Silence Please

Do Not turn over this page until advised to by the Invigilator

CIT Semester 1 Examinations 2018/19

Note to Candidates:	Check the Programme Title and the Module Description to ensure that you have received the correct examination. If in doubt please contact an Invigilator.
Module Title:	**Data Science and Analytics**
Module Code:	**DATA8001**
Programme Title(s):	HDip Sc Data Science Analytics MSc Data Science & Anlytcs P1 HDipSc Data Science Anlytcs PT
Block Code(s):	**SDAAN_8_Y5** **SDAAN_91Y5** **SDAANP8_Y5**
External Examiner(s):	**Dr. Niall Fitzgerald, Dr. Katarina Domijan**
Internal Examiner(s):	Mr. Michael Phelan, Mr. Aengus Daly
Instructions:	Answer Question One and Any Two Other Questions.
Duration:	2 Hours
Required Items:	

Q.1 Compulsory Question - Total 40 Marks – Answer any 4 parts.

a)	Explain the terms supervised, unsupervised learning and classification, regression in statistical learning. Give an example of a different algorithm for each term.	10 Marks
b)	Explain how KNN algorithm, (K-Nearest Neighbour) for both regression and classification works.	10 Marks
c)	Explain was is meant by overfitting and underfitting? Outline one strategy how these can be avoided.	10 Marks
d)	In regression tasks what is Leave-One Out cross validation, (LOOCV)? Why it is used?	10 Marks
e)	Standardising predictor variables is important in statistical learning. Explain why this is the case and give one example of how to standardise a variable.	10 Marks
f)	Give an explanation and a definition of : i) the expected value of a random variable **and** ii) the maximum likelihood estimator.	10 Marks

[Total 40 Marks]

Answer any 2 of the remaining 3 questions (all questions carry equal marks)

Q.2 Total 30 Marks

a)	List 6 of the 8 data protection rules as given by the Irish Data Protection Commissioner.	6 Marks
b)	Give 4 disadvantages of using decision trees as a classification model.	6 Marks
c)	Explain how the random forest algorithm for classification works, detailing how the in-built OOB (out-of-bag) error rates are calculated and the main parameter for tuning.	18 Marks

Q.3 Total 30 Marks

a)	Write a note on the GDPR.	10 Marks
b)	Write a note on the confusion matrix for a two class (binary) classification problem. Detail at least four measures, including false positive rate (FPR) and false negative rate (FNR), which can be calculated from the confusion matrix.	10 Marks
c)	Draw an example of a ROC curve and label the x-axis and the y-axis. Explain what a ROC curve measures and why it is used.	10 Marks

Q.4 Overleaf.

Q.4 Total 30 Marks

a) Explain how the decision tree algorithm works for regression. 16 Marks

b) Explain how the k-means clustering algorithm works and give 3 limitations of the algorithm. 14 Marks