TECHNOLOGICAL UNIVERSITY DUBLIN

TU059/1 MSc. in Computing
TU060/1 MSc. in Computing
TU060/2 MSc. in Computing

SEMESTER 1 EXAMINATIONS 2020/2021

OPEN BOOK EXAMINATION

DATA MINING [DATA9910]

MR. BRENDAN TIERNEY
DR. DEIRDRE LILLIS
DR. IGNACIO CASTIÑEIRAS

DATE & TIME TBA.

INSTRUCTIONS TO CANDIDATES e.g.

Answer **Any Three** Questions.

Question One has 34 marks, all other questions have 33 marks.

Illustrate your answers with appropriate examples and diagrams.
Submit answer document in PDF format on BrightSpace. All answers, diagrams, examples, etc to be included in one PDF document

1. **(a)** You are about to embark on a data science project. The project lead or consultant says, "*Give me all your data, I'll mine it and tell you what is happening*". Discuss this statement and explain a better approach for your project.

[10 Marks]

(b) Explain the role professional and personal ethics plays in your role as a data scientist? Illustrate your answer with three different scenarios when personal ethics comes into play for data science projects.

[8 Marks]

(c) Explain what is an Ensemble Model. Give an example, based on a business scenario, of how and when you would build an Ensemble Model and what you would like to achieve by using the Ensemble Model.

[8 marks]

(d) ONNX and PMML are examples of model interchange languages. Discuss the role these play with the deployment of machine learning models and some of the potential issues with using them.

[8 Marks]

2. **(a)** Discuss the role of Descriptive Analytics in your Data Science projects, and its importance as a precursor to the Data Preparation stage of the CRISP-DM lifecycle.

[7 Marks]

(b) Explain the importance of having a representative sample of your data in the training data set to avoid the *overfitting* of the model.

[7 Marks]

(c) When investigating data you need to look at data at many different levels. For each of the following, explain what information you would be looking at, what are the typical issues and how you would address the issues.

- Attribute data
- Related data across attributes
- Record level data

Illustrate your answers with example data issues that exist in the following table.

Name	Age	Time as a customer	Income	Marital Status	Sex	Num Children	Yr_Started_Work	Num_Yr_Experience
Sean Penn	56	24	1,200,00		M	2	1974	30
Delan Kelly	46	28	65000	S	M	0	1970	28
Minnie Mouse	89	70	25,000,000	Unknown	F	Unknown	1928	Lots
George Smith	25	2	2		F	0	2009	8
R. Moore	89	30	750000	M	M		1945	72
Sean Reily	35	5	50,000	Married	M	1	2000	17
Mrk Leddy	42	15	93,000	Married	Male	3	1999	35

[12 marks]

(d) Explain the type of data transformations needed to convert a regression problem into a multi-class classification problem.

[7marks]

3. **(a)** Explain how the K-means clustering method works.

[8 marks]

(b) The following is a set of one-dimensional points : {1,1,2,3,5,8,13,21,33,54}. Perform two iterations of k-means on these points using the two initial centroids 0 and 11.

[9 marks]

(c) Discuss, using a suitable example, how Clustering can be used in data preparation to calculate a suitable value for missing data. Discuss how this approach is more appropriate than taking the mean or mode for an attribute when the data set is segmented

[8 marks]

(d) Explain the possible benefits of using Clustering as a precursor to other data mining tasks to support better predictive accuracy.

[8 marks]

4. Consider the data set shown in the following table containing market basket transactions.

Customer ID	Transaction ID	Items Bought
1	1001	{i1, i4, i5}
1	1024	{i1, i2, i5}
2	1012	{i1, i2, i4, i5}
2	1031	{i1, i3, i4, i5}
3	1015	{i2, i5}
3	1022	{i2, i4, i5}
4	1029	{i3, i4}
4	1040	{i1, i2}
5	1033	{i1, i4, i5}
5	1038	{i1, i2, i5}

(a) Using the data in the above table, illustrate how the Apriori algorithm would process this data to find frequent item sets. Include descriptions of each stage/step.
Calculate the Support for Itemset {i2, i4, i5}.

[15 marks]

(b) Explain the following terms used to measure the association rules:

- support
- confidence
- lift

[9 marks]

(c) The Corse of Dimensionality can be a problem for Market Basket Analysis. Explain what this means, the impact it can have and how product hierarchies and aggregations can be used to address this.

[9 marks]