CO544: Machine Learning and Data Mining

Lab 1: Python Data Science Toolbox

E/20/197 Kawya A.H.D.

Exercise 1: NumPy Advanced Operations

Exercise	Observation	Insight
Generating Random Values	Unless we set a seed, each run gives different values.	np.random.randint() is a key tool in data science workflows. It is used for generating test inputs.
Boolean Indexing to Filter Values ≥ 50	This is useful for data filtering, conditional selection and data cleaning.	Boolean indexing is used to select elements based on conditions without loops.
Demonstrate broadcasting	NumPy automatically expanded the dimensions of the smaller array to match the larger array.	Broadcasting is a core concept for vectorized computations.
Compute dot product of two arrays	Dot product represents the sum of element-wise multiplication.	The dot product is fundamental in machine learning. This is how inputs are combined with weights.

Exercise 2: Matplotlib Subplots

Exercise	Observation	Insight
Prepare data for sine and cosine functions	X values have generated, and sine and cosine values have calculated for each x-value.	This simulates periodic wave behavior.
Create subplots	Two side_by_side subplots have been created. Titles and labels clarify the purpose of each subplot.	Using subplots helps in comparing multiple datasets or functions within a single figure. Sharing axes improves readability and removes redundant visual elements.

Exercise 3: Pandas Cleaning & Preprocessing

Exercise	Observation	Insight
Load Titanic dataset	We have loaded a real-world dataset containing passenger data from the Titanic voyage.	Always start by understanding data before any modeling.
Imputing Missing Values	For 'Age- Numerical', missing values are filled with the median. It is less sensitive to outliers than the mean. For 'Embarked- Categorical', filled with the mode, which is the most frequently occurring part	Imputation preserves data integrity without deleting rows, which is especially important in datasets with limited samples.

Dropping Duplicates	Any duplicate rows have been deleted to ensure each record is unique.	These duplicate rows can be bias analysis in model training. It leads to model overfitting.
Detecting Outliers in Fare	Values which are significantly higher and lower than most fares are called outliers	Outlier detection is crucial for model performance.

Exercise 4: Pandas Essentials

Exercise	Observation	Insight
Create and inspect Series	Two Pandas series have been created with default integer index & manually defined indexed labels.	However we have defined the series, they can be accessed by default or manually defined labels.
Build Data Frame and summarize	Small Structured Data Frame has been created, and it represents a simple table with names and their scores. A random Numerical Data Set has been created, and it contains 100 rows and 3 columns filled with random numbers from a normal distribution.	Manually constructed data frames are suitable for testing, prototyping or handling small datasets. Randomly generated data helps test statistical methods, visualize distributions, or benchmark machine learning algorithms.
Indexing, sorting, and dropping	loc is for label access, not position. iloc is position-based indexing. . sort_values is for sorting the entire data frame. .drop is for dropping a column.	Using these functions allows flexible data success by label or index, while other sorting and dropping helps revel patterns by ordering rows and simplifies the data frames by removing unnecessary columns or rows.
Handle missing data	These functions are used to handle missed values.	These functions remove incomplete data to ensure clean records & missing values with default but possibly introducing bias.
Excel I/O	The tail() function lets us quickly inspect if the last few records are intact or contain anomalies. Saving to a new Excel file ensures we don't overwrite the original file.	Excel I/O in Pandas allows easy integration of tabular data workflows.

Exercise 5: Loading Open Dataset from UCI Repository

Exercise	Observation	Insight
Load Wine dataset	The data is loaded directly from the UCI Machine Learning Repository. The dataset has no header row, so column names are manually assigned.	Loading and inspecting a labeled dataset like the Wine dataset is essential for understanding its structure.
Group by class	We can compare feature averages across wine class. It provides a numerical summary.	Grouping by class and calculating feature means is a fundamental step in data analysis.

Exercise 6: scikit-learn Iris Dataset (Extended)

Exercise	Observation	Insight
Load and preview Iris	The dataset has 150 samples and 4 features per sample.	Loading and structuring the Iris dataset into a labeled Data Frame allows for easy analysis.
Train/test split	The dataset is divided into training and test sets to evaluate model performance fairly.	Splitting the data into training and test sets ensures that model evaluation is fair and unbiased.
Model training and evaluation	The classification report gives a detailed performance breakdown for each Iris class.	Logistic Regression, when applied to a well-balanced and linearly separable dataset-like Iris. It offers high interpretability and reliable performance.