**Faculty of Computing**



**Lab 02: Introduction to Pandas and Scikit-Learn**

**CS471 Machine Learning**

**BESE – 13A**

**03rd February 2025**

**Lab Engineer: Mr. Junaid Sajid**

**Instructor: Dr. Hashir Moheed Kiani**

### **Lab 02: Introduction to Pandas and Scikit-Learn**

#### **Lab Objectives:**

1. Develop hands-on experience in manipulating datasets using Python libraries such as Pandas and Scikit-Learn.
2. Perform exploratory data analysis (EDA) to identify patterns, trends, and insights in datasets.
3. Visualize data effectively to support analysis and reporting.

#### **End Goal:**

Students will be able to manipulate datasets, perform exploratory data analysis, and create insightful visualizations using Python.

### **Datasets:**

1. **Titanic Dataset**:
   * Description: Data on passengers of the Titanic, including survival information.
   * Source: Downloadable from Kaggle ([Titanic Dataset](https://www.kaggle.com/datasets/yasserh/titanic-dataset?select=Titanic-Dataset.csv)).
2. **NYC Airbnb Dataset**:
   * Description: Listings data with features such as price, location, and availability.
   * Source: Downloadable from Kaggle ([New York City Airbnb Open Data](https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data)).

### **Pandas:**

* Importing Pandas:

import pandas as pd

* Creating a Series:

data = [10, 20, 30, 40]

series = pd.Series(data, index=['A', 'B', 'C', 'D'])

print(series)

* Creating a DataFrame:

data = {'Name': ['Alice', 'Bob', 'Charlie'],

'Age': [25, 30, 35],

'City': ['New York', 'London', 'Paris']}

df = pd.DataFrame(data)

print(df)

* Loading a dataset:

df = pd.read\_csv('data.csv')

print(df.head())

* Basic operations:

print(df.info())

print(df.describe())

print(df['Age'].mean())

### **scikit-learn:**

* Loading a sample dataset:

from sklearn.datasets import load\_iris

data = load\_iris()

print(data.keys())

print(data['feature\_names'])

print(data['target\_names'])

* Splitting the dataset:

from sklearn.model\_selection import train\_test\_split

X = data['data']

y = data['target']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

print(X\_train.shape, X\_test.shape)

* Training a machine learning model:

from sklearn.ensemble import RandomForestClassifier

model = RandomForestClassifier()

model.fit(X\_train, y\_train)

* Making predictions and evaluating:

from sklearn.metrics import accuracy\_score

y\_pred = model.predict(X\_test)

accuracy = accuracy\_score(y\_test, y\_pred)

print(f'Accuracy: {accuracy \* 100:.2f}%')

### **Matplot Lib:**

* Importing Matplotlib:

import matplotlib.pyplot as plt

* Plotting a basic graph:

x = [1, 2, 3, 4, 5]

y = [10, 20, 30, 40, 50]

plt.plot(x, y, marker='o')

plt.title("Basic Line Plot")

plt.xlabel("X-axis")

plt.ylabel("Y-axis")

plt.show()

### **Exploratory Data Analysis on Iris Dataset:**

* **Load the Iris Dataset**

from sklearn.datasets import load\_iris

import pandas as pd

# Load the dataset

iris\_sklearn = load\_iris()

iris\_df = pd.DataFrame(data=iris\_sklearn.data, columns=iris\_sklearn.feature\_names)

iris\_df['species'] = iris\_sklearn.target

iris\_df['species'] = iris\_df['species'].replace({0: 'setosa', 1: 'versicolor', 2: 'virginica'})

# Display the first few rows

print(iris\_df.head())

* **Data Visualization with Matplotlib**
* **Scatter Plot**:

plt.scatter(iris\_df['sepal length (cm)'], iris\_df['sepal width (cm)'], c='blue', label='Sepal')

plt.scatter(iris\_df['petal length (cm)'], iris\_df['petal width (cm)'], c='red', label='Petal')

plt.title("Scatter Plot of Sepal and Petal Dimensions")

plt.xlabel("Length (cm)")

plt.ylabel("Width (cm)")

plt.legend()

plt.show()

* **Histogram**:

plt.hist(iris\_df['sepal length (cm)'], bins=10, color='green', alpha=0.7)

plt.title("Histogram of Sepal Length")

plt.xlabel("Sepal Length (cm)")

plt.ylabel("Frequency")

plt.show()

* **Bar Plot**:

species\_counts = iris\_df['species'].value\_counts()

species\_counts.plot(kind='bar', color=['blue', 'orange', 'green'])

plt.title("Species Count")

plt.xlabel("Species")

plt.ylabel("Count")

plt.show()

**Lab Tasks:**

#### **Task 1: Data Loading and Exploration**

1. Load the **Titanic Dataset**
2. Inspect the dataset:
   * Display column names, data types, and summary statistics using Pandas.
   * Identify missing values and discuss their implications.

#### **Task 2: Data Cleaning and Manipulation**

1. Handle missing values:
   * Fill missing values in the Age column with the median age.
   * Drop rows with zero values in the Fare column.
2. Create new features:
   * Add a family\_size feature by combining sibsp (siblings/spouses) and parch (parents/children).
   * Create a travel\_alone feature indicating whether the passenger traveled alone.
3. Filter data:
   * Select passengers aged between 18 and 50 and save the subset as a new DataFrame.

#### **Task 3: Exploratory Data Analysis**

1. Analyze categorical variables:
   * Visualize the survival rate by gender and passenger
2. Analyze numerical variables:
   * Plot a histogram for the Age column and overlay the median.
   * Create a box plot of Fare across different passenger classes.
3. Detect correlations:
   * Compute and visualize the correlation matrix using heatmap.

#### **Task 4: Advanced Data Visualization**

1. Plot geospatial data using the **NYC Airbnb Dataset**:
   * Visualize the geographic distribution of listings using a scatter plot of latitude and longitude.
   * Create a bar plot of average price per neighborhood.
2. Identify outliers in the price column using a box plot.

### **Lab Summary and Deliverables:**

1. **Deliverables**: Python code, visualizations, and a report summarizing and one report on self reflections:
   * Data cleaning steps.
   * Insights derived from EDA and visualizations.
   * Observations regarding patterns and anomalies in the datasets.
2. **Learning Outcomes**:
   * Gained practical skills in data manipulation using Pandas and NumPy.
   * Conducted exploratory data analysis to extract insights.
   * Applied effective visualization techniques to support data storytelling.

### **Self-Reflection Report Guidelines**

#### **Report Template (To Be Included in Notebook):**

Students will answer the following questions as part of their notebook submission:

1. **Understanding of Concepts**:
   * Summarize the key concepts of data manipulation and EDA you applied in this lab.
   * What new skills or knowledge did you gain?
2. *Example*:  
   *"Through this lab, I learned how to handle missing values effectively using Pandas and how to create new features to extract meaningful insights. The correlation matrix visualization was particularly insightful for identifying relationships among variables."*
3. **Challenges Faced**:
   * Describe any difficulties encountered during the lab tasks.
   * How did you resolve them?
4. *Example*:  
   *"I initially struggled with visualizing correlations using a heatmap. After referring to the Pandas and Seaborn documentation, I realized I needed to preprocess the data to exclude non-numerical columns."*