NATIONAL UNIVERSITY OF SINGAPORE

Department of Statistics and Data Science

DSA1101 Introduction to Data Science

(Semester 1 - AY 2024/2025)

Individual Assignment

Due Date: 23:59 pm, Saturday 02 November 2024

INSTRUCTIONS TO STUDENTS

1. Students are supposed to submit your work on time. Any submission after the due time of the due date are marked as late.
2. **10% of the given mark will be deducted for each 2 hours late in submission.**
3. **No extension on the deadline for any circumstances.**
4. Students are required to complete this assignment individually.
5. submission is done online.
6. Your submission has **two separate files**. One is a .pdf file of the report, and the second file is of the R code (.R file or .Rmd file). Make sure that there is no error when the graders open and run your R code.
7. Be sure to lay out systematically the various parts and steps in your report.
8. Please use **set.seed(1101)** for your work.

Heart disease is the top disease that causes high death rate in the world. Data set given in the file `heart-disease-dsa101.csv` is a clean data set of 300 people with some variables that help to predict heart disease.

The description on a few variables is given below.

`sex`: 1= male; 0 = female

`chest.pain`: indicates the type of chest pain experienced by the patient (0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic).

`bp`: The patient's resting blood pressure measured in mm Hg upon admission.

`chol`: Patient's serum cholesterol level in mg/dl.

`fbs`: Fasting blood sugar level greater than 120 mg/dl (1 = true, 0 = false).

`rest.ecg`: The results of the resting electrocardiogram (0 = normal, 1 = ST-T wave abnormality, 2 = possible or definite left ventricular hypertrophy).

`heart.rate`: The highest heart rate achieved during exercise testing.

`angina`: Whether the patient experienced angina induced by exercise (1 = no, 0 = yes).

`st.depression`: ST depression observed in the electrocardiogram during exercise relative to rest, measured in mm.

`vessels`: Number of major vessels (ranging from 0 to 4) visible by fluoroscopy.

`blood.disorder`: A blood disorder (1 = normal, 2 = fixed defect, 3 = reversible defect; 0 = missing value).

`disease`: Indicates the presence or absence of heart disease (1 = presence, 0 = absence).

Purpose of this assignment: Write a statistical report to show your work on choosing a classification method for predicting heart disease status; and propose the best classifier. That means, for each classifier fitted, you need to investigate on the goodness of fit of it.

Suggestion for the main part of the report

Part 0 Introduction

1. In this part, you need to introduce the problem (main goal of the report), introduce the data set and the steps that you plan to follow in the subsequent parts of the report.

Part I EDA: Exploring the variables and association

2. You should summarize/describe the response variable as well as brief understanding about each input variable.

3. You should check the association between the response and each input variable. Give your comment on the strength of the association. This step is to identify the potential inputs to add to a model/classifier.

   Hint: Topic 2 could help you know how to check the association between 2 variables.

Part II Methods: Building Model/Classifier

4. For this assignment, you are required to perform 3 methods KNN, Decision trees (DT), and Logistic regression (LR).

   For each method, you may consider to find the best arguments to use (for example, find the best `"misplit"` for DT; find the best k for KNN). While finding the best argument to use, you should use 5-fold CV to evaluate based on TPR - True Positive Rate (which is also known as sensitivity).

   It is advised that the 5 folds should be unchanged for every classification methods used (KNN, DT).

5. After finding the best argument for KNN and DT, you will compare the performance of KNN, DT and LR with each other. That means, for each classifier, derive its goodness of fit by TPR, precision, ROC and AUC, on the full data set of 300 observations.

6. Comments on pros and cons of each classifier fitted (KNN, DT and LR).

Part III Conclusion: The Best Model/Classifier

7. Propose the best classifier (your choice of classifier).

Format of the report

1. Your report is a .pdf file, limited to **no more than SIX printing pages, font size 12, margins of not less than 0.75 inches**.
2. Table and/or figure in the report should be numbered clearly.
3. If you submit the report without submitting R code file, your mark will be deducted by half of the mark given to your report.
4. If you add any R code into your report, it will still be counted within the six pages allowed. Hence, it's advised not to add R code into your report.

END OF ASSESSMENT