CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Examinations 2018 – Autumn

Module Title: Data Science and Analytics

Module Code:	**DATA8001**
School:	**Science and Informatics**
Programme Title:	Higher Diploma in Science in Data Science and Analytics Master in Science in Data Science and Analytics
Programme Code:	CR_SDAAN_8 , CR_SDAAN_9
External Examiner(s):	**Prof. Michael Wallace, Dr Niall Fitzgerald**
Internal Examiner(s):	**Dr Michael Phelan, Mr Aengus Daly**
Instructions:	**Answer Question One and Any two other questions**
Duration:	**2 hours**
Sitting:	**Autumn 2018**

Requirements for this examination:

Note to Candidates: Please check the Programme Title and the Module Title to ensure that you are attempting the correct exaination.
If in doubt please contact an Invigilator.

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

a) Write a note on data cleaning; explain why it is important and give 3 techniques. 10 Marks

b) Give 4 ways that data visualisation can aid data analytics. 10 Marks

c) Give 3 advantages and 3 disadvantages of using Excel instead of R, Rapidminer or Python for data analytics. 10 Marks

d) Explain why testing and evaluation are important in Crisp-DM and detail 2 methods that are used in Data Analytics. 10 Marks

e) In data analytics explain the difference between classification and regression for supervised learning and give 2 methods/algorithms for each. 10 Marks

f) In decision/classification trees what is overfitting? Explain how it can be corrected. 10 Marks

[Total 40 Marks]

Answer any 2 of the remaining 4 questions (all questions carry equal marks)

Q.2 Total 30 Marks.

a)	Give 4 advantages and 4 disadvantages for using a datawarehouse.	12 Marks
b)	Give one method for assessing a model accuracy for supervised learning for regression and one for classification, giving details of both.	12 Marks
c)	List 6 of the 8 data protection rules as given by the Irish Data Protection Commissioner.	6 Marks

Q.3 Total 30 Marks.

a)	Write a note on anomaly detection giving 3 different methods that can be used.	10 Marks
b)	Write a note on bootstrapping detailing how it works and give 2 uses for it.	10 Marks
c)	Write a note on the GDPR.	10 Marks

Q.4 Total 30 Marks.

a)	In statistics explain what is meant by expected value.	10 Marks
b)	According to the Khatri and Brown (2010) article, Designing Data Governance Data Lifecycle is an important. Explain why this is the case.	10 Marks
c)	Write a note on entropy and information gain in regards to the decision tree classification algorithm. OR Discuss the "Art" & the "Science" in Data Science.	10 Marks

Q.5 Total 30 Marks.

a) Table 1 provides data on the height (x_i in cm) and weight (y_i in kg) for a sample group of students.

Calculate the intercept, slope and regression equation based on the data in Table 1. Using your regression equation, if a student is 157cm in height, what would their expected weight (in kg) be? — 13 Marks

b) What is TF-IDF document representation & how does it work? — 10 Marks

c) Detail the main characteristics of NoSQL databases and list the 4 different types of NoSQL databases. — 7 Marks

Table 1. Student Height & Weight Data

Note: you can use the empty grey cells to save your calculations.

Student	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	147	52.21				
2	150	53.12				
3	160	58.57				
4	170	64.47				
5	180	72.19				
Sum						
Mean						

Linear Regression Equations

- Regression Equation: $\hat{y} = w_0 + w_1 x$
- Slope: $w_1 = \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2}$ (N = 5 for this example)
- Intercept: $w_0 = \bar{y} - w_1 \times \bar{x}$

Where $\hat{y}$ is the predicted value of y (i.e., statistics grade), $\bar{x}$ is the mean x value (i.e., aptitude test) and $\bar{y}$ is the mean y value