INSTITUTE OF TECHNOLOGY
BLANCHARDSTOWN

Year	Year 3
Semester	Semester 1
Date of Examination	[To be completed by the School Administrator]
Time of Examination	[To be completed by the School Administrator]

Prog Code	BN302	**Prog Title**	Bachelor of Science in Computing in Information Technology	**Module Code**	Comp H3018
Prog Code	BN013	**Prog Title**	Bachelor of Science in Computing in Information Technology	**Module Code**	Comp H3018
Prog Code	BN104	**Prog Title**	Bachelor of Science (Honours) in Computing in Information Technology	**Module Code**	Comp H3024
Prog Code	BN311	**Prog Title**	Bachelor of Science in Computing in Information Security and Digital Forensics	**Module Code**	ISDF H3018

Module Title	Data Mining

Internal Examiner(s): *Geraldine Gray, Laura Keyes*
External Examiner(s): *Mr Michael Barrett, Dr Tom Lunney*

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 tables above.**
2) **Question One Section A is COMPULSORY. Candidates should attempt Question One and ANY other two questions in Section B.**
3) **This paper is worth 100 marks. Question One is worth 40 marks and all other questions are worth 30 marks each.**

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

SECTION A: COMPULSORY QUESTION

Question 1: **(40 marks)**

Attempt **ALL** ten parts.

a) List and briefly explain each of the processes in the **CRISP_DM** methodology.
(4 marks)

b) Measurements can be classified as: *Ratio scale; Nominal scale; Interval scale; Ordinal scale and Categorical scale*. Give examples for each of these five types of measurement, and order them in terms of information content. Why is it important to know the category of the data before mining?
(4 marks)

c) Describe how a boxplot can give information about whether the value of an attribute is normally distributed or skewed.
(4 marks)

d) Explain, with the aid of an example, one technique for reducing the impact of noise on a dataset.
(4 marks)

e) What is the problem of *overfitting* in classification?
(4 marks)

f) Explain what is meant by *pre-pruning* and *post-pruning* a decision tree.
(4 marks)

g) Briefly describe two approaches for training and testing a model.
(4 marks)

h) What is *normalisation* and when would it be used in data mining.
(4 marks)

i) Discuss, with the aid of examples, the role of both *supervised* and *unsupervised* learning in data mining. Which data mining methods are associated with supervised learning? Which are associated with unsupervised learning?
(4 marks)

j) In clustering analysis, for what type of dataset would DBSCAN be a better choice than K-means?
(4 marks)

SECTION B: ANSWER ANY TWO QUESTIONS

Question 2: Data Preparation (30 marks)

ExampleSet (5000 examples, 1 special attribute, 21 regular attributes)

Role	Name	Type	Statistics	Range	Missing
regular	stalkSurfaceAR	binominal	mode = s (2391), least = f (213)	s (2391), f (213)	2396
regular	stalkColorBR	binominal	mode = w (1938), least = p (1194)	w (1938), p (1194)	1868
regular	ringType	binominal	mode = p (1962), least = e (1658)	p (1962), e (1658)	1380
regular	veilColor	binominal	mode = w (4800), least = n (96)	w (4800), n (96)	104
regular	ringNumber	binominal	mode = o (4589), least = t (375)	o (4589), t (375)	36
label	safe	binominal	mode = p (2816), least = e (2184)	e (2184), p (2816)	0
regular	capShape	polynominal	mode = x (2193), least = c (3)	x (2193), b (230), f (1	0
regular	capSurface	polynominal	mode = y (2135), least = g (2)	s (1355), y (2135), f (	0
regular	capColor	polynominal	mode = n (1407), least = u (16)	y (804), w (411), n (1	0
regular	bruises	binominal	mode = f (3396), least = t (1604)	t (1604), f (3396)	0
regular	odor	polynominal	mode = n (1980), least = m (36)	a (137), p (86), l (141	0
regular	gillAttachment	binominal	mode = f (4790), least = a (210)	f (4790), a (210)	0
regular	gillSpacing	binominal	mode = c (4430), least = w (570)	c (4430), w (570)	0
regular	gillsize	binominal	mode = b (3482), least = n (1518)	b (3482), n (1518)	0
regular	gillColor	polynominal	mode = b (1172), least = r (10)	k (141), n (499), g (57	0

Figure 3. Meta data for the Mushroom dataset

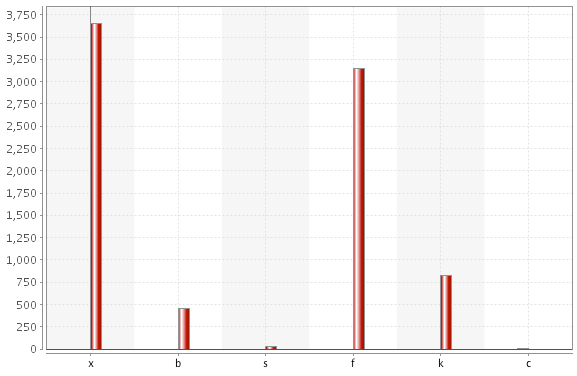
a) As is illustrated in Figure 3 above, five attributes listed in the meta data have missing values. For each attribute, explain what you would do to address the missing values. Justify all choices made. Where you recommend filling the missing values, explain two alternative techniques you could use.

(12 marks)

b) Explain how you would decide if sampling is appropriate for the mushroom dataset above. Also in your answer give details of two sampling techniques that could be used.

(11 marks)

c) Interpret the histogram below for the attribute '**capShape'**. Does it suggest any issues with this attribute?



(7 marks)

Question 3: Classification **(30 marks)**

a) Explain the difference between an eager and a lazy learner. Give scenarios when each would be a suitable classifier to use.

(6 marks)

b) You have been asked to mine a data set with the following characteristics:

There are 2000 rows of data. The fifteen attributes are numeric only, and the class label is binary. The objective is to both understand the patterns in the data, and also build a model that will accurately predict the class label. The patterns in the dataset are not expected to change much over time.

Which classification algorithms are suited to this task? Will you need to use more than one classifier to meet the objectives above? Explain you answer.

(6 marks)

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

If the actual class label for this row is '**edible'**, does ***k***-Nearest Neighbour classify it correctly at k=3?

(*14 marks)*

Training data:

Row Number	Cap Shape	GillSize	Ring Type	Type
1	Round	Narrow	Oval	poisonous
2	Round	Narrow	Ring	poisonous
3	Round	Broad	Ring	poisonous
4	Flat	Broad	Ring	poisonous
5	Flat	Narrow	Ring	edible
6	Flat	Broad	Oval	edible
7	Round	Broad	Oval	edible

Test data:

Round	Narrow	Oval	?

d) Discuss how to find the optimal value of ***k*** when using ***k***-Nearest Neighbour.

(4 marks)

Question 4: Clustering and Evaluation (30 marks)

a) Using the information provided in Figure 1 and Figure 2 below, calculate the *z-score* normalisation for the first <u>two</u> values of the attribute **sepalwidth**.

Role	Name	Type	Statistics	Range	Missings
label	iris	polynominal	mode = Iris-setosa (49)	Iris-setosa (49), Iris-ver:	0
regular	sepallenght	real	avg = 5.849 +/- 0.836	[4.300 ; 7.900]	0
regular	sepalwidth	real	avg = 3.056 +/- 0.432	[2.000 ; 4.400]	0
regular	petallength	real	avg = 3.755 +/- 1.768	[1.000 ; 6.900]	0
regular	petalwidth	real	avg = 1.195 +/- 0.761	[0.100 ; 2.500]	0

Figure 1 Meta data for the Iris dataset

ExampleSet (146 examples, 1 special attribute, 4 regular attributes)

Row No.	iris	sepallenght	sepalwidth	petallength	petalwidth
1	Iris-setosa	4.900	3	1.400	0.200
2	Iris-setosa	4.700	3.200	1.300	0.200
3	Iris-setosa	4.600	3.100	1.500	0.200
4	Iris-setosa	5	3.600	1.400	0.200
5	Iris-setosa	5.400	3.900	1.700	0.400
6	Iris-setosa	4.600	3.400	1.400	0.300

Figure 2 Excerpt of Iris dataset

(5 marks)

b) What is *Clustering analysis*? In your answer contrast the following approaches to clustering: hierarchical vs. partitional vs. density, exclusive vs. overlapping; and complete vs. partial.

(7 marks)

a) Explain the *K-means* clustering algorithm. Include in your answer the advantages and disadvantages of k-means clustering and describe <u>one</u> method that can be used for cluster evaluation.

(8 marks)

c) The following confusion matrices present the classification results that were produced using a Decision Tree and Neural Network in the prediction of Iris type for the Iris dataset. Which data mining method performs the best for the particular task? Discuss each evaluation measure you use to assess the performance of each model.

(10 marks)

accuracy: 93.81% +/- 5.68% (mikro: 93.84%)

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	49	0	0	100.00%
pred. Iris-versicolor	0	45	6	88.24%
pred. Iris-virginica	0	3	43	93.48%
class recall	100.00%	93.75%	87.76%	

Figure 5 Decision Tree Classification

accuracy: 97.24% +/- 4.51% (mikro: 97.26%)

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	49	0	0	100.00%
pred. Iris-versicolor	0	46	2	95.83%
pred. Iris-virginica	0	2	47	95.92%
class recall	100.00%	95.83%	95.92%	

Figure 6 Neural Network Classification