INSTITUTE OF TECHNOLOGY BLANCHARDSTOWN

Academic term	2013-14
Year of study	Year 3
Semester	SEMESTER ONE
Date of examination	
Time of examination	

Programme code	**Programme title**	**Module code**
BN302	Bachelor of Science in Computing in Information Technology	COMP H3027
BN013	Bachelor of Science in Computing in Information Technology	COMP H3027
BN104	Bachelor of Science (Honours) in Computing	COMP H3027

Module title	**Data Mining**

Internal Examiner(s)	**Geraldine Gray**
External Examiner(s)	Dr. Tom Lunney Mr. Michael Barrett

Instructions to candidates:

1.	**To ensure that you take the correct examination, please check that the module and programme which you are following is listed in the table above.**
2.	**Questions 1 in Section A is COMPULSORY. Candidates should attempt Questions 1, and any two of three questions in Section B.**
3.	**There are 100 marks on the paper. Question 1 is worth 40 marks. All other questions are worth 30 marks each.**
4.	**Show all your work**

DO NOT TURN OVER THIS PAGE UNTIL YOU ARE TOLD TO DO SO

SECTION A

Question 1: (Compulsory)

a) Explain the importance of both the **Data Understanding** and **Data Preparation** phases of the CRISP-DM methodology.

(4 marks)

b) Discuss the information content of each of the following data types in the context of training a classification model: **nominal**, **categorical**, **interval** and **ordinal**.

(4 marks)

c) What is the role of both the training dataset and the test dataset when training a classification model?

(4 marks)

d) Explain the difference between class **precision** and class **recall**.

(4 marks)

e) Discuss the impact of having ***k*** set too high, or too low, when training a k-Nearest Neighbour classifier.

(4 marks)

f) Explain the terms **input layer**, **output layer** and **hidden layer** in the context of an Artificial Neural Network.

(4 marks)

g) Describe how **binning** is used as a data preparation technique. In what circumstances you would consider using binning?

(4 marks)

h) When would you consider normalizing (scaling) an attribute? Explain how **Z-transform** scales an attribute's values.

(4 marks)

i) What is the aim of a clustering algorithm when analyzing a dataset?

(4 marks)

j) Explain the difference between subjective and objective cluster evaluation.

(4 marks)

Total: 40 marks

SECTION B

Question 2:

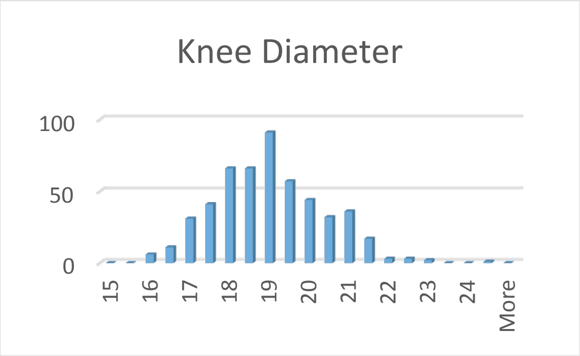
The table below shows the meta data for a dataset of skeletal measures, used to determine **gender**. The dataset has 8 attributes, and 2000 rows:

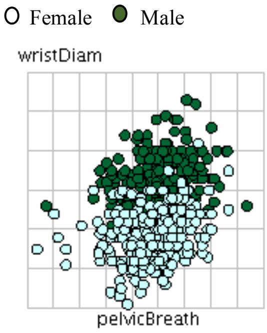
Role	Name	Type	Statistic	Range	Missing values
Label	Gender	Binominal	Mode=0(1012)	0 (1102) 1(988)	0
Regular	Age	Real	30±9.6	[18,67]	1300
Regular	Pelvic Breath	Real	27.83±2.2	[18.7,34.7]	2
Regular	Chest Depth	Real	19.2±2.5	[14.3,27.5	3
Regular	Chest Diameter	Real	27.9±22.7	[22.2,35.6]	6
Regular	Elbow Diameter	Real	13.38±1.3	[9.9,16.7]	200
Regular	Wrist Diameter	Real	10.54±0.9	[8.1,13.3]	0
Regular	Knee Diameter	Real	18.8±1.3	[15.7,24.3	0
Regular	Height	Real	171±9.3	[147.2,198.1]	0

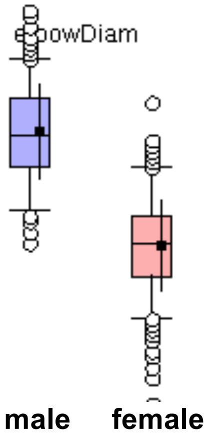
a) For each of the five attributes with missing data, recommend a suitable approach for handling their missing values. Justify each of your recommendations. **(9 marks)**

b) Recommend two other preprocessing techniques to use on the dataset above. Give a detailed explanation of each technique, and justify why they are an appropriate choice for this dataset. **(10 marks)**

c) Interpret each of the three plots below. The histogram is for **Knee Diameter**. The scatter plot illustrates **Wrist Diameter** by **Pelvic Breath** and is colour coded by **Gender**. The box plots are for **Elbow Diameter**, split by **Gender**. **(11 marks)**







Total: 30 marks

Question 3:

The dataset below is an extract from a car dataset, containing attributes that describe each car. There is a binary class label, to **Buy**, yes or no.

Training data:

	Safety	Value For Money	Boot Size	Buy
1	High	Good	Large	Yes
2	Medium	Good	Meduim	Yes
3	Low	Average	Medium	Yes
4	Low	Poor	Large	No
5	High	Poor	Small	No
6	Medium	Good	Medium	No

Test row:

Safety	Value For Money	Boot Size	Buy
Low	Poor	Small	?

a) Given the six rows of training data above, explain how ***k***-Nearest Neighbour would classify the row of test data shown above if ***k*** is set to 3. Include all calculations in your answer.

If the actual class label for this row is '**no'**, does ***k***-Nearest Neighbour classify it correctly at k=3?

(12 marks)

b) Interpret the following confusion matrix from a ***k***-Nearest Neighbour classifier, trained on 40 rows of the car dataset:

	Predicted Yes	**Predicted No**
Actual Yes	20	5
Actual No	10	5

i. What is the overall **accuracy** of the classifier? **(2 marks)**

ii. Calculate the **precision** for each class. Which class has the best precision? **(5 marks)**

iii. Calculate the **recall** for each class. Which class has the best recall? **(5 marks)**

c) Compare k-Nearest Neighbour with one other classification algorithm you have studied in terms of: input and output data types supported; how easy the output is to understand; how well the algorithm can handle missing data; and training time. **(6 marks)**

Total: 30 marks

Question 4:

a) Calculate the **Euclidean** distance between each of the three rows of data below. Note: attributes are already scaled to the range [0, 10]. Which two rows are the most similar?

	Age	**Level of education**	**Income**
Row 1:	5	2	3
Row 2:	2	2	7
Row 3:	6	8	7

(7 marks)

b) Explain the difference between a **partition based** clustering algorithm and a **density based** clustering algorithm.

(6 marks)

c) One of the disadvantages of **k-means** clustering is that the number of clusters must be specified in advance. Assuming a dataset has 3 clusters, explain in detail how k-means clustering identifies the three cluster.

(10 marks)

d) Explain how **DBScan** identifies **core points**, **border points** and **noise points** in a dataset. How are these labels used to define clusters in the dataset?

(7 marks)

Total: 30 marks