Arkansas Tech University
Graduate College, Information Technology
INFT 6903 Data Science, Summer 2022
Assignment 6 / Due date: July 13th, 2022 before midnight. Total points: 100. Good luck!

Important Note: You can use any kind of Python compiler for this assignment. Also note that, some of online compilers **do not** support/provide some data science libraries like Numpy and Pandas. We will continue using *Anaconda* and *Jupyter notebook* in data analysis and machine learning algorithm development (clustering and classification). Therefore, I recommend using *Anaconda* and *Jupyter notebook* in this assignment. Please also find the *Instructions to Download Anaconda and Jupyter Notebook* pdf file from Blackboard under the Week5 folder.

Anaconda and Jupyter notebook: https://www.anaconda.com/
Google Colab: https://research.google.com/colaboratory/
Onlinegdb: https://www.onlinegdb.com/online_python_compiler
Replit: https://replit.com/languages/python3
Downloading&Installing a Python editor from the Python website: https://www.python.org/

---

1. (*30 points*) Download the *sample_dataset1.txt* file to your desktop from the Week6 folder on Blackboard.

a) Use Pandas to read the file and store them in a DataFrame (DF) object named data. Then, display first few rows of the data frame (*Hint*: You can see the Slide 28 from our PowerPoint presentation file from the Week6 folder to write the path of the data set).

b) For each quantitative attribute (first attribute, second attribute, third attribute, and fourth attribute) calculate its average, standard deviation, minimum, and maximum values.

c) For each qualitative attribute (class), count the frequency for each of its distinct values.

d) Display the summary for all the attributes simultaneously in a table using the *describe()* function.

e) Display the histogram for the first attribute by discretizing it into 8 separate bins and counting the frequency for each bin.

f) Show the distribution of values for each attribute.

g) For each pair of attributes, use a scatter plot to visualize their joint distribution.

h) Plot the distribution values with parallel coordinates.

2. (*15 points*) Download the *sample_dataset2.csv* file to your desktop from the Week6 folder on Blackboard.

a) Load this data set using Pandas. Then, draw a line plot of its daily time series.

b) Draw the monthly time series line plot of the same data set.

c) Group the daily precipitation time series and aggregate it by year to obtain the annual precipitation values.

3. (*15 points*) Download and read the Wisconsin Breast Cancer data set from the UCI Machine Learning data set repository using Pandas (*Hint*: You can see the Slide 16 from our PowerPoint presentation file from the Week6 folder).

a) Display the first five records of the data set. Select and show randomly a sample of size 6 from the data set.

b) Select randomly 2% of the data set, then display the selected samples (You can use the *random_state* parameter of the *sample* function).

c) Plot histogram of the data set. Use the *value_counts()* function to count the frequency of each attribute value.

4. (*10 points*) Why we use the Principal Component Analysis (PCA) algorithm?

5. (*10 points*) Apply the Principal Component Analysis (PCA) algorithm for the following array in Python.

[8, 21], [1, 5], [17, 3]

6. (*10 points*) Explain the flowchart of the K-means algorithm with your own sentences.

7. (*10 points*) Apply the K-means algorithm for the following data set in Python. Try the k values for k=2, k=3, and k=4.

```
Data = {'x': [14, 21, 46, 90, 87, 3, 22, 38, 55, 67],
        'y': [9, 33, 20, 51, 39, 45, 60, 83, 54, 72]
       }
```