Subject: 23CSE301

Lab Session: 02

Notes:

1. Please read the assignment notes carefully and comply to the guidelines provided.
2. Code should be checked into the GitHub. These details shall be provided in the Lab.
3. If you have not completed the prerequisite assignments, please complete them before the next lab session.

Coding Instructions:

1. The code should be modularized; The asked functionality should be available as a function. Please create multiple functions if needed. However, all functions should be present within a single code block, if you are using Jupyter or Colab notebooks.
2. There should be no print statement within the function. All print statements should be in the main program.
3. Please use proper naming of variables.
4. For lists, strings and matrices, you may use your input values as appropriate.
5. Please make inline documentation / comments as needed within the code blocks.

Main Section (Mandatory):

Refer to lecture portions on Linear Algebra.

A1. Please refer to the "***Purchase Data***" worksheet of ***Lab Session Data.xlsx***. Please load the data and segregate them into 2 matrices A & C (following the nomenclature of AX = C). Do the following activities.

- What is the dimensionality of the vector space for this data?
- How many vectors exist in this vector space?
- What is the rank of Matrix A?
- Using Pseudo-Inverse find the cost of each product available for sale.
  (Suggestion: If you use Python, you can use ***`numpy.linalg.pinv()`*** *function to get a pseudo-inverse.)*

A2. Mark all customers (in "***Purchase Data***" table) with payments above Rs. 200 as RICH and others as POOR. Develop a classifier model to categorize customers into RICH or POOR class based on purchase behavior.

A3. Please refer to the data present in "***IRCTC Stock Price***" data sheet of the above excel file. Do the following after loading the data to your programming platform.

- Calculate the mean and variance of the Price data present in column D.
  (Suggestion: if you use Python, you may use ***`statistics.mean()`*** *&* ***`statistics.variance()`*** *methods).*
- Select the price data for all Wednesdays and calculate the sample mean. Compare the mean with the population mean and note your observations.
- Select the price data for the month of Apr and calculate the sample mean. Compare the mean with the population mean and note your observations.
- From the Chg% (available in column I) find the probability of making a loss over the stock.
  (Suggestion: use lambda function to find negative values)

- Calculate the probability of making a profit on Wednesday.
- Calculate the conditional probability of making profit, given that today is Wednesday.
- Make a scatter plot of Chg% data against the day of the week

A4. **Data Exploration:** Load the data available in "**thyroid0387_UCI"** worksheet. Perform the following tasks:

- Study each attribute and associated values present. Identify the datatype (nominal etc.) for the attribute.
- For categorical attributes, identify the encoding scheme to be employed. (Guidance: employ label encoding for ordinal variables while One-Hot encoding may be employed for nominal variables).
- Study the data range for numeric variables.
- Study the presence of missing values in each attribute.
- Study presence of outliers in data.
- For numeric variables, calculate the mean and variance (or standard deviation).

A5. **Similarity Measure:** Take the first 2 observation vectors from the dataset. Consider only the attributes (direct or derived) with binary values for these vectors (ignore other attributes). Calculate the Jaccard Coefficient (JC) and Simple Matching Coefficient (SMC) between the document vectors. Use first vector for each document for this. Compare the values for JC and SMC and judge the appropriateness of each of them.

JC = $(f11) / (f01 + f10 + f11)$
SMC = $(f11 + f00) / (f00 + f01 + f10 + f11)$
$f11$= number of attributes where **the attribute carries value of 1 in both the vectors.**

A6. **Cosine Similarity Measure:** Now take the complete vectors for these two observations (including all the attributes). Calculate the Cosine similarity between the documents by using the second feature vector for each document.

If **A** and **B** are two document vectors, then

$$\cos(\mathbf{A}, \mathbf{B}) = \langle \mathbf{A}, \mathbf{B} \rangle / \|\mathbf{A}\| \|\mathbf{B}\|$$

$$< A, B >= \sum_{k=1}^{n} a_k * bk$$

$\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ are lengths of vectors A & B

A7. **Heatmap Plot:** Consider the first 20 observation vectors. Calculate the JC, SMC and COS between the pairs of vectors for these 20 vectors. Employ similar strategies for coefficient calculation as in A4 & A5. Employ a heatmap plot to visualize the similarities.

Suggestion to Python users →

```
import seaborn as sns
```

```
sns.heatmap(data, annot = True)
```

A8. **Data Imputation:** employ appropriate central tendencies to fill the missing values in the data variables. Employ following guidance.

- Mean may be used when the attribute is numeric with no outliers
- Median may be employed for attributes which are numeric and contain outliers
- Mode may be employed for categorical attributes

A9. **Data Normalization / Scaling:** from the data study, identify the attributes which may need normalization. Employ appropriate normalization techniques to create normalized set of data.

Optional Section:

O1. Create 2 separate square matrices from the purchase data matrix. Repeat experiments A2 & A3 with both these matrices. Do the X values obtained from the square matrices match to the one obtained from the whole purchase data matrix?

O2. Repeat experiments (A4 to A6) by taking observation samples from different regions of the data. You may try a random sampling to pick 20 vectors randomly form the set.

O3. Try the same exercise on data available on "**marketing_campaign**" worksheet.

Report Assignment:

1. Write your understanding of your project in the introduction section of the report.
2. Download at least 10 published papers (from IEEE Xplore, Springer, Elsevier or Science Direct) for your project. Study these papers use them for literature survey section of your report.
3. Using the learnings so far, design a system that could be used for customer / patient segmentation. Enrich your answer with:
    a. Flow diagram to depict the data flow. Example: input handling, preprocessing, similarity scoring, output.
    b. Architecture diagram for the system should be in methodology or system description section. Detail what happens in each block.
    c. define parameters to be used in the system; assign values for these parameters and justify them.

    Since this is a Design work, the solution should be provided in the methodology section of the IEEE format of report. The results may be taken from above experiments and discussed to conclude the paper.

4. In the "Results Analysis & Discussion" section of your report, write about the following topics:
    a. Discuss the importance of rank of an observation matrix in model building for classification.
    b. Discuss on regression (Ex: A2) and classification (Ex: A3) tasks. How would you differentiate between them.
    c. Observing the stock data provided, record your suggestions to build a system that may be able to predict the price and Change % into future.