E/20/420 Wanasinghe J.K.
CO544 : Machine Learning and Data Mining
LAB 01 – Report

1) Insights and observations from each exercise.

Exercise 1: NumPy Advanced Operations:

Arrays of random integers were generated and boolean indexing was used to filter values greater than or equal to 50. Broadcasting with reshaped arrays demonstrated how operations scale across dimensions. A dot product of two numerical arrays was also generated.

Insights

- NumPy is ideal for numerical computing and vectorized operations. Boolean indexing and broadcasting are powerful for efficient data manipulation.
- In NumPy, broadcasting refers to the ability to perform arithmetic operations between arrays of different shapes in a way that avoids making unnecessary copies of data.

Exercise 2: Matplotlib Subplots

We used matplotlib to visualize sine and cosine waves side-by-side using subplots. Shared x-axes and appropriate labels made the plot intuitive and readable.

Insights

- Matplotlib is a very useful tool in python, which allows for the plots to be drawn which will make the working with data more efficient and productive.
- Visualizing functions or trends helps uncover data behavior. Subplots are effective for comparing patterns side-by-side.

Exercise 3: Pandas Cleaning & Preprocessing

Titanic dataset was worked upon and handled missing values using median and mode imputation. Duplicates were dropped. The 'Fare' column was rounded to integers and detected outliers using the Interquartile Range (IQR).

Insights

- Pandas is an open-source Python library that provides high-performance, easy-to-use data structures and data analysis tools.
    - Two core data structures:
        - Series – a one-dimensional labeled array.
        - DataFrame – a two-dimensional labeled table, like a spreadsheet.
    - Useful for:
        - Data manipulation: Reading, writing, filtering, and transforming tabular data (like CSV or Excel).
        - Data cleaning: Handling missing data, duplicates, outliers, and formatting issues.
        - Data analysis: Summarizing, grouping, pivoting, and aggregating data.
        - Integration: Works well with other libraries like NumPy, Matplotlib, and scikit-learn.
- Real-world datasets often contain missing values and outliers. Imputation and IQR-based detection help in preprocessing and improving model input quality.

Exercise 4: Pandas Essentials

Creating and inspecting Series and DataFrames, indexing using loc and iloc, sorting, and dropping columns. Missing data handling was explored with dropna() and fillna(), and performed basic Excel I/O.

Insights

- Pandas offers a versatile toolkit for data analysis and manipulation, including handling missing data and exporting results.
- A Series is a one-dimensional labeled array.
- A DataFrame is a 2D table with rows and columns, the most used structure in Pandas.

Exercise 5: Loading Open Dataset from UCI Repository

The wine dataset was loaded and grouped by wine class to compute the mean of each chemical property. The analysis revealed that different wine classes have distinguishable feature profiles.

- loc[] selects data by label/index name.
- iloc[] selects data by position/index number.
- Sorting helps in ordering data based on values.
- Dropping columns is useful to remove irrelevant or redundant information.
- df_nan.dropna() : Drops all rows that have at least one NaN value. So it keeps only rows where all values are present.
- df_nan.fillna(0) : Replaces all NaN values with 0.
- Pandas can read and write Excel files using to_excel() and read_excel().

Exercise 6: Iris Dataset Classification with scikit-learn

Iris dataset was used and trained a Logistic Regression model. The dataset was split into training and test sets using train_test_split(), and evaluated predictions using a classification report.

```
Model Accuracy: 1.0

Classification Report:
              precision    recall  f1-score   support

      setosa       1.00      1.00      1.00        19
  versicolor       1.00      1.00      1.00        13
   virginica       1.00      1.00      1.00        13

    accuracy                           1.00        45
   macro avg       1.00      1.00      1.00        45
weighted avg       1.00      1.00      1.00        45
```

- The model classified all samples correctly, showing how clean, well-structured data supports high accuracy.

2) Strategies for handling missing data and outliers.

Missing Data:

- Median imputation was used for numerical columns like 'Age'. It is less sensitive to outliers than mean.
- Mode imputation for categorical columns like 'Embarked' ensures valid category values.
- dropna() is applied when missing values are minimal and need to be removed completely.

Outliers:

- Outliers were identified using the IQR method in the Titanic dataset's Fare_int column.
- Managing outliers helps prevent skewing of statistical analysis and improves model stability.

3) Interpretation of pivot/group-by results

GroupBy (Wine Dataset):

- Grouping the wine data by 'Class' provided a clear view of how feature means vary per class.
- For example, Class 1 had significantly higher alcohol and flavonoids.

Pivot Tables:

- While not used in this lab, pivot tables allow more detailed summaries by multiple categorical keys and are useful in business analytics.

4) Reflection on Model Performance Metrics

From Exercise 6:

- Accuracy: The model achieved 1.0 (100%), indicating no misclassifications.
- Precision: 1.0 for each class—no false positives.
- Recall: 1.0—no missed predictions.
- F1-score: 1.0—perfect balance of precision and recall.

Reflection:

- This ideal performance reflects both the simplicity and cleanliness of the Iris dataset.
- In real-world scenarios, performance would likely decrease due to noise, imbalance, or complexity.