UTD

CS 4375 - Assignment#1
Due Date: 2/24/23, 11:59 pm

1-Explore the attached dataset (Heart Failure Prediction Dataset) and answer the following questions

- How many objects and attributes are there in the dataset ?
- Specify the type of each attribute (Nominal, Ordinal, Interval, or Ratio).
- Is there any missing values or duplicate data objects in the dataset? If so, using techniques we discussed in the class improve the quality of the dataset.
- After improving the quality of the dataset, using histogram, boxplot, ggplot, or barplot, display the distribution of attributes (at least 5 attributes). Report your findings (at least 3 lines)
- After improving the quality of the dataset, using boxplot or ggplot, show the relationship between a continuous attribute and a discrete attribute (draw at least 4 plots). Report your findings (at least 3 lines)

2-Extract continuous features and sex feature from the improved dataset. Compute the averages for them, grouped (aggregated) according to sex. Report your findings (at least 3 lines).

3-Extract continuous features of the improved dataset. Standardize the scale of them to make them comparable. Calculate Euclidean distance between 10 objects before and after Standardization. Report your findings (at least 3 lines).

4-Using random sampling with replacement, create a sample (300 objects) from the improved data set. How many duplicated objects are there in the sample? Report your findings (at least 3 lines).

5-Extract RestingBP, Cholesterol and FastingBS features from the improved dataset. Draw a scatterplot3d for them. Then, project these data points into a 2 dimension area using PCA method. Report your findings (at least 3 lines).

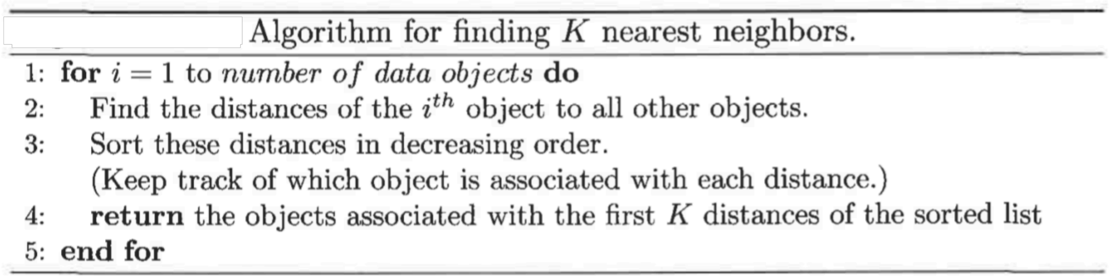
6-Discretize the Age attribute to 4 categories using equal interval and equal frequency methods. For both methods, report the interval width and number of objects in each interval (at least 3 lines).

7-Extract 2 continuous features and 50 objects from the improved dataset. Calculate Pearson correlation matrix for them. Is there any linear relationship between them? Why ? (at least 3 lines)

8-For measures the following vectors, x and y, calculate the indicated similarity or distance . PLease show your works

x: (1, 1,0,0), y : (1,1,0,0) cosine, correlation, Euclidean, Jaccard

9-Consider the problem of finding the K nearest neighbors of a data object. A programmer designs the following algorithm for this task.



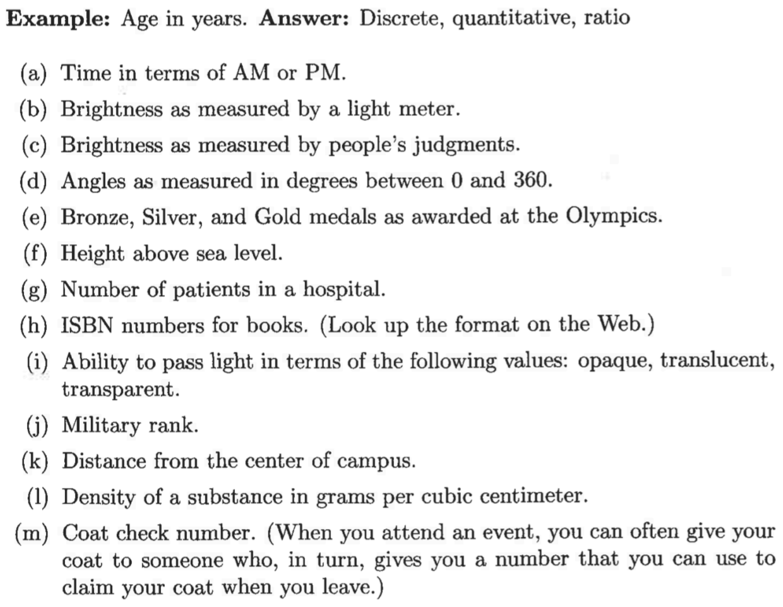
Algorithm for finding K nearest neighbors.

```
for $i = 1$ to number of data objects do
    Find the distances of the $i^{th}$ object to all other objects.
    Sort these distances in decreasing order.
    (Keep track of which object is associated with each distance.)
    return the objects associated with the first $K$ distances of the sorted list
end for
```

(a) Describe the potential problems with this algorithm if there are duplicate objects in the data set. Assume the distance function will only return a distance of 0 for objects that are the same.

(b) How would you fix this problem?

10-Classify the following attributes as binary, discrete, or continuous. Also, classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.



Example: Age in years. **Answer:** Discrete, quantitative, ratio

(a) Time in terms of AM or PM.

(b) Brightness as measured by a light meter.

(c) Brightness as measured by people's judgments.

(d) Angles as measured in degrees between 0 and 360.

(e) Bronze, Silver, and Gold medals as awarded at the Olympics.

(f) Height above sea level.

(g) Number of patients in a hospital.

(h) ISBN numbers for books. (Look up the format on the Web.)

(i) Ability to pass light in terms of the following values: opaque, translucent, transparent.

(j) Military rank.

(k) Distance from the center of campus.

(l) Density of a substance in grams per cubic centimeter.

(m) Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave.)