NATIONAL UNIVERSITY OF SINGAPORE

DSA1101 INTRODUCTION TO DATA SCIENCE

(Semester 1 : AY 2025/2026)

Time Allowed: 2 Hours

INSTRUCTIONS TO STUDENTS

1. Please indicate only your student number on your answer file. **Do not indicate your name.**
2. This exam paper contains **THREE (3)** problems and comprises **SIX (6)** printed pages (including the cover page).
3. The data file used for this exam is given in the folder "For Finals" on Canvas/DSA1101/Files.
4. This is an OPEN BOOK BLOCKED INTERNET exam. You may refer to your lecture notes, tutorials, textbooks or any notes that you have made, but you are not allowed to perform any online search.
5. The use of offline Large Language Models (LLMs) is prohibited for this exam.
6. All calculators in the List of Approved Calculators are allowed.
7. Students are required to answer ALL questions. Total mark is 100.
8. During the exam, you are not allowed to communicate with any person other than the invigilators.
9. At the end of the exam,
   - Copy and paste your R code into the Examplify text box. Indentation and alignment may not be retained when pasting, but that is acceptable.
   - Save an exact copy of your R file on your laptop, and upload it to Canvas/DSA1101/Assignments/Final Exam Submission immediately after the exam.
10. Do not modify the code in your submission file. Any difference found (except for indentation or alignment) between Examplify and Canvas submissions will be penalized.

1. (*20 points*) Please type out your answers in the R file. No explanation is needed.

(a) The coefficient of determination of a multiple linear regression model, R^2, can decrease when more predictors are added to the model. True or False?

(b) The k-nearest neighbors classification works equally well regardless of whether features are measured in different units. True or False?

(c) Information gain is always non-negative when splitting a node in a decision tree. True or False?

(d) The Naïve Bayes classifier functions well even with features that are highly associated with one another. True or False?

(e) The coefficients in both linear and logistic regression models are estimated through minimizing the sum of squared errors. True or False?

(f) The k-means algorithm will give the same clusters regardless of your starting centroids. True or False?

(g) In association rules, *Leverage* cannot be negative. True or False?

(h) Ethical data science is solely about following laws and regulations. True or False?

(i) Data scientists should consider potential malicious uses of their models as part of ethical responsibility. True or False?

(j) In association rules, consider the rule $A \to B$. Which of the following statements is True?

A. Support measures how often B occurs given that A occurs.

B. Confidence measures how often A and B occur together in the data set.

C. Lift greater than 1 indicates that A and B occur together more often than expected by chance.

D. A low positive leverage implies that A and B occur together less often than expected by chance.

2. (*10 points*) Given a numeric vector of 10 values below.

$$22, 27, 12, 19, 22, 25, 28, 16, 13, 24$$

(a) Define a function in R, named **ave**, which returns the mean of a numeric vector. You are not allowed to use the two built-in functions `mean` or `summary` in R in your code.

Using your function, calculate and report the mean of the vector above.

(b) Define a function in R, named **med**, which returns the median of a numeric vector. You are not allowed to use the two built-in functions `median` or `quantile` in R in your code.

Using your function, calculate and report the median of the vector above.

Notes:

(i) Functions **ave** and **med** must have at least one argument, named **numvec**, to specify the numeric vector.

(ii) Definition of median: If the values in the numeric vector is sorted from smallest to largest, then median is the middle value if the length of that vector is an odd number and median is the average of the two middle values if the length of that vector is an even number.

3. (*70 points*) Consider a random sample of 70,000 people. Their health information is recorded for a study about factors that may affect the presence of cardiovascular disease. The data set is given in the file `cardio.csv`. The table below gives the description for some variables in this study.

Variable	Description
`age`	age (in days)
`gender`	gender: 1 = Female, 2 = Male
`height`	height (in cm)
`weight`	weight (in kg)
`ap_hi`	systolic blood pressure
`ap_lo`	diastolic blood pressure
`cholesterol`	cholesterol level: 1 = normal, 2 = above normal, 3 = well above normal
`gluc`	glucose level: 1 = normal, 2 = above normal, 3 = well above normal
`smoke`	smoking: 0 = No, 1 = Yes
`alco`	alcohol intake: 0 = No, 1 = Yes
`active`	physical activity: 0 = No, 1 = Yes
`cardio`	presence of cardiovascular disease: 0 = No, 1 = Yes

For the questions below,

- Load the data set into R and name it as **data**.
- Report the final numerical answer to three significant figures if its absolute value is smaller than one (e.g. 0.0123) and to three decimal places if its absolute value is larger than one (e.g. -2.345).
- Do not split the data set given into train set and test set.

Part I: Regression Models (10 points)

1. Write code to fit a regression model using only the numeric input features, to be named as **LM**, for predicting the probability of having cardiovascular disease. You are to choose the more appropriate model between a linear regression model and a logistic regression model. Using a significance level of 0.05, are there any insignificant features in that fitted model?
2. Using **LM**, compute and report the probability of having a cardiovascular disease for a 20,000 days old male, who is 170cm and 65kg, has systolic blood pressure reading of 120 with diastolic blood pressure reading of 90, has normal cholesterol and glucose levels, does not smoke or take alcohol, and does not do physical activity.
3. Using **LM**, interpret the coefficient for weight.
4. Write a code to plot the ROC curve of the model **LM** in red color based on its predictions. Derive and report the value of AUC. How many different values of threshold δ was used in creating the ROC curve?

Part II: Decision Trees (20 points)

All the decision tree models below are using Information Gain for branch split.

5. Using only complex parameter $cp = 0.005$ to specify the complexity of the tree, write code to fit a decision tree, named **DT**, with all the inputs given in the data set.
6. Write code to plot the fitted tree. Report the names of the inputs that appear in the fitted tree. Among those inputs, which one is the most important?
7. From the plot of the tree, obtain and report the predicted probability of having cardiovascular disease for the person mentioned in Question 2.

8. Write code to predict the status of cardiovascular disease of people in **data.** Calculate and report the accuracy and Type I error rate for the prediction.

9. Write code to predict the probability of having cardiovascular disease of people in **data** where the vector of probabilities is named as **pred.dt.prob**.

10. Write code to plot the ROC curve of the tree **DT** in black color based on its prediction for **data.** Derive and report the value of AUC.

11. Let **delta** denote a vector of values for threshold in the range from 0.3 to 0.7 inclusive and rounded to 1 decimal place. For each value of threshold in **delta**, write code to

    (i) predict the status of having cardiovascular disease for people in **data**

    (ii) calculate accuracy and Type I error rate

    Conclude which threshold in **delta** gives highest accuracy. Separately, conclude which threshold gives lowest Type I error rate.

Part III: Naive Bayes Classifier (20 points)

12. We now fit a naive Bayes classifier for **data** with all the inputs given. Write code to fit the classifier, to be named as **NB.**

13. Write code to get the predicted probability of having cardiovascular disease for the person mentioned in Question 2. Report the probability.

14. Write code to predict the probability of having cardiovascular disease for people in **data** where the vector of probabilities is named as **pred.nb.prob.**

15. Write code to plot the ROC curve of the classifier **NB** in blue color based on its prediction for **data.** Derive and report the value of AUC.

16. Let **alpha** denote the vector of threshold values used for the ROC curve in Question 15. Write code to plot a figure that shows how the TPR and the FPR change when the threshold changes.

17. Propose a value of the threshold α (rounded to three decimal places) such that the classifier **NB** can attain a TPR of at least 0.8 while the FPR is as low as possible.

18. With the proposed value of α in Question 17 and the predicted probabilities in **pred.nb.prob**, write code to calculate the accuracy of the prediction.

Part IV: KNN classifier (20 points)

Notes:

(i) For this part, we consider **cholesterol** and **gluc** as numeric inputs.

(ii) R might take a while to run each command `knn()`.

19. Write code to create a new data frame, called **data.KNN** which its columns are all the numeric inputs after standardization and its last column is the response column.

20. Run command `set.seed(666)`.

    Write code to fit a KNN classifier with $k = 3$ for **data** to predict the status of having cardiovascular disease for people in **data**.

    Write code to derive the accuracy and Type I error rate for the prediction.

21. Write code to find the predicted probability of having cardiovascular disease for people in **data**.

22. Write code to plot the ROC curve of the 3-NN classifier in red color based on its prediction for **data**. Derive and report the value of AUC.

23. In the plot in Question 22, write code to add the ROC curve of **DT** in Question 10 in black color and ROC curve of **NB** in Question 15 in blue color.

    Which curve has the highest AUC value?

–END OF PAPER–