FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF MATHEMATICS AND STATISTICS

MID-SEMESTER ASSESSMENT

MODULE CODE: MA4128

SEMESTER: Spring

MODULE TITLE: Advanced Data Modeling

DURATION OF EXAM: 1 hour

LECTURER: Mr. Kevin O'Brien

GRADING SCHEME: 50 marks
20% of module grade

INSTRUCTIONS TO CANDIDATES

Scientific calculators approved by the University of Limerick can be used.
Statistical tables provided at the end of the exam paper.
Students must attempt ALL questions

Question 1. (10 marks) Nelson Rules for Control Charts

The **Nelson Rules** are a set of eight decision rules for detecting "out-of-control" or non-random conditions on control charts. These rules are applied to a control chart on which the magnitude of some variable is plotted against time. The rules are based on the mean value and the standard deviation of the samples.

(i) (4×2.5 Marks) Discuss any four of these rules, and how they would be used to detect "out of control" processes. Support your answer with sketches.

In your answer, you may make reference to the following properties of the Normal Distribution. Consider the random variable X distributed as

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

where μ is the mean and σ^2 is the variance of an random variable X.

- $\Pr(\mu - 1\sigma \leq X \leq \mu + 1\sigma) = 0.6827$
- $\Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.9545$
- $\Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.9973$

Marking Scheme: For Each Rule: A statment of the the rule is worth 1 Mark. A derviation of the probability is worth 1 Mark. A Sketch is worth 0.5 Marks.

Question 2. (10 marks) Binary Classification

(a) ***ROC Curves (4 Marks)***
What is a ROC curve? Explain its function, how it is determined, and the means of interpreting the curve. Support your answer with a sketchs.

(b) ***Performance Metrics (6 Marks)***
For following binary classification outcome table, calculate the following appraisal metrics.

(i.) (1 Mark) Accuracy;

(ii.) (1 Mark) Recall;

(iii.) (1 Mark) Precision;

(iv.) (1 Mark) F-measure.

	Predict Negative	Predict Positive
Observed Negative	9600	20
Observed Positive	300	80

(v.) (2 Marks) Explain why the F-measure is considered a more informative measure of performance than the Accuracy score.

Question 3. (10 marks) Hierarchical Clustering

(i.) (1 Mark) Compute the Euclidean distance between the following points.

$$A = (4, 6, 8, 2)$$

$$B = (3, 6, 1, 6)$$

(ii.) (2 Marks) Why do you standardize variables before carrying out a cluster analysis. Explain why using the standardized value may not be suitable in some cases? Give another example of numeric transformation.

(iii.) (4 Marks) Compare and contrast any three linkage methods. Support your answer with sketches

(iv.) (1 Mark) In the context of hierarchical cluster analysis, distinguish between agglomerative clustering and divisive clustering.

(v.) (2 Marks) How do we determine the appropriate number of clusters? Give two different visualization methods that are used to display the outcome of a cluster analysis.

Question 4. (10 marks) K-Means Clustering

(i.) (5 Marks) Explain the process of k-means clustering, starting with initial cluster allocation. You may work on the basis of a two-cluster solution. Support your answer with several sketches.

(ii.) (2 Marks) Compare and contrast k-means clustering and hierarchical clustering in terms of the number of cluster determined.

(iii.) (3 Marks) For a 4 cluster solution, Interpret the ANOVA table below.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Net profit	495,145	3	1419,744	3	,349	,795
Own funds	2878,202	3	2537,200	3	1,134	,460
Assets	842788,443	3	9987,138	3	84,387	,002
Client deposits	634017,636	3	35643,498	3	17,788	,021
Loans	957411,333	3	37401,709	3	25,598	,012

Question 5. (10 marks) Modelling Count Variables

(i.) (2 Marks) What is Poisson regression used to model. State any assumptions that must be checked before it can be used as an analysis.

(ii.) (1 Mark) The `R` Code output given below is used to predict the number of awards won by students.

- Information is provided on which of the three school programs the student takes part in (*General*, *Vocational* or *Academic*).
- Also we are given the mathematics test score.

State the mathematical formula used to predict the numer of awards won. *You can denote* ***progAcademic***, ***progVocational*** *and* ***math*** *as* x_1,x_2 *and* x_3 *respectively.*

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -5.2471     0.6585   -7.97  1.6e-15 ***
progAcademic     1.0839     0.3583    3.03   0.0025 **
progVocational   0.3698     0.4411    0.84   0.4018
math             0.0702     0.0106    6.62  3.6e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(iii.) (2 Marks) Use the model in Part (ii) to predict the number ofawards won by a general program student, witha maths score of 60.

(iv.) (3 Marks) What is Zero Inflation? Explain the Modelling Process for a Zero Inflated Model. Give an Example of Zero-Inflated Count Process. *Support your answer with a sketch, if necessary.*

(v.) (1 Mark) Describe a situation whereby Negative Binomial Regression Models would be used instead of Poisson Models.

(vi.) (1 Mark) What is Vuong Test used for?

Formulas and Tables

Critical Values for Dixon Q Test

N	$\alpha = 0.10$ *Confidence*= 0.90	$\alpha = 0.05$ *Confidence*= 0.95	$\alpha = 0.01$ *Confidence*= 0.99
3	0.941	0.97	0.994
4	0.765	0.829	0.926
5	0.642	0.71	0.821
6	0.56	0.625	0.74
7	0.507	0.568	0.68
8	0.468	0.526	0.634
9	0.437	0.493	0.598
10	0.412	0.466	0.568
11	0.392	0.444	0.542
12	0.376	0.426	0.522
13	0.361	0.41	0.503
14	0.349	0.396	0.488
15	0.338	0.384	0.475
16	0.329	0.374	0.463