Programme Code: TU856, TU857, TU858
Module Code: CMPU 4010
CRN: 25771, 22416, 31082

TECHNOLOGICAL UNIVERSITY DUBLIN

CITY CAMPUS

TU856 BSc. (Honours) Degree in Computer Science

TU857 BSc. (Honours) Degree in Computer Science (Infrastructure)

TU858 BSc. (Honours) Degree in Computer Science (International)

Year 4

SEMESTER 1 EXAMINATIONS 2021/22

Machine Learning for Data Analytics

Internal Examiner: Dr. Svetlana Hensman

Dr. Paul Doyle

External Examiner: Sanita Tifentale (TU856)

Pauline Martin(TU857)

Pamela O'Brien (TU858)

Duration: 2 hours

ANSWER **ALL** QUESTIONS.
QUESTION 1 IS 40 MARKS.
QUESTION 2 IS 30 MARKS.
QUESTION 3 IS 30 MARKS.

1. (a) Over-fitting the training data is one of the major issues to be aware of when developing a machine learning model. Explain what is over-fitting and discuss the steps we can take to prevent it.

(10 marks)

(b) Outline the key outcomes of the data exploration process.

(10 marks)

(c) Discuss what is meant by the curse of dimensionality in the context of data analytics, and suggest how we can deal with it.

(10 marks)

(d) The table below shows 14 predictions made by a spam classification model for a categorical target feature (the prediction is SPAM=true or SPAM=false).

ID	TARGET	SPAM
1	false	false
2	false	false
3	true	false
4	true	false
5	true	true
6	true	true
7	true	false

ID	TARGET	SPAM
8	false	false
9	false	true
10	false	true
11	true	false
12	true	true
13	false	false
14	true	false

(i) Create the confusion matrix for the results shown in the table above.

(5 marks)

(ii) Calculate the **precision**, **recall** and **F1 measure**.

(5 marks)

2. (a) The table below describes a set of 6 patients in terms of their **WEIGHT** (kgs) and **HEIGHT** (meters), and whether or not they have **DIABETES**.

ID	WEIGHT (KG)	HEIGHT (M)	DIABETES
1	68	1.7	yes
2	55	1.6	no
3	70	1.6	yes
4	100	1.9	no
5	50	1.5	no
6	92	1.8	no

A doctor has a new patient with **WEIGHT** 77kg and **HEIGHT** 1.7m.

i) If the doctor inputs the data for the new patient in a 3-Nearest Neighbours classifier built based on the data above to check if the patient is at risk of diabetes, will the model predict **yes** or **no** that the patient?
The model uses *Euclidean distance* as a similarity measure.

(10 marks)

(Q2 continues on next page)

ii) Some medical professionals use BMI as a combined metrics for a patient's **WEIGHT** and **HEIGHT**. BMI is calculated as

$$BMI = \frac{weight\ (kg)}{height\ (m) * height\ (m)}$$

Assuming the data for the patients in table above has been updated, and **WEIGHT** and **HEIGHT** features replaced with **BMI** feature instead, what prediction would a 3-Nearest Neighbours classifier return for the new patient based on the updated data.

(10 marks)

(b) The table below contains a dataset of 14 training examples about decision making factors for a game of tennis to go ahead.

DAY	OUTLOOK	WIND	DECISION
1	Sunny	Weak	No
2	Sunny	Strong	No
3	Overcast	Weak	Yes
4	Rain	Weak	Yes
5	Rain	Weak	Yes
6	Rain	Strong	No
7	Overcast	Strong	Yes
8	Sunny	Weak	No
9	Sunny	Weak	Yes
10	Rain	Weak	Yes
11	Sunny	Strong	Yes
12	Overcast	Strong	Yes
13	Overcast	Weak	Yes
14	Rain	Strong	No

You are asked to construct a decision tree for this data using ID3 algorithm and entropy-based information gain. Which feature will be tested first in the root of the tree?

There is no need to construct the full tree, just show how the first feature to be tested on will be selected and the resulting partial tree.

On the next page are some formulas that may be useful.

(Q2 continues on next page)

$$H(t, \mathcal{D}) = -\sum_{l \in levels(t)} (P(t = l) \times log_2(P(t = l)))$$

$$rem(d, \mathcal{D}) = \sum_{l \in levels(d)} \underbrace{\frac{|\mathcal{D}_{d=l}|}{|\mathcal{D}|}}_{\text{weighting}} \times \underbrace{H(t, \mathcal{D}_{d=l})}_{\substack{\text{entropy of} \\ \text{partition } \mathcal{D}_{d=l}}}$$

$$IG(d, \mathcal{D}) = H(t, \mathcal{D}) - rem(d, \mathcal{D})$$

(10 marks)

3. (a) The table below lists a dataset containing details about an uptake of a recent promotion (Yes or No) by 5 customers of an airline. The descriptive features included in the table are gender (GENDER), whether the customer is a frequent flyer (FREQUENT FLYER) and whether the customer has purchased a flight in the last 12 months (FLIGHT 12 MONTHS).

ID	GENDER	FREQUENT FLYER	FLIGHT 12 MONTHS	PROMOTION
1	female	true	true	Yes
2	male	false	false	Yes
3	male	false	true	No
4	female	true	true	Yes
5	female	false	false	No

(i) Calculate the probabilities required by a **Naïve-Bayes model** built based on this dataset.

(10 marks)

(ii) What would be the outcome predicted by the model for new instance

GENDER = female, FREQUENT FLYER = true, FLIGHT 12 MONTHS = true

(5 marks)

(iii) Do you notice any issues with the calculated probabilities in part (i)? Discuss how you could address them.

(5 marks)

(b) Discuss the possible limitation of a **linear regression model**, and explain how **basis functions** could be used to address it.

(10 marks)