**CS612-Lab2**

**Question1: 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**

1. Time in terms of AM or PM

**Binary, qualitative, ordinal**

1. Brightness as measured by a light meter.

**Continuous , Quantitative and ratio**

1. Brightness as measured by people’s judgments:

**Discrete, qualitative, ordinal ( we can choose from a discrete set of ratings)**

1. Angles as measured in degrees between 0◦ and 360:

**Continuous, quantitative, ratio**

1. Bronze, Silver, and Gold medals as awarded at the Olympics.

**Discrete, qualitative, ordinal**

1. Number of patients in a hospital.

**Discrete, quantitative, ratio**

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

**Discrete, qualitative, ordinal**

1. Military rank.

**Discrete, qualitative, ordinal**

1. Density of a substance in grams per cubic centimeter.

**Continuous, quantitative, ratio**

1. 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.

**Discrete, qualitative, nominal**

**Question 2: You have been given a data file called “AlzheimerData.xlsx”. It is a matrix file with 275 rows (representing Drugs) and 26 columns (representing properties of the drug) for Alzheimer Disease to do data mining and create a model to predict if a drug can be used for Alzheimer disease or not. Step 1 is to clean the data. You are required to write a python program that does the following:**

* Remove the columns that only include Zero
* Remove the columns which have more than 5 non-digit characters. Otherwise if they have less than 5 junk characters, replace the junk character with zeros.
* Remove the rows which have more than 5 non-digit characters. Otherwise they have less than 5 junk characters, replace the junk character with zeros.

After cleaning the data, create two separate matrices. The first matrix should scale the data

**x\_standarization = (x – mean(X)) / (StdDev(X))**

The second matrix should normalize the data

**x\_Normalization = (x – min(X)) / ((max(X) – min(X)))**

Place all the above files in one folder, name your folder based on your first and your last name ex: (A1-CS612-JessicaSharma)

Note: You should not write your program like C++. You ***MUST*** use the as many as available built-in functions located in the following libraries to optimize and minimize your code. The most important libraries are:

* NumPy, Scikit-Learn, SciPy, Pandas, Keras, PyTorchm etc

**Test your program with a data set that is provided in this lab.**

* **Place the questions with the answers to question 1 in one file, call it “Lab2-Q1”**
* **the code for the program in another file call it “Lab2-Q2-Code.py”**
* **the snapshot of the output of the program in the third file., call it “Lab2-Q2-CleanData.txt”**
* **Place the standardized file in “Lab2-Q2-Rescaled.txt”**
* **Place the normalized file in “Lab2-Q2-Normalized.txt”**
* **Place all the files in a folder and call the folder based on your name and your partner’s name (ex:Jack-and-Nancy-Lab2)**
* **Zip the folder and only one person in the team is required to submit the work.**