Programme Code: TU860
Shared with: TU883
Module Code: COMP H4030
CRN: 25058

TECHNOLOGICAL UNIVERSITY DUBLIN

BLANCHARDSTOWN CAMPUS

TU860 (shared with TU883) – Computing (Information Technology)

Year 4

SEMESTER 1
EXAMINATIONS 2024/25

Data Analytics

Internal Examiner: Dr Aurelia Power

External Examiner: Dr Owen Foley

Exam Duration: 2 hours

Instructions:

1) To ensure that you take the correct examination, please check that the module and programme which you are following matches the one(s) above.
2) The paper consists of five questions. Candidates should complete ANY FOUR of the five questions.
3) The paper is worth 100 marks. Each question is worth 25 marks.
4) You can use a calculator to work out the solutions of mathematical calculations.

Question 1:

a) Given a dataset of students (each row represents a student) from a technological university, identify whether the following five student attributes (columns) are **nominal**, **categorical**, **ordinal**, **interval,** or **ratio**.

- Energy level on scale of {low, moderate, high}.
- Country.
- Average hours studying daily.
- Sleep duration before on the night before exam.
- Numeric grade (out of 100).

(5 marks)

b) Figure 1 shows a scatterplot of two student attributes: **the average hours spent studying daily** on the x axis and **the sleep duration on the night before exam** on the y axis; the scatter plot is coloured by class label column: *performance category*, which has three possible values*: Excellent, Average*, and *Poor*. Describe how they relate to one another and to the three classes. Do you think combining these two attributes are useful in separating the three classes? Explain your answer.

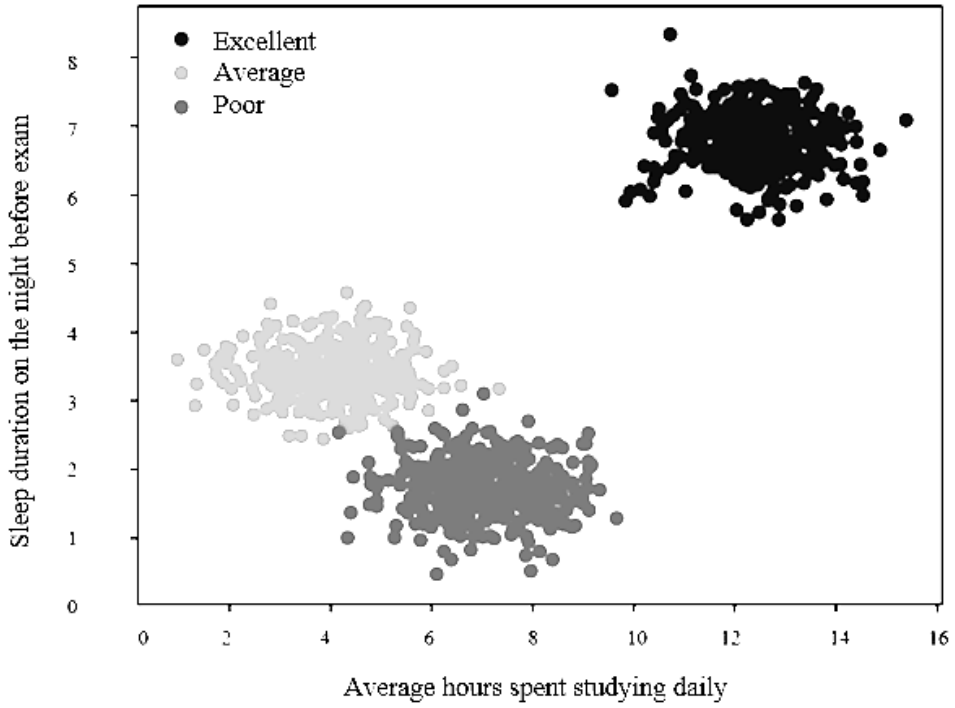

Figure 1: Average number of hours spent studying daily and the number of hours slept on the night before exam, coloured by performance category.

(8 marks)

c) A technological university has students from four different countries. Discuss what the grouped bar chart in Figure 2 indicates with respect to proportions of **male students** and **female students** from those four different countries.

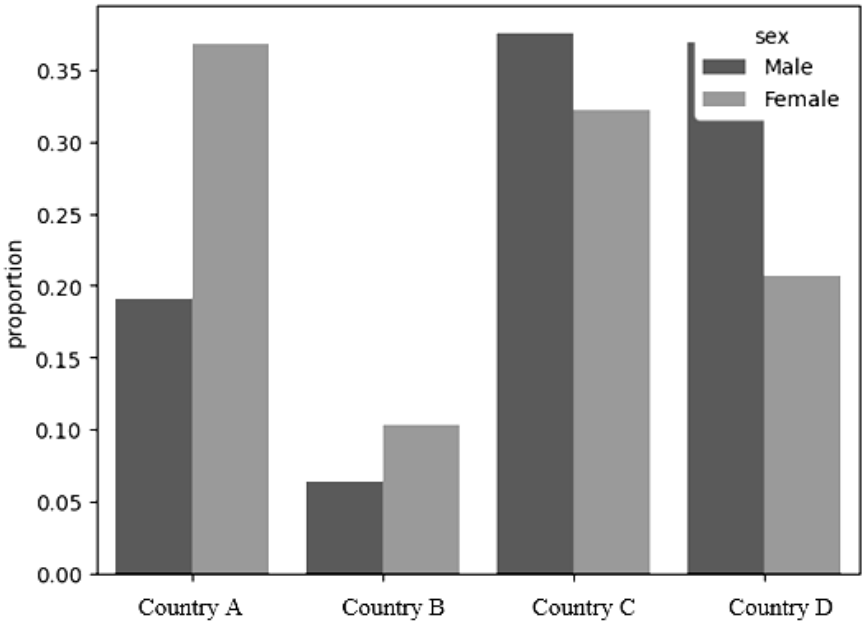

Figure 2: University proportions of male students and female students from 4 different countries.

(6 marks)

d) Describe in detail **one of the phases** of the CRISP-DM methodology.

(6 marks)

Total: 25 marks

Question 2:

Table 1 shows the confusion matrix generated after applying a classification model to a test dataset with 350 instances (students). The class label is *performance category*, which has three possible values*: Excellent, Average*, and *Poor*.

Table 1. Confusion Matrix.

		Predicted class		
		Average	**Excellent**	**Poor**
Actual class	**Average**	208	7	2
	Excellent	10	75	0
	Poor	5	0	43

a) Use the data from the confusion matrix in Table 1 to answer the following questions, showing clearly all your calculations (Note: you only need to report your results with a precision of 2 decimal places):

i. What is the overall % model **accuracy**?

ii. What is the % **recall** for each class?

iii. What is the % **precision** for each class?

(7 marks)

b) Briefly describe each of the following sampling methods: **random sampling**, **stratified sampling** and **Kennard-Stone sampling**. How does each of these methods affect the **class distribution**? Provide examples to support your answer.

(12 marks)

c) Describe the **two main reasons for excluding columns (attributes)** during the preparation phase.

(6 marks)

Total: 25 marks

Question 3:

Table 2 shows the summary statistics of 2 student numeric attributes and the class label – *performance category* – which is nominal. The dataset comes from a technological university and has **3755** instances where each instance represents a student.

Table 2. Summary Statistics.

Attribute	dtype	Missing	Min	Max	Mean	Standard deviation	25th percentile	50th percentile	75h percentile
average hours spent studying daily	float	730	0	16	5.7	7.09	3.4	4.1	6.4
sleep duration on the night before exam	float	25	0	8	6.1	2.01	5.2	6.8	7.3

Class Label Attribute	Type	Missing	Unique	Least frequent	Most frequent	Other Value(s)
performance category	object	0	3	Poor (434)	Average (2516)	Excellent (805)

a) Interpret the **summary statistics** in Table 2 for the two numeric attributes.

(8 marks)

b) Two of the attributes in Table 2 have **missing values**. Suggest appropriate techniques to deal with them. Justify your choices.

(7 marks)

c) The **class label attribute** in Table 2 has three unique values. Explain what issue you can identify and what can potentially lead to. Suggest one way you can address it.

(6 marks)

d) Demonstrate how to apply **min-max normalisation** to the following list of 5 values: 7, 3, 4, 7, 8. Show all steps.

(4 marks)

Total: 25 marks

Question 4:

Table 3 is a sample from a dataset capturing attributes used to predict whether a student will complete an honours degree. The binary class label, *completes_hounours*, has two unique values: *Yes* and *No*. The attribute *Average Hours Spent Studying Daily* has been scaled to the range [0, 1], the *Gender* attribute has two unique values: *Female* and *Male*, and *Energy Level* attribute has three unique values: *Low, Moderate,* and *High*. The table also includes a row of unlabelled data.

Table 3: Sample Data.

Training Data			
Normalised Average Hours Spent Studying Daily	**Gender/Sex**	**Energy Level**	**completes_honours**
0.46	Female	Moderate	No
0.23	Female	High	No
0.77	Female	Moderate	No
0.55	Female	Low	Yes
0.57	Male	Moderate	Yes
0.61	Male	High	Yes
Unlabelled Data			
Normalised Average Hours Spent Studying Daily	**Gender/Sex**	**Energy Level**	**completes_honours**
0.52	Male	High	?

a) Using table 3, demonstrate how **a k-nearest neighbour algorithm** uses the 6 rows of data to classify the row of unlabelled data. Show all calculations.

i) If $k=2$, what is the predicted class?

ii) If $k=3$, what is the predicted class?

(9 marks)

b) Using the training data from table 3, demonstrate how **a decision tree algorithm** will compute the information gain for the *Gender/Sex* and *Energy Level* attributes. Draw diagrams to show how the training data is split in each branch. Include all calculations.

NOTE You do not need to calculate logarithms, only select the correct one from the following: $\log_2(3/4) = 0.42$; $\log_2(1/4) = -2$; $\log_2(2/3) = -0.58$; $\log_2(1/3) = -1.56$.

(16 marks)

Total: 25 marks

Question 5:

Table 4: the results of a **linear regression model** for predicting *Numeric Grade* based on 2 attributes: *Average Hours Spent Studying Daily* and *Sleep Duration on the Night Before Exam*.

Attribute	Coefficient	Standard Error	t-stat	p-Value
Average Hours Spent Studying Daily	4.3	0.7	15.4	0.02
Sleep Duration on the Night Before Exam	7.3	0.5	16.3	0.0

a) Table 4 above shows the results of a linear regression model. For both attributes, *Average Hours Spent Studying Daily* and *Sleep Duration on the Night Before Exam,* interpret each score in relation to the dependent variable *Numeric Grade* and, based on these, determine whether the attributes should be included in the model.

(12 marks)

b) Define the concept of **coefficient of determination (R^2)**; how does a score of 0.86 relate to the model in Table 4? (Note: You do not need to include formulas in your answer)**.**

(4 marks)

c) With respect to **association analysis**, suppose a website that sells educational software for level 3 students has gathered 100,000 transactions. Among these transactions, 40,000 contain statistics software and 15,000 contain data modelling software. Of the 40,000 that contain statistics software, 10,000 contain also data modelling software.

i) Derive the association rule that apply to customers who buy statistics software and data modelling software as part of the same transaction; ensure to identify which is the antecedent and which is the consequent.

ii) Compute the confidence in this rule.

iii) Compute the support for this rule.

iv) Is the rule generated in i) interesting according to the lift framework? (Note: you need to first compute the lift for this rule and then interpret it).

(9 marks)

Total: 25 marks