UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

WST 212 Semester Test 3

21 June 2021

Monday Test submission

Instructions:

- Complete all questions in Section A and save all your code in a single R script, named **`st3_a.R`** (**Please ensure you use the capital R file extension**)
- Complete the questions in Section B and save all your code in a single R script, named **`st3_b.R`** **(Please ensure you use the capital R file extension)**
- Three submissions are required for this Test: submit **`st3_a.R`** under st3_A (practical) and **`st3_B.R`** under and complete interpretation submission under st3 (fill-in).
- Load the libraries as provided in the templates.
- Check the template for variable names answer sections.

Submission: Code 💻

- Multiple code submissions are allowed and your autograded results will be available shortly after each submission.
- Ensure all variables are named correctly, as incorrectly named variables will not be awarded any marks. (Remember variable names are case sensitive.)
- Ensure your code does not consist of any syntax errors. If your code produces errors when run, the autograder will not be able to mark it.
- Any code commented out (code is commented out when # is typed in front of it) will be considered rough work and will not be marked.
- Once you have completed your submission, ensure the file is submitted on Gradescope, with the correct file name. The autograder will only be able to grade your submission if you use the correct filename.

Guidelines:

- Two templates to be used for this assignment have been provided on ClickUP.
- Remember to assign your code to the variables indicated in this document.

Section A (st3_a.R)

Question 1 [2]

Consider the Super database. Write a query that returns the company name, contact person, email address, contact number, for only those numbers within the Pretoria area, i.e., area code 012. Organize the report by company name in alphabetical order. Save your query into the object named q1

Make use of the following table(s) in the Super database: • super_bursary

Question 2 [3]

Consider the Super database. The property managing agency would like you to compile a report of the average value of vacant and non-vacant flats for each owner. The report should only contain information where the average value of flats in suburb 1 is more than 3000 and should be ordered by the vacancy and owner id. The final report must include the following information only: owner id, vacancy status, and average price. Save your query into the object named q2

Make use of the following table(s) in the Super database: • super_flat_unit

Question 3 [4]

Consider the Super database. The property managing company would like a list of all tenants with behavioural complaints against them. Write an SQL query displaying the tenant surname, name, contact number (cellphone), behavioural complaint description and resolved status. Organize the report by tenant surname in alphabetical order. Save your query into the object named q3

Make use of the following table(s) in the Super database:

- super_tenant
- super_behaviour

Question 4 [4]

Consider the Super database. A South African investment group wants a breakdown of all international applicants. Write a query that includes the applicant surname, name, origin country, name of city and suburb of interest. Order the report by applicant origin country and surname. Save your query into the object named q4

Make use of the following table(s) in the Super database:

- super_applicant
- super_city
- super_country
- super_suburb

Section B (st3_b.R)

For this section, you have been provided a trainset and testset of the BreastCancer data where 0 indicates a Benign tumour patient and 1 indicates a Malignant tumour patient.

Question 1 [2]

Evaluate the class imbalance if any, of the training set

a) What is the proportion of the Benign class in the dataset (round of to 3 decimal places eg. 0.13259 rounds to 0.133)

b) What is the proportion of the Malignant class in the dataset (round of to 3 decimal places eg. 0.13259 rounds to 0.133)

Question 2 [1]

Is the data imbalanced (input a string: yes or no)"

Question 3 [1]

Fit a logistic regression model on your trainData with Class as your labels

Question 4 [1]

Predict the response variables of your test data

Question 5 [1]

Apply the following threshold to your predictions:

Anything above 0.5 as Benign and anything less than or equal to 0.5 as Malignant (use ifelse for this question)

Question 6 [4]

If 1 is negative case. Evaluate the following (round of to 3 decimal places eg. 0.13259 rounds to 0.133)

a) Precision
b) Recall
c) True Positive Rate
d) Accuracy

Question 7 [2]

Your supervisor requests you to fit a k-means learner to see if it will be able to pick up patterns of patients falling in the Benign and Malignant classes

a) Set seed to 200 and evaluate the total within sum of squares for 1 up to 10 cluster centres and plot your scree plot.
b) Set seed to 42 and train a k-means learner on 2 cluster centres.

Question 8 [2]

Calculate the proportions of the two clusters and compare them with the actual proportions of the testClass to answer the following

a) Are the extrapolated clusters representative of the actual classes (yes/no)
b) Between the logistic and the k-means model, which is the most reliable (input only the letter A or B)