INSTITUTE OF TECHNOLOGY
BLANCHARDSTOWN

Year	Year 3
Semester	Semester 1 Repeat
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 <u>two</u> 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) Briefly explain the importance of the CRISP_DM methodology.

(4 marks)

b) Why perform Exploratory Data Analysis (EDA) and data preparation on a dataset? Why not proceed directly into the modeling phase?

(4 marks)

c) Describe two ways the distribution of values in a variable can be illustrated.

(4 marks)

d) What is the *curse of dimensionality*? Explain how this can be addressed.

(4 marks)

e) Explain what is meant by the term *outliers*. Illustrate your answer with an example.

(4 marks)

f) How can *regression* be used to reduce the level of noise in a dataset? Illustrate your answer with an example

(4 marks)

g) In data modeling, describe the difference between a training set and a test set.

(4 marks)

c) Explain the role of a *dependent variable (*or *class label)* when mining a data set.

(4 marks)

h) Briefly describe the following listing two algorithms used for each:

- Clustering
- Classification.

(4 marks)

j) Is mining a sample of data as effective as mining the entire dataset? Discuss your answer.

(4 marks)

SECTION B: ANSWER ANY TWO QUESTIONS

Question 2: Data Preparation (30 marks)

ExampleSet (1000 examples, 0 special attributes, 127 regular attributes)

Role	Name	Type	Statistics	Range	Missing
regular	PolicBudgPerPop	integer	avg = 4.221 +/- 4.088	[0.000 ; 10.000]	846
regular	community	integer	avg = 59.798 +/- 108.557	[1.000 ; 840.000]	529
regular	communityname	integer	avg = 44640.454 +/- 25075	[70.000 ; 94597.000]	529
regular	householdsize	real	avg = 2.714 +/- 0.351	[1.600 ; 5.280]	3
regular	racepctblack	real	avg = 9.429 +/- 14.546	[0.030 ; 96.670]	3
regular	racePctWhite	real	avg = 84.157 +/- 16.669	[2.680 ; 99.340]	3
regular	racePctAsian	real	avg = 2.432 +/- 3.784	[0.030 ; 34.330]	1
regular	racePctHisp	real	avg = 7.860 +/- 15.195	[0.140 ; 95.290]	1
regular	state	polynominal	mode = Lebanoncity (3), least	Glendalecity (3), Jacksoncity (3),	0
regular	county	polynominal	mode = CA (112), least = DE (	NJ (100), PA (56), OR (13), NY (	0
regular	fold	integer	avg = 5.282 +/- 2.894	[1.000 ; 10.000]	0
regular	population	integer	avg = 49242.836 +/- 16158	[10005.000 ; 3485398.000]	0
regular	agePct12t29	real	avg = 27.558 +/- 6.109	[9.380 ; 69.670]	0
regular	agePct16t24	real	avg = 13.966 +/- 5.889	[4.640 ; 61.340]	0
regular	agePct65up	real	avg = 12.060 +/- 5.059	[1.660 ; 52.770]	0
regular	numbUrban	integer	avg = 43521.087 +/- 16278	[0.000 ; 3485398.000]	0

Figure 1. Meta data for the Crime & Community dataset

Figure 1 above is an extract from the meta data generated from the crime&Community dataset, a US based dataset to investigate community related attributes and their relationship to Crime in that community. Answer the following questions based on this meta data:

Note: The dataset has 127 attributes in total.

a) Eight attributes listed in the meta data have missing values. Explain what you would do to address these missing values. Justify all choices made.

(7 marks)

b) The dataset above is to be used for cluster analysis. Apart from filling missing values, give details of TWO other preprocessing techniques you would recommend for the dataset. Explain the purpose of each technique, how it works, and justify why it is appropriate based on the metadata above.

(14 marks)

c) The histograms shown on the next page were generated as part of the Exploratory Data Analysis of the Crime&Community dataset. Discuss the two histograms with reference to:
   i. Variable distribution
   ii. Presence of outliers

(6 marks)

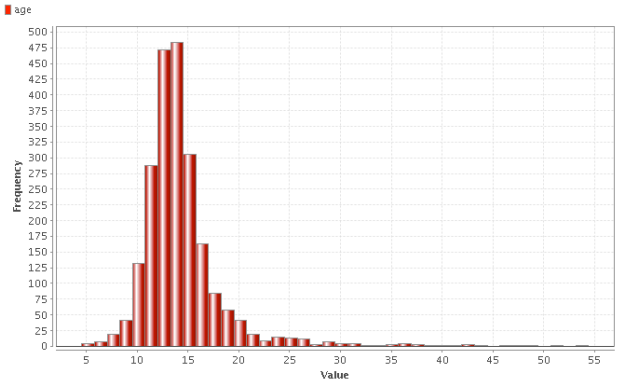
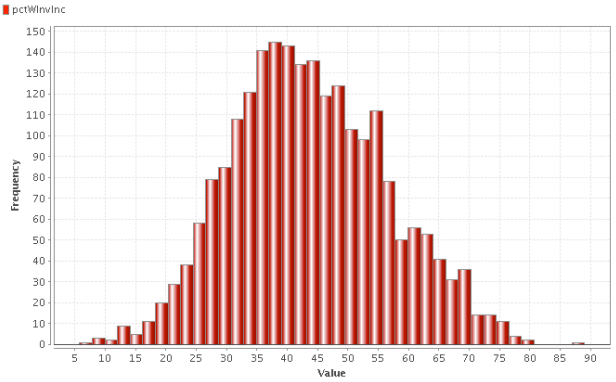


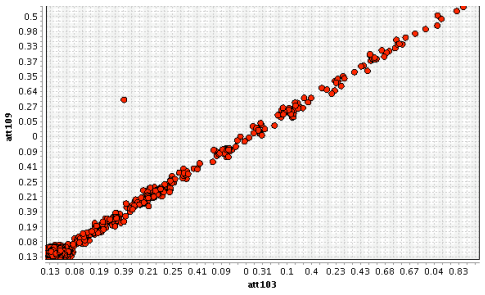
Figure 2. Histogram for Age



Histogram for Income

d) Below is a scatter matrix of two attributes from the Crime&Community dataset. What does this tell you about the relationship between the two attributes, and what is the significance of this in a data mining context?

(3 marks)



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

a) Why would you use cross validation when training a classification model? In your answer, include how cross validation works, and the purpose of both a training and a test dataset.

(10 marks)

b) Explain how an impurity measure such as **entropy** can be used to decide which attribute to select for each node on a decision tree. Use the data given below to illustrate your answer by calculating the entropy for **CapShape** and **GillSize**, and use that to determine which should be used as the root node of a tree to classifying mushrooms as edible or poisonous.

$$\text{Entropy}(S) = -p(\text{yes})\log_2 p(\text{yes}) - p(\text{no})\log_2 p(\text{no})$$

CapShape	**GillSize**	**Edible?**	
Round	Narrow	yes	
Round	Narrow	no	Note:
Round	Broad	no	Entropy(3,2)=0.97
Flat	Broad	yes	
Flat	Narrow	yes	
Flat	Broad	yes	
Round	Broad	no	
Round	Narrow	no	

(14 marks)

c) Explain what is meant by **pre-pruning** a decision tree. If mining a dataset that is known to be noisy, would you recommend generating a full decision tree or a pruned decision tree? Explain your answer.

(6 marks)

Question 4: Clustering and Evaluation **(30 marks)**

a) Calculate the Euclidean distance for the data shown in the table below. What needs to be done if the scales of the attributes are different?

row1	5	2	8
row2	7	5	3

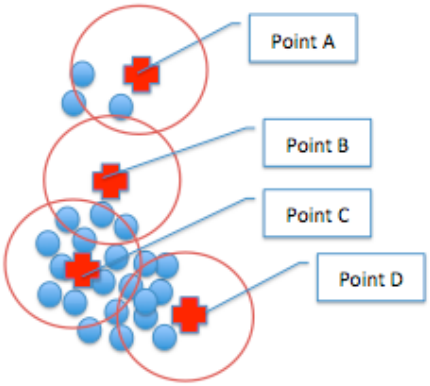
Note: $d(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$

(6 marks)

b) Differentiate between *hierarchical*, *partitional* and *density* based clustering.

(6 marks)

c) Define the DBScan clustering algorithm. For what type of dataset would DBScan be a better choice than K-Means? Using the diagram below, how would a DBScan algorithm label the four points (core, border or noise) assuming MinPts = 5, and Eps is illustrated by the four circles.



(9 marks)

d) Given the classification results in the following confusion matrix, complete the classification *accuracy*, *precision* and *recall* scores.

(9 marks)

Classified As:		**Correct**
Positive	**Negative**	
25	5	**Positive**
2	100	**Negative**