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

Final Examination– Term1, 2021

30th April, 2021

COMP9321 Data Service Engineering

Total Exam Mark: 40

Total Number of Questions: 20 + 10

Exam Duration: 2 Hours +15 minutes (reading and submitting)

****** 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 10 Marks)

Use Moodle Quiz to Answer all the 20 Questions. The last submission is going to be marked. Make sure you click submit at the end so the submission will be considered.

https://moodle.telt.unsw.edu.au/mod/quiz/view.php?id=3789585

PartB: Written Answer Questions (Total 30 Marks)

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

Question1 (2 marks):

In your own words, explain the principle of HATEOAS (Hypermedia As The Engine Of Application State) enrich your explanation with an example.

Question2 (2 marks)

In your own words, explain when we should consider Deep Learning versus when should we consider a more traditional machine learning algorithm (e.g., kNN, decision trees, linear regression…etc).

Question3 (1 mark)

Mention 3 security consideration you may prioritize when securing your RESTful API. In your own words, explain one of them with an example.

Question4 (2 marks)

For the Data snippet below, list the data quality problem(s). Mention the record(s) and the field(s).

Student ID	Name	Program	Cohort	Gender	Age	Date of Birth	City
1234	John Smith	3344	UGRD	M	21	23-04-1999	Sydney
1122	Jane Doe	3322	PGRD	F	20	01-18-2000	Syddey
1234	Smith, John	3344	UGRD	M	21	23-04-1999	Brisbane
3234	Jack Sparrow	2211	postgrad	M	21	17-01-2999	Australia
5223	Ada Wong	N/A	None	F	-19	7-11-2001	Sydney

Question5: (2 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	-

Question6: (3 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

Question7: (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? Explain in your own words when it would be more suitable to prioritize precession over recall and vice versa.

Question8: (2 Marks)

You have a coffee service where you handle coffee orders (drink type, size, number of shots):

A. Consider the following HTTP request invoking a POST method of the Coffee RESTful API. Write down the content of the HTTP response that you would return as the result.

```
POST /orders

HTTP/1.1

Host: api.coffeehouse.com

Content-Type: application/json

{

“drink”  :  ”latte”,

“size”    :  ”small”,

“shots” :   2

}
```

B. In the light of part A of the question, consider the following HTTP request invoking a GET method of the Coffee RESTful API. Write down the content of the HTTP response that you would return as the result. Do not be concerned about the specific content as long as it is relevant.

```
GET /orders?order_by=+id&filter=id,drink

HTTP/1.1

Host: api.coffeehouse.com

Content-Type: application/json
```

Question9: (5 marks)

You have taken the role of cyber security analyst in an organization. You are considering utilizing your data service engineering knowledge to help the team. You have access to a vulnerability database from a security service provider including information about software vulnerabilities discovered, the type of the vulnerability, the software that this vulnerability has affected, the version affected with the vulnerability, the severity score, textual description of the vulnerability, and the date this vulnerability was discovered (snippet provided in Table1). You also have data about the software products used in the organization containing the software used in the organization, the version of the software, whether the software is used in production environment or not, and the date of the last maintenance to the systems with the software (table2). As this is a big organization so they prioritize the vulnerabilities that they need to fix according to how severe the vulnerability is. You can consider the following categorization:

Critical = vulnerabilities with Vulnerability Severity Score higher or equal to 8.5

High = vulnerabilities with Vulnerability Severity Score higher than 5.6 and lower than 8.5

Medium = vulnerabilities with Vulnerability Severity Score higher than 3.6 and lower or equal to 5.5

Low= vulnerabilities with Vulnerability Severity Score lower or equal to 3.5

The security service provider usually provides quick update about new discovered vulnerabilities but due to the need to test and measure severity they delay providing the "**Vulnerability Severity Score"** by more than 10 days. During these 10 days the organization do not know how dangerous the vulnerabilities are, so they just randomly start fixing them. You acknowledge that this is not the best security practice, and you want to help solve this problem.

1. Provide a brief description of how you are going to approach the problem
2. Explain what are the specific pre-processing steps that you need to perform to increase the value of the data provided.
3. Explain what is the machine learning algorithm you are going to consider? Explain why you chose this Machine learning algorithm and why do you see it fit to solve the problem? Explain if there are any modification to the data you need to perform to be able to fit the model. Be advised, since this is your small project, so you are limited in terms of computation power and resources.

Table1

Vul ID	Vulnerability Type	Affected Software	Affected Software Version	Vulnerability description	Vulnerability Severity Score	Date Discovered
1	RCE	Microsoft Excel	6.9 to 9.1	Textual description of vulnerability	10	2021-3-10
2	Buffer overflow	OpenSSH	2.3.1 to 3.3	Textual description of vulnerability	7.5	2020-1-9
3	Memory Leak	Linux Kernel	before 5.0.3	Textual description of vulnerability	7.8	2021-2-3
4	XSS	Apache Tomcat	7.0.1 to 8.0.91	Textual description of vulnerability	4.5	2020-4-4
5	Memory corruption	Windows 10	1607 to 1903	Textual description of vulnerability	9.3	2021-3-28

Table2

Software	Version	Live (production)	Last Maintenance
Microsoft Excel	7.1	Yes	2021-24-04
Ubuntu Linux	6.0	No	2021-18-03
OpenSSH	3.2	Yes	2021-10-04
Windows 10	1708	Yes	2021-16-03
Apache Tomcat	7.0	No	2021-13-04

Question10: (4 Marks)

Consider a database containing information about movies: genre, director, and decade of release. We also have information about which users have watched 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	Drama	D_1	40
B	2010s	Drama	D_1	500
C	2000s	Action	D_2	300
D	1990s	Action	D_2	25
E	2010s	Drama	D_3	1

Consider user U_1 is interested in the time period 2000s, the director D_2 and the genre Drama. 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.