### NATIONAL UNIVERSITY OF SINGAPORE

Department of Statistics and Data Science

### DSA1101 Introduction to Data Science

(Semester 2 : AY 2023/2024)

Individual Assignment

Due Date: 23:59 pm, Sunday 14 April 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. Make sure that there is no error when the graders open and run your R code file.
- 7. Be sure to lay out systematically the various parts and steps in your report.
- 8. Your submission files should be named as A0123456B.pdf and A0123456B.R where A0123456B is your student number.
- 9. Please use set.seed(1101) for your work.

Diabetes is among the most prevalence chronic diseases in the world. Data set given in the file diabetes-dataset.csv is a clean data set of 100,000 survey responses, provided by the author Mohammed Mustafa.<sup>1</sup>

The description on a few variables is given below.

hypertension: 0 = No; 1 = Yes

heart\_diease: 0 = No; 1 = Yes

smoking\_history: current = currently is smoking; ever = smoked sometimes but not often; former = smoked before but has completely quitted; never = never before and after; not current = before not smoking but not sure for future

gender = Female, Male and Other (LGBT)

**Purpose of this assignment**: Write a statistical report to show your work on choosing a classification method for predicting diabetes status; and propose the best classifier. Investigate on the goodness of fit of the classifiers fitted.

 $<sup>^{1}</sup>$ https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

# Suggestion for the main part of the report

## Part I Exploring the data set

- 1. You should summarize/describe the response variable as well as input variables.
- 2. You should check the association between the response and each input variable before fitting any models/classifier. Comment on the strength of the association. This step is to identify the potential features for the model/classifier.
- 3. It is advised to separate the full data set into two parts for training and for testing with the ratio of 8:2, respectively.

## Part II Building Model/Classifier and Conclusion

- 4. Propose some models/classifiers.
- 5. For each model/classifier, examine its goodness of fit: by ROC and AUC, and at least one of the 5 metrics introduced in Topic 4.
- 6. Comparing the goodness of fit between models/classifiers fitted, propose the best one (final model).
- 7. Comments on pros and cons of each model fitted.
- 8. You might consider (**optional**) to use N-fold CV to find the best value for parameter in the model you fit (such as best k for KNN; or best cp or minsplit for decision tree.)

## Few Notes

1. Note 1: Each student must report your work on at least three different models/classifiers.

2. Note 2: Different student might have different choice for the final model. However, you need to justify your choice clearly.

## Format of the report

- 1. Your report is a .pdf file, limited to no more than SIX printing pages, font size 12.
- 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