University of Windsor
Faculty of Science

COMP 8157 Advanced Database Topics
University of Windsor, School of Computer Science
Lab 1
Weight: 10 %
Due: May 16th [11:59]

Aim: This assignment will assess your understanding of data Mining concepts.

Submission:

1. **A report (word file) to answer all questions and**
2. **the R file**

Due:

To accomplish this assignment, you need to:

- Download the "**Heart Attack**" from Brightspace
- Use R studio for analyzing and visualizing the dataset.
- Grades are given for each question.

Place this confidentiality statement in your submission report.

CONFIDENTIALITY AGREEMENT & STATEMENT OF HONESTY
I, ___________________ verify that the submitted work is my own, original work, and that I did not use Generative AI tools (e.g., ChatGPT, Bard) to produce this lab report. I confirm knowing that a mark of 0 may be assigned for sharing or copying this work.

________________________ ________________________ ________________________
Student Signature Student Name Student I.D.

________________________ ________________________ ________________________
Student Signature Student Name Student I.D.

________________________ ________________________ ________________________
Student Signature Student Name Student I.D.

Part 1: Data Exploration (12 marks)

1. Import the **Heart Attack** datasets **(2 marks)**.
2. Summarize that Heart Attack dataset and explain the output **(4 marks).**
3. Show the structure and dimension of the dataset and explain it **(2 marks)**.
4. Show the first 8 rows and the last 5 rows of the dataset **(2 marks)**.
5. Show the column names of the Heart Attack dataset **(2 marks)**.

Part 2: Data Pre-Processing (28 marks)

6. What is the class variable in the Heart Attack dataset? What does it indicate **(4 marks)**.?
7. What is the datatype of the class variable **(4 marks)**?
8. Change the class type of the class variable of Heart Attack dataset to factor. Show the output after the conversion **(4 marks)**.
9. Find the sum of the missing values in Heart Attack dataset **(4 marks)**.
10. Find which columns contain missing values in the dataset. What is the total missing values for each column **(4 marks)**?
11. Replace the missing values in the Heart Attack by 0. Check what if the missing values was replaced successfully **(4 marks)**.
12. Rename the sex attribute from (0 and 1) to (Male and Female). Show the conversion output of the specific attribute **(4 marks)**.

Part 3: Data Visualization (60 marks)

13. Create a scatter plot. The plot should show the relationship between the cholesterol and the age attributes **(10 marks)**.
    a. Add labels, title, and color to the plot. The color should be blue.
    b. Add open red triangles to the plot.
14. Use the ggplot function to plot any two variables **(10 marks)**.
    a. The points shape should be filled square.
15. barplot the ‘age’ variable of the Heart Attack dataset **(10 marks)**:
    a. Add labels, title, and color to the plot.
16. Create a histogram of the ‘cp’ attribute **(10 marks)**:
    a. Find the minimum and maximum of the attribute.

b. Add a break function and use the seq(x, y, z) function.
c. Add labels, title = (Chest Pain type), and color to the plot.

17. Boxplot the ‘age’ attribute and explain the output **(10 marks)**.
18. Create a correlation plot of the whole dataset variables and explain the output. Do not forget to convert some of the variable’s datatype **(10 marks).**