Data and Data Preprocessing

Problem 1: Types of Attributes (14 points)

Classify the following attributes as nominal, ordinal, interval, ratio. **Explain why.**

(a) Rating of an Amazon product by a person on a scale of 1 to 5

(b) The Internet Speed

(c) Number of customers in a store.

(d) UCF Student ID

(e) Distance

(f) Letter grade (A, B, C, D)

(g) The temperature at Orlando

Problem 2: Exploring Data Preprocessing Techniques (26 points)

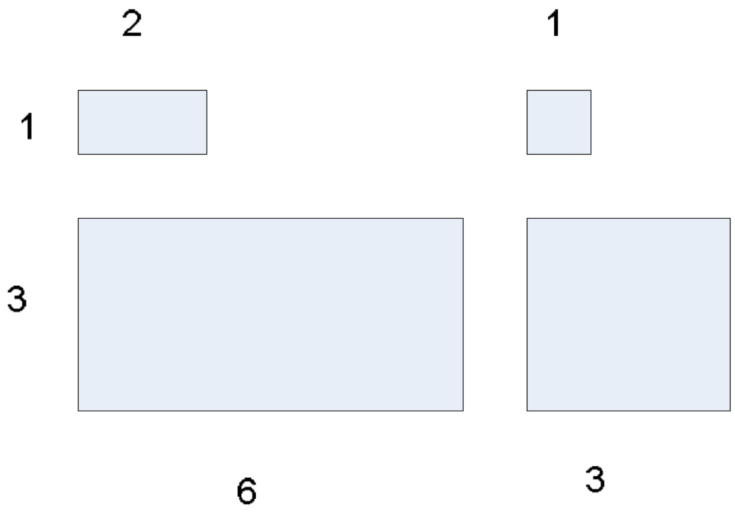
Read the solution post of the Kaggle Titanic Dataset: https://www.kaggle.com/code/preejababu/titanic-data-science-solutions. Run the code and reproduce the data preprocessing and classification modeling steps.

Q1 (Reproduce): Please read, understand, run the code and reproduce the model accuracies. Please briefly explain whether you can reproduce the classification accuracies of 'Support Vector Machines', 'KNN', 'Logistic Regression', 'Random Forest', 'Naive Bayes', 'Perceptron', 'Stochastic Gradient Decent', 'Linear SVC', 'Decision Tree'. (10 points)

Q2 (Improve): Is the data preprocessing process proposed in the Kaggle post the best preprocessing solution? If yes, please explain why. If not, can you leverage what you learned in the class and your previous experiences to improve data processing, to obtain better accuracies for all these classification models? Describe what is your improved data preprocessing, and what are your improved accuracies? (16 points)

Problem 3: Distance/Similarity Measures (10 points)

Given the four boxes shown in the following figure, answer the following questions. In the diagram, numbers indicate the lengths and widths and you can consider each box to be a vector of two real numbers, length and width. For example, the top left box would be (2,1), while the bottom right box would be (3,3). Restrict your choices of similarity/distance measure to Euclidean distance and correlation. **<u>Please explain your choice.</u>**



Which proximity measure would you use to group the boxes based on their shapes (length-width ratio)?

Which proximity measure would you use to group the boxes based on their size?

Please submit a PDF report. In your report, please answer each question with your explanations, plots, results in brief. DO NOT paste your code or snapshot into the PDF. At the end of your PDF, please include a website address (e.g., Github, Dropbox, OneDrive, GoogleDrive) that can allow the TA to read your code if any.