QQI

HIGHER DIPLOMA IN SCIENCE IN DATA ANALYTICS

AUTUMN 2019 EXAMINATIONS

Module Code:	**B8IT109**
Module Description:	**Advanced Data Analytics**
Examiner:	**Dr Shahram Azizi Sazi**
Internal Moderator:	**Mr Paul Laird**
External Examiner:	**Dr Ralf Bierig**

Date: Monday, 2nd September 2019
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 different websites to respond the questions.*
- *Please select four questions out of five questions. Explicitly specify your selected questions on the top of 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: *Naïve Bayes, Decision Tree, Random Forest*

Loading the package **'datasets',** use the dataset **'readingSkills'** and consider **nativeSpeaker** as the output variable.

a) Split the dataset into 80% as the train-set and 20% as the test-set. (use set.seed(104)) **(2.5 Marks)**

b) Apply Random Forest (RF) algorithm to train the classifier using train-set with 20 trees. **(5 Marks)**

c) Predict the test-set using the trained model of classifier. **(2.5 Marks)**

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

e) Redo parts b-d to apply either Naïve Bayes or Decision Tree. Which model does provide the higher accuracy? **(10 Marks)**

(TOTAL: 25 Marks)

Question 2: *time series analysis*

Use **data('EuStockMarkets')** to load the in-built dataset 'EuStockMarkets' in R, consider ***DAX*** as your time series variable:

(a) Validate the assumptions using graphical visualization.

(5 Marks)

(b) Fit the optimized ARIMA model for ***DAX*** and provide the coefficient estimates for the fitted model. **(10 Marks)**

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

(5 Marks)

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

(5 Marks)

(Total: 25 Marks)

Question 3: *Regression analysis and support vector machine*

Loading the package **‘datasets’,** use the dataset ‘**trees’**, and consider **‘Girth’** as the output variable and select the others as the input variables. Split the dataset into 80% trainset and 20% as the testset (use set.seed(1456)).

a) Perform linear regression (LR) analysis and derive the optimal predictive model based on the trainset. (Hint: Use $\alpha = 0.01$ for the attribute selection). Predict the test values using the predictive model. **(10 Marks)**

b) Apply support vector regression (SVR) with the kernel ‘poly’ and predict the output variable of the testset. **(5 Marks)**

c) Use RMSE measure to evaluate the accuracy of two models in 100 Monte Carlo runs. Which method does provide more accurate prediction? ($Hint: RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$). **(10 Marks)**

(TOTAL: 25 Marks)

Question 4

Use dataset available on *'http://users.stat.ufl.edu/~winner/data/biodiesel_transest.csv'* , then:

(a) Based on the attributes of this dataset, propose an appropriate GLM to model **prop1_ec** as the target variable to other numerical variables. Express your reason.

(7 Marks)

(b) Specify the significant input variables on **prop1_ec** at the level of α=0.05, and estimate their corresponding coefficients.

(8 Marks)

(c) Train the model using 80% of this dataset, and predict 20% test dataset using the trained model. What is the best predictive model at the level of α=0.05. (Hint: use set.seed(1781))

(10 Marks)

(Total: 25 Marks)

Question 5 : *multivariate analysis and unsupervised learning methods*

Use dataset available on
http://users.stat.ufl.edu/~winner/data/hybrid_reg.csv

(a) Use LDA to classify the dataset into few classes so that at least 85% of information of dataset is explained through new classification. (**Hint**: model the output variable "**carclass_id"** to input variables "**msrp**", "**accelrate**", and "**mpg**"). How many LDs do you choose? Explain the reason. **(10 Marks)**

(b) Apply PCA to input variables, and identify the important principle components involving at least 90% of dataset variation. Explain your decision strategy? Plot principle components versus their variance (**Hint**: to sketch the plot use the Scree plot). **(5 Marks)**

(c) Use K-means clustering analysis to input variables and identify the most important classes. How many classes do you select? Why?

(5 Marks)

(d) Split the dataset into two sets of variables so that **X**=(msrp, mpgmpge) and **Y**=(accelrate, mpg). Apply canonical correlation analysis to find the cross-correlation between **X** and **Y**. What is the correlation between ***msrp*** and ***mpg***?

(5 Marks)
(Total: 25 Marks)

END OF EXAMINATION