Dublin Business School

QQI

HIGHER DIPLOMA IN SCIENCE IN DATA ANALYTICS

FINAL EXAMINATIONS

Module Code:	**B8IT109**
Module Description:	**Advanced Data Analytics**
Examiner:	**Dr Shahram Azizi Sazi**
Internal Moderator:	**Dr Amir Esmaeily**
External Examiner:	**Catherine Mulwa**

Date: 15th June 2020
Time: 18:30-20:30

INSTRUCTIONS TO CANDIDATES

- *This is an open book- material exam, students are allowed to use their own laptop, lecture notes, code, and websites to respond to the questions. Appropriate referencing must be used.*
- *Please select four questions out of five questions. Explicitly specify your selected questions on the top of the exam paper.*
- *R code and necessary outputs (i.e. graphs/plots/curves) need to be saved in word format and submit to Moodle.*

Question 1

`Use` **`mtcars`** `dataset` and consider **`disp`** `and` **`am`** as the attributes of interest.

a) Use the appropriate probability models to quantify the uncertainty in **disp** and **am**.
(5 Marks)

b) Estimate the parameters of your proposed models using the dataset.
(5 Marks)

c) Predict the future values of **disp** and **am** using (a) and (b).
(10 Marks)

d) Using (a), (b), find P(**disp** > 0.7). **(5 Marks)**

(Total: 25 Marks)

Question 2

Using the dataset available on,
`http://data.princeton.edu/wws509/datasets/cuse.dat`,
consider '**wantsMore**' as the output variable.

a) Split the dataset into 80% as the train-set and 20% as the test-set.
(2.5 Marks)

b) Apply Naïve Bayes (NB) algorithm to train the classifier using the train-set. **(2.5 Marks)**

c) Predict the test-set using the trained model of classifier. Express the functional form of the optimal NB classifier. **(5 Marks)**

d) Provide the confusion matrix and accuracy of predictions. **(5 Marks)**

e) Redo parts (b)-(d) to apply logistic regression algorithm. (**Hint**: consider $\alpha = 0.2$ and include **age**, **education**, and **notUsing** as input variables to implement the logistic classifier).
(10 Marks)

(Total: 25 Marks)

Question 3

Use the dataset '**quakes**', and consider **'mag'** as the output variable and select the set of input variables from the remaining columns. Split the dataset into 80% trainset and 20% as the testset.

a) Perform linear regression (LR) analysis and derive the optimal predictive model based on the trainset. (**Hint**: Use $\alpha = 0.05$ for the attribute selection). Predict the values of testset using the predictive model. **(7.5 Marks)**

b) Apply support vector regression (SVR) to predict the values of testset. **(7.5 Marks)**

c) Use RMSE to evaluate the accuracy of two models in 1000 Monte Carlo runs. Which method does provide a better prediction? **(10 Marks)**

(Total: 25 Marks)

Question 4

Use dataset available on http://www.stat.ufl.edu/~winner/data/clotthes_expend.csv , apply time series analysis, consider **sales.b** as your time series variable:

a) Validate the assumptions using graphical visualization.

(5 Marks)

b) Fit the optimized model for **sales.b** and provide the coefficient estimates for the fitted model. **(7.5 Marks)**

c) What is the estimated order for AR and MA?

(5 Marks)

d) Forecast h=10 step ahead prediction of **sales.b** on the plot of the original time series.

(7.5 Marks)

(Total: 25 Marks)

Question 5

Use dataset available on

http://www.stat.ufl.edu/~winner/data/iran_rock.csv,

a) Perform ANOVA and interpret the output. **(10 Marks)**

Load the dataset available on http://www.stat.ufl.edu/~winner/data/esp_studies1.csv,

b) Apply PCA, and identify the important principle components involving at least 80% of dataset variation. Explain your decision strategy. **(7.5 Marks)**

c) Use LDA to classify the dataset into few classes so that at least 85% of information of dataset is explained through new classification. **(7.5 Marks)**

(Total: 25 Marks)

End of Examination