

Chapter 9: Data Analysis using Python

Introduction

Data is at the core of Artificial Intelligence. However, raw data alone is not useful until it is cleaned, structured, and analyzed. This chapter introduces **data analysis using Python**, focusing on libraries such as **Pandas**, **NumPy**, and **Matplotlib**. You'll learn how to import data, process it, clean it, and draw meaningful insights. These skills are essential for every AI developer or data scientist.

9.1 Introduction to Data Analysis

Data analysis is the process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.

Types of Data Analysis

- **Descriptive:** Summarizes past data.
 - **Diagnostic:** Explains why something happened.
 - **Predictive:** Predicts future outcomes.
 - **Prescriptive:** Suggests actions.
-

9.2 Python Libraries for Data Analysis

9.2.1 NumPy (Numerical Python)

- Core library for scientific computing in Python.
- Provides a high-performance multidimensional array object.

```
import numpy as np
```

```
arr = np.array([1, 2, 3, 4])  
print(arr.mean()) # Output: 2.5
```

9.2.2 Pandas (Panel Data)

- Built on NumPy; used for data manipulation and analysis.
- Provides two key data structures:
 - Series – 1D labeled array.
 - DataFrame – 2D labeled data structure.

```
import pandas as pd
```

```
data = {'Name': ['Alice', 'Bob'], 'Age': [24, 27]}  
df = pd.DataFrame(data)  
print(df)
```

9.2.3 Matplotlib

- Used for data visualization.
- Plots like bar charts, line graphs, histograms, etc.

```
import matplotlib.pyplot as plt
```

```
x = [1, 2, 3]  
y = [2, 4, 1]
```

```
plt.plot(x, y)  
plt.title("Line Graph")  
plt.xlabel("X Axis")  
plt.ylabel("Y Axis")  
plt.show()
```

9.3 Loading and Exploring Datasets

9.3.1 Reading Data from CSV

```
df = pd.read_csv("students.csv")  
print(df.head())
```

9.3.2 Understanding Dataset Properties

- `df.head()`: First 5 rows
- `df.tail()`: Last 5 rows
- `df.shape`: Rows and columns
- `df.columns`: Column names
- `df.info()`: Data types and nulls
- `df.describe()`: Summary stats

9.4 Data Cleaning

Data cleaning is crucial for accurate analysis. Common tasks include:

9.4.1 Handling Missing Values

```
df.isnull().sum()  
df.fillna(0, inplace=True)
```

9.4.2 Removing Duplicates

```
df.drop_duplicates(inplace=True)
```

9.4.3 Changing Data Types

```
df['Age'] = df['Age'].astype(int)
```

9.5 Data Manipulation

9.5.1 Selecting Columns and Rows

```
df['Name']          # Single column  
df[['Name', 'Age']] # Multiple columns  
df.iloc[0]         # First row
```

9.5.2 Filtering Data

```
df[df['Age'] > 25]    # Rows where Age > 25
```

9.5.3 Sorting Data

```
df.sort_values('Age', ascending=False)
```

9.6 Data Aggregation

9.6.1 Grouping Data

```
df.groupby('Gender')['Marks'].mean()
```

9.6.2 Pivot Tables

```
df.pivot_table(index='Gender', values='Marks', aggfunc='mean')
```

9.7 Data Visualization with Matplotlib

9.7.1 Line Chart

```
plt.plot(df['Age'])  
plt.title("Age Plot")  
plt.show()
```

9.7.2 Bar Chart

```
plt.bar(df['Name'], df['Marks'])  
plt.title("Student Marks")  
plt.show()
```

9.7.3 Histogram

```
plt.hist(df['Marks'], bins=5)
plt.title("Marks Distribution")
plt.show()
```

9.7.4 Pie Chart

```
df['Gender'].value_counts().plot.pie(autopct='%1.1f%%')
plt.title("Gender Distribution")
plt.show()
```

9.8 Mini Project: Analyzing Student Data

Objective: Analyze a CSV file containing student names, gender, age, and marks.

Steps:

1. Load the data.
2. Clean it (handle missing values).
3. Find average marks by gender.
4. Visualize the result using a bar chart.
5. Save cleaned data.

```
import pandas as pd
import matplotlib.pyplot as plt
```

```
df = pd.read_csv("student_data.csv")
df.fillna(df.mean(numeric_only=True), inplace=True)
avg_marks = df.groupby("Gender")["Marks"].mean()
```

```
avg_marks.plot(kind="bar", color=['skyblue', 'lightgreen'])
plt.title("Average Marks by Gender")
plt.ylabel("Marks")
plt.show()
```

```
df.to_csv("student_data_cleaned.csv", index=False)
```

Summary

- **Python libraries** like Pandas, NumPy, and Matplotlib are essential tools for data analysis.
- You learned how to **load**, **clean**, **manipulate**, and **visualize** data.
- Practical knowledge of data analysis builds the foundation for **Machine Learning** and **Artificial Intelligence**.
- This chapter sets the stage for using real-world datasets in AI systems and preparing them for intelligent analysis and predictions.
