CO544 – Machine Learning and Data Mining

Lab 1 Report: Python Data Science Toolbox
Registration Number: E/20/280

Insights and Observations

Exercise 1: NumPy Advanced Operations
This exercise demonstrated advanced array operations, including random number generation, boolean indexing, broadcasting, and computing the dot product.

- Broadcasting simplifies element-wise operations between arrays of different shapes. So we do nt have to use loops to do it.

- Calculated the dot product between two 1D arrays, confirming it as the sum of products of corresponding elements.

Exercise 2: Matplotlib Subplots
We created subplots for sine and cosine functions, to illustrate Matplotlib's multi panel plottings.

- Using plt.subplots() improves the readability and structure of visualizations.

- Sharing the x-axis (sharex=True) makes comparative analysis intuitive.

Exercise 3: Pandas Cleaning & Preprocessing
Here it is focused on data cleaning tasks on the Titanic dataset, such as handling missing values, removing duplicates, and detecting outliers.

- Using .fillna() to fill missing data and .drop_duplicates() to drop duplicate data, ensures data quality and consistency.

- The IQR method effectively identifies extreme values in skewed distributions

Exercise 4: Pandas Essentials
Learned to create and manipulate Series and DataFrames, as well as perform indexing and sorting, manage missing data, and work with Excel I/O.

- Read from files and write to files easily.
- .loc and .iloc allow precise data access.

- Works well to fill in gaps with .fillna() when a value cannot simply be dropped without additional actions taken on the data.

Exercise 5: Loading Open Dataset from UCI Repository
The wine dataset is classified by the different types of wines available and grouped by classes to showcase the various statistical figure differences present.

- Grouping by class reveals clear statistical differences among wine types.

- For feature analysis and pattern detection, aggregating with .mean() is beneficial.

Exercise 6: scikit-learn Iris Dataset
We trained a Logistic Regression model on the Iris dataset, classifying it with certain metrics later on.

- With classification_report, the insights derived from the model can be evaluated with ease.

- When the dataset has clear linear divides such as in the case of the Iris, logistic regression gives accurate results.

Strategies for Handling Missing Data and Outliers

Missing Data:

- Median Imputation: Used for the 'Age' column in the Titanic dataset as it is less sensitive to outliers than the mean.
- Mode Imputation: Used for categorical data like 'Embarked'.
- Alternative: If many values are missing, consider dropping columns or using more advanced techniques like KNN or regression imputation.

Outliers:

- IQR Method: Values outside Q1 − 1.5×IQR and Q3 + 1.5×IQR were considered outliers (the left and right corners from the data curve).
- Strategy: Depending on the context, outliers can be removed, capped, or transformed (e.g., log scaling).
- Note: Outlier detection is crucial before applying machine learning models, especially distance-based models like KNN.

Interpretation of Pivot/Group-by Results

Using the groupby function on the Wine dataset grouped data by wine class and calculated mean feature values.

- Features such as ‘Flavanoids’ and ‘Alcohol’ vary significantly across wine classes, indicating they can be useful predictors.
- **Use Case:** Group-by operations are fundamental in exploratory data analysis (EDA) to understand class-wise distributions and correlations.

Reflection on Model Performance Metrics

The classification report generated for the Iris dataset includes:

- **Precision**: True positives over predicted positives.
- **Recall**: True positives over actual positives.
- **F1-Score:** Harmonic mean of precision and recall.
- **Accuracy**: Overall correct predictions.

Observations:

- The logistic regression model achieved high precision, recall, and F1-scores across all classes.
- Indicates balanced class distribution and model effectiveness.
- **Limitation: F**or multi-class imbalanced datasets, more nuanced evaluation (e.g., confusion matrix, macro/micro averages) may be needed.