume that a DataFrame named `FlightsDF` of flight statistics is defined in PySpark, with the following code processed.

```
FlightsDF.printSchema()
Out:
root
 |-- DEST_CITY: string (nullable = true)
 |-- DEST_COUNTRY_NAME: string (nullable = true)
 |-- ORIGIN_CITY: string (nullable = true)
 |-- ORIGIN_COUNTRY_NAME: string (nullable = true)

DF.show(2)
Out:
+----------------------+-----------+--------------+
|DEST_CITY|DEST_COUNTRY|ORIGIN_CITY|ORIGIN_COUNTRY|
+---------+------------+-----------+--------------+
|Sydney   |Australia   |Melbourne  |Australia     |
|Auckland |New Zealand |Singapore  |Singapore     |
+---------+------------+-----------+--------------+
only showing top 2 rows
```

Based on `FlightsDF`, write down the code in PySpark to implement the following operation: Find the country or countries with most *domestic* flights. (Note. A domestic flight has the same original and destination country.)

(4 marks)

Question 6 (10 marks)

(6.1) Why a classical Perceptron (i.e., a single layer of linear threshold units) is not preferable to use? Provide your reasons.

(3 marks)

(6.2) Based on the transfer learning pipeline example in the file named “11 Spark to TensorFlow: Transfer learning pipeline” on this subject’s Moodle site (which can also be directly accessed via the following link), explain the advantages of transfer learning.

https://documents.uow.edu.au/~guoxin/CSCI316/Petastorm-Spark-Converter-TensorFlow.html

In particular, discuss the situations when transfer learning is most applicable.

Note. Your answer must relate to the example; otherwise no mark will be provided.

(7 marks)

End of Examination