HYDERABAD INSTITUTE OF ARTS, SCIENCE, AND TECHNOLOGY

Artificial Intelligence Lab -2
Instructor: Miss Ayesha Eman
Date: 12/09/2025

Lab Title: Data Handling & Preprocessing for Artificial Intelligence.

Lab Objectives

By the end of this lab, students should be able to:

- Import and explore datasets using Pandas.
- Clean and preprocess data for AI tasks.
- Handle missing values effectively.
- Perform descriptive statistics.
- Visualize data distributions and relationships.

Lab Topics

1. Reading datasets (CSV, Excel) with Pandas.
2. Handling missing data (`dropna`, `fillna`).
3. Sorting, filtering, and grouping data.
4. Descriptive statistics (mean, median, std, etc.).
5. Visualizing data with histograms, bar charts, and scatterplots.

Lab Exercises

Exercise 1: Load a Dataset

Goal: Read a dataset and view its structure.

```
import pandas as pd

# Load dataset
students = pd.read_csv("students.csv")
print(students.head())
print(students.info())
```

Exercise 2: Handle Missing Data

Goal: Explore and clean missing values.

```
# Check missing values
print(students.isnull().sum())

# Fill missing marks with average
students['Marks'] = students['Marks'].fillna(students['Marks'].mean())

# Drop rows with missing names
students = students.dropna(subset=['Name'])
print(students)
```

Exercise 3: Descriptive Statistics

Goal: Summarize dataset features.

```
# Summary statistics
print(students.describe())

# Find top scorer
print("Top Scorer:", students.loc[students['Marks'].idxmax()])

# Find average marks
print("Average Marks:", students['Marks'].mean())
```

Exercise 4: Data Filtering & Grouping

Goal: Extract meaningful subsets.

```
# Filter students with marks > 80
high_scorers = students[students['Marks'] > 80]
print(high_scorers)

# Group by Grade
print(students.groupby('Grade')['Marks'].mean())
```

Exercise 5: Visualization

Goal: Visualize distributions and relationships.

```
import matplotlib.pyplot as plt

# Histogram
students['Marks'].hist(bins=10, color='skyblue')
plt.title("Distribution of Marks")
plt.xlabel("Marks")
plt.ylabel("Frequency")
plt.show()
```

```
# Scatterplot
plt.scatter(students['Hours_Studied'], students['Marks'], color='green')
plt.title("Hours Studied vs Marks")
plt.xlabel("Hours Studied")
plt.ylabel("Marks")
plt.show()

# Bar chart
students['Result'].value_counts().plot(kind='bar', color='orange')
plt.title("Pass vs Fail Count")
plt.xlabel("Result")
plt.ylabel("Count")
plt.show()
```

Discussion Points

- Why is data preprocessing considered the most time-consuming step in AI?
- How do missing values affect AI model performance?
- When should we use mean vs median for replacing missing values?
- Why is visualization important before applying AI models?
- What risks exist if we skip proper preprocessing?

Assessment Questions

1. Import a dataset of your choice and clean missing values using at least two different methods.
2. Calculate and explain the standard deviation of a numeric column in your dataset.
3. Create a scatterplot showing the relationship between two continuous features.