W228A/101C, W228B/101C

DUBLIN INSTITUTE OF TECHNOLOGY

DT228A/1 MSc. in Computing
DT228B/1 MSc. in Computing
DT228B/2 MSc. in Computing

WINTER EXAMINATIONS 2017/2018

DATA MINING [DATA9910]

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

DATE & TIME TBA.

TWO HOURS

Answer **TWO** Questions.

All questions carry equal marks.

Illustrate your answers with appropriate examples and diagrams.

1. **(a)** Explain the following terms, clearly distinguishing the differences between them, in relation to data analytics, data mining and machine learning:

- Descriptive Analytics
- Predictive Analytics
- Prescriptive Analytics

[15 marks]

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

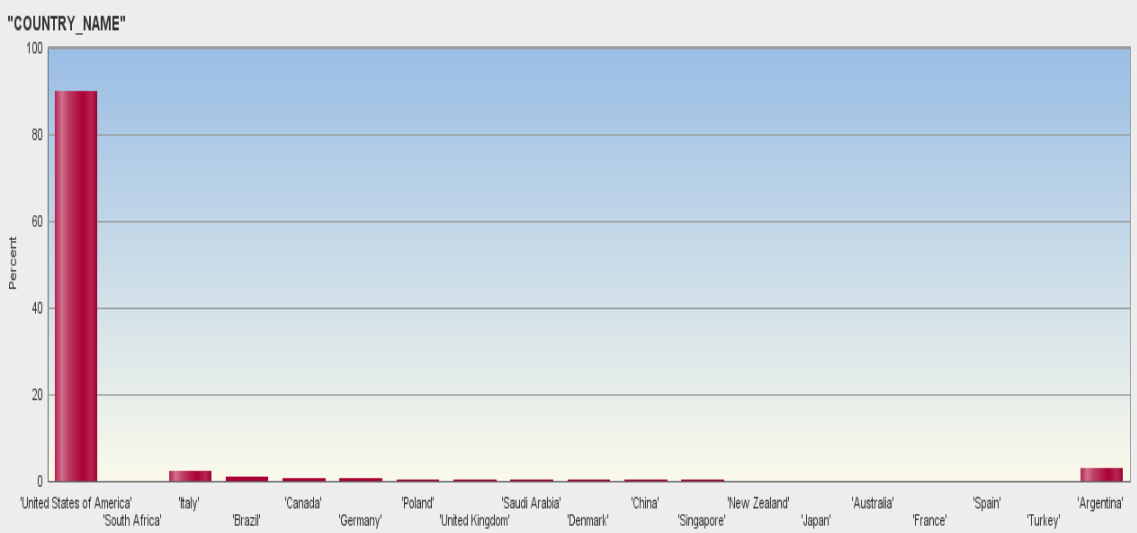
- Attribute data
- Related data across attributes
- Record level data

Illustrate your answers with example data issues that exists 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

[15 marks]

(c) The following graph and table gives the distribution of customers by country for a company. Identify the problem that exists for creating a customer churn predictive model for this data and what recommendations you would give.



Country Name	Num Customers
United States of America	134400
Italy	3700
Brazil	1400
Canada	900
Germany	800
Poland	700
United Kingdom	600
Denmark	500
Saudi Arabia	500
China	400
Singapore	400
New Zealand	300
Japan	200
Australia	200
Turkey	100
France	100
Spain	100
South Africa	100

[10 marks]

(d) Discuss the impact and issues that the EU GDPR will have on the use of advanced analytics and machine learning on the profiling of customer data.

[10 marks]

2. **(a)** Explain the following quote from George Box regarding the creation of classification models.

"*Essentially, all models are wrong, but some are more useful*"

A model is a simplification or an approximation of reality and hence will not reflect all of reality.

[10 marks]

(b) Explain how the Random Forests algorithm works for both the building of the model and the scoring of the data.

[10 marks]

(c) Discuss how a *Confusion Matrix* can be used to evaluate a classification model.

[10 marks]

(d) Explain the following terms and their importance for evaluating classification models. For the following table calculate the value for the following terms:

- Precision
- Recall / Sensitivity
- Accuracy

Predicted	Class = N	Class = Y
Actual		
Class = N	221125	6305
Class = Y	28881	13094

What conclusions can be drawn from the results calculated for Precision, Recall/Sensitivity and Accuracy.

[20 marks]

3. **(a)** Data can be prepared in two different ways for Association Rule analysis. These include:

- Single record
- Parent-Child relationship

Explain both of these methods and how these approaches address the issue of sparsity in the data.

[10 marks]

(b) Explain how the Apriori pruning principle addresses the issue of in-frequent item sets, to reduce the search space for frequent item-sets.

[10 marks]

(c) For the following data illustrate how the Apriori algorithm would process the following data to find the frequent item sets. Explain what happens at each step.

ID	Basket Contents
T101	Beer, Crisps, Milk
T102	Crisps, Bread
T103	Crisps, Nappies
T104	Beer, Crisps, Bread
T105	Beer, Nappies
T106	Crisps, Nappies
T107	Beer, Nappies
T108	Beer, Crisps, Nappies, Milk
T109	Beer, Crisps, Nappies

[18 marks]

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

- support
- confidence
- lift

[12 marks]