NATIONAL UNIVERSITY OF SINGAPORE

Department of Statistics and Data Science

DSA1101 Introduction to Data Science

(Semester 1 - AY 2025/2026)

Individual Assignment

Due Date: 23:59 pm, Thursday 06 November 2025

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.**
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 as well as in your R code file.
8. Please use **set.seed(611)** 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.[1]

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. That means, for each classifier fitted, you need to investigate on the goodness of fit of it.

[1] https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

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 into model/classifier.

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

Part II Methods: Building KNN, DT and LR Classifiers

4. Finding the best version for each method.

   (i) It is advised that you should use 5-fold CV to evaluate the goodness of fit of each method. The 5 folds hence should remain unchanged for 3 methods used.

   (ii) You may consider to find the best version of each method before comparing them with each other. For example, for DT, find the best "`misplit`" or best "`maxdepth`"; For KNN, find the best k; For LR, find the best model such that it should not be too complex.

   (iii) In order to find the best version of each method, you could use Type 1 error as the comparing criterion.
5. After finding the best version for each method, you will compare the performance of KNN, DT and LR with each other on the full data set. That means, for each method, derive its goodness of fit by TPR, precision, ROC and AUC, on the full data set of 100,000 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) for this data set.

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**.

   The standard for a typed A4 document is 1 inch for each margin.

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.

Notes: This is a statistical report, hence, please present it in a report style. It should not be presented as a list of answers/solutions for a list of questions.

END OF ASSESSMENT