Methodology

1) Purchase Data -Finding Items Costs

- I started with loading the Purchase data sheet of the Excel file.
- I have chosen Candies, Mangoes and Milk Packets as the input features (X) and Payment as the output (y).
- Then I computed X rank to verify the information of the matrix.
- The pseudo-inverse method was then applied to estimate the unit cost of each item (Candies, Mangoes, Milk Packets).

2) Customer Classification RICH or POOR

- I have added a new column that is called Class.
- In case the payment exceeds 200 rupees, the customer is declared as RICH.
- Otherwise, the customer is rated as POOR.

3) IRCTC Stock Data Mean, Variance, Probability

- I loaded the Stock Price sheet of the IRCTC.
- I took out the Price column and did the following calculations:

1. Mean
2. Variance

- To find mean and variance, I computed it in 2 ways:

1. Using NumPy
2. Using my own functions

- Then I compared their time which was taken with 10 runs.
- I also found:

1. Average price on Wednesdays
2. Average price in April
3. With the Chg% column, I determined:
4. Probability of loss (Chg% < 0)
5. Wednesday (Chg% > 0 on Wed) probability of making profits.

4) Visualization

- I plotted a scatter graph:
- X-axis: Day
- Y-axis: Chg%

- This assists in the knowledge of the stock change variation throughout the week days.

5) Thyroid Data set: Data Checking

- I loaded thyroid0387 UCI sheet.
- I checked:
- Types of data (categorical and numerical).
- The minimum and maximum values of numeric column.
- Absent values in the columns.
- Outliers using IQR method
- Mean and variance of all numeric features were also computed by me.

6) Similarity Measures

- The two initial records I chose are of the dataset on thyroid.
- On binary (0/1) columns, I calculated:
- Jaccard similarity
- Simple Matching Coefficient (SMC)
- On numeric columns, I worked out:
- Cosine similarity

7) Heatmap for Similarity

- I selected the first 20 records.
- I developed a Jaccard similarity matrix on them.
- I then used Seaborn to plot the similarity as a heatmap so that it could be easily visualized.

Summary

The dataset I used in this project was the lab Session Data.xlsx on the Excel . The initial section of the work is to the Purchase dataset, in which I had the amounts of items which were bought (candies, mangoes, and milk packets) as input variables and the amount of money as output. Using pseudo-inverse algorithm, I was able to determine the approximate cost contribution of every item, which serves as a naive linear regression model. This is also where I grouped customers into RICH or POOR groups according to the value that they pay, and this shows how raw numerical data may be translated into useful labels that are used in learning assignments.

In the second section, I conducted a statistical analysis of the IRCTC stock price data by determining valuable statistics such as the mean and the variance and compared the performance of NumPy with functions written in Python to learn about efficiency of computation. The additional findings I made included the average stock price on Wednesday, April, and probability of loss and profit based on the column of change percentage and illustrated with a scatter plot. Lastly, I accessed the thyroid disease dataset by doing data profiling (datatype check, ranges, missing values, and outliers), and used Jaccard coefficient, Simple Matching coefficient and Cosine similarity to calculate similarity between records. I also produced a Jaccard similarity heatmap in order to facilitate easy interpretation of similarities. In general, this project allowed me to train a full workflow that would include the stages of data loading, analysis, and visualization, and then similarity measurement, which is a solid foundation to use machine learning applications.