THE UNIVERSITY OF NEW SOUTH WALES
School of Computer Science and Engineering

Final Examination– Term1, 2020

5th May, 2020

COMP9321 Data Service Engineering

Total Exam Mark: 40

Total Number of Questions: 30 + 8

Exam Duration: 24 Hours

**** IMPORTANT NOTICE****

There are Two parts in this exam paper: Part A - Multiple Choice Questions, Part B - Written Answer Questions. Plan your time wisely and attempt to complete all parts.

You may submit your solutions as many times as you like. The last submission ONLY will be marked.

Questions (and sub-questions) are not worth equal marks. Answer all questions.

For multiple choice questions select the response which best answers the question. Keep your written answers clear and coherent. Messy or irrelevant answers will not be marked.

The Answers need to be according to your own effort and in your own words. If you do not follow these instructions, you will get zero marks for the exam and a possible charge of academic misconduct.

PartA: Multiple Choice Questions (Total 15 Marks)

Use Moodle Quiz to Answer all the 30 Questions. The last submission is going to be marked.

https://moodle.telt.unsw.edu.au/mod/quiz/view.php?id=2985563

PartB: Written Answer Questions (Total 25 Marks)

The written Answer Questions Paper is to be submitted using Give System as a PDF file named z{id}.pdf

Question1: (4 Marks)

An organization has two datasets one for devices, their location, the operator, and quality tests; and the other is about the technical support tickets opened for each device. The organization want to draw some insights in regard to the opened support tickets for each device and the relation with when the device was quality tested and who is the operator.

In the light of the datasets snippets shown below, what pre-processing (cleansing and manipulation) is needed to make sure that the organization can conduct the required task. Explain each step in the light of the datasets provided. You can use Python code, pseudo code, or you can explain as a series of steps. In the case of using code there is no need to preserve the syntax but it is a MUST to include proper commenting to explain each step. Be advised that the organization is low on resources (e.g., storage), so that need to be considered in the pre-processing.

Dataset 1			
Device ID	Quality Tested Date/Time	Location	Operator
B1834	2019-01-16:23:59:12	K17-401-08	Albert
B9872	2019-01-03:09:15:17	K17-401-08	Albert
N2543	2019-01-27:06:39:01	K17-502-12	Jill
n/a	2019-01-18:06:39:01	NaN	NaN
M4328	2019-03-27:09:30:01	K17-401-09	Chris
B9872	2019-01-29:08:19:17	K17-401-08	Albert

Dataset 2		
Device ID	Support Ticket Date/time	Ticket Handled by
B1834	2019-21-01:11:59:12 AM	Morty
N2543	2019-01-03:03:39:01 PM	Morty
M4328	2019-23-05:01:30:01 PM	Morty
B9872	2019-16-03:08:19:17 AM	Morty
M4328	2019-23-05:01:30:01 PM	-

Question2: (4 Marks)

Suppose a shop owner wants to divide her customers into different groups. She has the number of purchases they made in the last year and based on the number of purchases, she wants to segment them into groups. There is no fixed target here as to how many groups to have. the shop owner does not know what type of customers should be assigned to which group. Below is a data sample.

Shopper ID	Purchases Made
A	18
B	7
C	22
D	12
E	24

A) What Machine learning Algorithm are you going to use to solve this problem? Why?
B) Illustrate the calculation steps of how the algorithm is going to work and groups are going to be formulated. Explaining each step.
C) Explain how you will determine the number of Groups eventually

Question3: (2 Marks)

Suppose we want to compute 10-Fold Cross-Validation error on 100 training examples. We need to compute error N1 times, and the Cross-Validation error is the average of the errors. To compute each error, we need to build a model with data of size N2, and test the model on the data of size N3.

What are the appropriate numbers for N1, N2, N3? Why?

Question4: (2 Marks)

Consider the following confusion matrix

		Current Answer	Current Answer
		True	False
Predicted Answer	True	8	2
Predicted Answer	False	12	11

For the above “confusion matrix” what is the precision, recall and F1-score? Illustrate the procedure for the calculation.

Question5: (3 Marks)

An organization is considering allowing the consumption of the data they have through a REST API. The organization want to only allow the consumption of their API to authenticated users. They are a security conscious organization and they want to make sure to minimize the attack window if the credentials are leaked.

A) What would you advise them to use for REST API authentication? And Why? Explain the authentication scheme with example

B) Additionally, If the organization want to track the usage of their API and do some rate limiting. What would you advise them to use? Why? Explain with an example.

Question6: (2 Marks)

Consider the following HTTP request invoking a POST method of a RESTful API:

```
POST /orders HTTP/1.1

Host: api.coffeehouse.com

Content-Type: application/xml

<order>

<drink>latte</drink>

</order>
```

Write down the content of the HTTP response that you would return as the result. Explain your answer.

Question7: (4 Marks)

Let's assume that a certain pandemic due to an infectious virus (let's call it MOVID-99) has caused some social isolation restrictions in the country. The Government in an attempt to track the activities of people who are infected and determine if they have been in contact with other people (who might be infected as well) released an App to be installed on people mobile devices(let's call it MOVID-Saffe). Assume that the App records the data (snippet shown in table 1). You have access to the App recorded data in addition to data from the Health Department (snippet shown in Table2). Assume that it is currently unknown how the virus spread and who could be at risk and why, you need to use machine learning to predict who are the people at risk.

A) What Machine Learning Model are your going to Choose to solve this problem? Explain why in details and mention any limitations if any.

B) What are the features you are going to use (and if you need any transformation)

Table 1

Identifier of people close to Device	Distance (meter)	Duration of contact (minutes)	GPS Location	Timestamp (DD/MM/YY;HH:MM:SS)
X11992291	2	10	41°24'12.2"N 2°10'26.5"E	03/02/2020; 13:00:10
N72783912	8	2	32°24'12.2"W 2°10'26.5"N	03/02/2020; 13:30:12
X11992291	1.5	30	41°24'12.2"N 2°10'26.5"E	03/02/2020; 12:00:10
L11892277	3	20	44°39'12.2"N 2°11'26.5"W	04/02/2020; 11:20:44
Z98192922	6	3	41°24'12.2"N 2°10'26.5"E	03/02/2020; 12:00:10

Table 2

Identifier of people	Name	Age (years)	Contact number	Infection Tested
X11992291	John	22	04111111	positive
N72783912	Sally	27	04222222	negative
L11892277	James	48	04222333	negative
Z98192922	Jane	70	04555888	positive

Question8: (4 Marks)

Consider a database containing information about movies: genre, director, and decade of release. We also have information about which users have seen each movie. The rating for a user on a movie is either 0 or 1. Here is a summary of the database:

Movie	Release decade	Genre	Director	Total numbers of rating
A	1970s	Comedy	D_1	40
B	2010s	Comedy	D_1	500
C	2000s	Action	D_2	300
D	1990s	Action	D_2	25
E	2010s	Comedy	D_3	1

Consider user U_1 is interested in the time period 2000s, the director D_2 and the genre Comedy. We have some existing recommender system R that recommended the movie B to user U_1.

The recommender system R could be one or more of the following options:

- User-based collaborative filtering
- Item-based collaborative filtering
- Content-based recommender system

A) Given the above dataset, which one(s) do you think R could be? (If more than one option is possible, you need to state them all.) Explain your answer.

B) If some user U_2 wants to watch a movie, under what conditions can our recommender system R recommend U_2 a movie? If R recommends a movie, how does it do it? If R cannot recommend a movie, give reasons as to why it can't. State any additional information R might want from U_2 for predicting a movie for this user, if required.