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

Ratio: because rate can be distincted,can be ordered. Many people's rates can be added together. To get the average, rates can be divided with number of people.

(b) The Internet Speed

Interval:1. It can be distincted, same number means same speed. Different number means different speed.2. Can be ordered. Higher number means higher speed, lower number means lower speed.3. Can be added. Total speed of one building can be get from adding speed together in this building. 4. Can do multiplication: to get the average speed, can divided thetotal speed in the building by the number of device.

(c) Number of customers in a store.

Ratio:1. It can be distincted, same number means same amount of customer in store. Different number means different amount of customer in store.2. Can be ordered. Higher number means more customers in store.3. Can be added. Total number of customer in one year can be get from adding number of customers together in one year. 4. Can do multiplication: to get the average number of customers, can divided the total customers in one year by days of one years.

(d) UCF Student ID

Nominal: 1. It can only be distincted.

(e) Distance

Ratio:1. It can be distincted, same number means same distance. Different number means different distance.2. Can be ordered. Higher number means higher distance.3. Can be added. Total distance in a travel can be get from adding distance in each section together. 4. Can do multiplication: to get the average distance, It can be get by dividing the distance by days.

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

Ordinal: 1. It can be disntincted A is not euqal to B.2. Grade A is better than Grade B so it is ordered.But it can not do addition.

(g) The temperature at Orlando

Ratio:1. It can be distincted, same number means same temperature. Different number means different temperature.2. Can be ordered. Higher number means higher temperature.3. Can be added. Total distance in one year can be get from adding number of customers together in one year. 4. Can do multiplication: to get the average number of customers, can divided the total customers in one year by days of one years.

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.

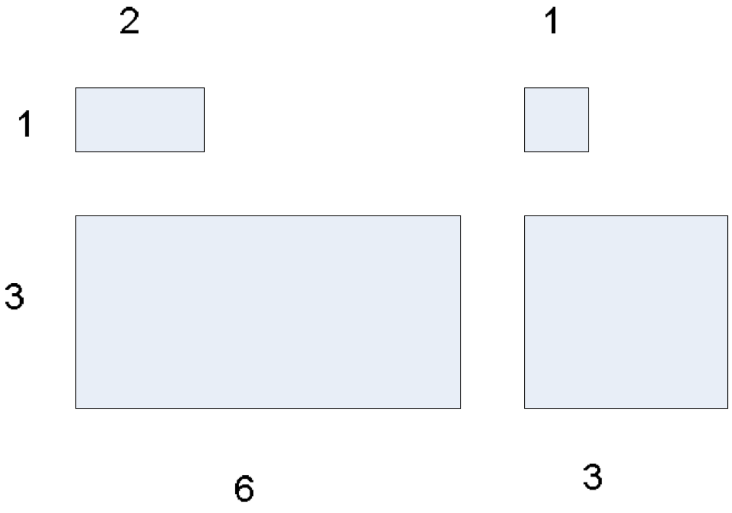
Q1 (Reproduce): Please read, understand, run the code and reproduce the model accuracies. Please briefly explain whether you can reproduce the classification accuracies of 'Support Vector Machines', 'KNN', 'Logistic Regression', 'Random Forest', 'Naive Bayes', 'Perceptron', 'Stochastic Gradient Decent',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

Q1:
Yes, I can reproduce the code. Because the enviroment and package are the same. Execpt for these, I do the same data process to the same dataset. So that I can reproduce it and have the same result.
Q2:
I think the Kaggle's solution is the best for now. Since it creates new feature; remove the feature that does not contribute to the model and convert the non-numerical feature to numerical feature.It also extra features from existing feautre which has much error and could reduce the accuracy of the model.

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. **Please explain your choice.**



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?

Question1:
I will choose correlation. Because the length-width ratio represent the shape of boxs,
the correlation will tell us if the length and width will increase with the same ratio.
Question2:
I will use Euclidean distance,because these are two dimension's figuare, using Euclidean
distance is easier to understand and compare.

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.

Github link is:https://github.com/zhenqi72/HW_datamining.git