Assignment – 1 (Chapter1-2)

Part A – Theory Questions

Q1.
(a) From the dataset of a hospital (age, blood pressure, disease type, treatment success: Yes/No), mark which features are numerical and which are categorical.
(b) Why might using a categorical column (like disease type) as a number cause wrong conclusions?
(c) Which ML algorithm would naturally handle categorical data well? Which one fits numerical better?

Q2.
(a) A company records the salaries of employees. If one CEO earns 1 crore while most others earn below 50,000, would you use mean or median to describe central tendency? Why?
(b) Define variance in your own words with an example.
(c) How does standard deviation give more intuition about data spread compared to variance?

Q3.
(a) A student memorizes last year's question paper and scores well in mock tests but fails in the real exam. How is this similar to ML overfitting?
(b) List two strategies to improve generalization in models.
(c) Explain with an analogy why underfitting is as harmful as overfitting.

Q4.
(a) A survey has missing "income" data for 10% of participants. Suggest two ways to address this issue.
(b) In a dataset of house prices, one record shows a price 100× higher than others. What is this called? How can it affect predictions?
(c) Which is worse for a model — systematic missing values or random missing values? Explain briefly.

Q5.
(a) A histogram of exam scores shows a long left tail. What does it suggest about student performance?
(b) A boxplot shows several dots above the whisker — what does that mean?
(c) Why might scatter plots help you guess relationships before fitting a model?

Q6.
(a) In predicting house prices, which would be the features and which is the target?
(b) If one feature is strongly correlated with the target, how does it help the model?
(c) Can we ever have multiple targets? Give an example.

Q7.
(a) Classify these scenarios: grouping songs by similarity, predicting next month's rainfall, teaching a robot to balance.
(b) Why is supervised learning more common in industry?
(c) Give one benefit of unsupervised learning despite being harder to evaluate.

Q8.
(a) A recruitment dataset contains more male than female applicants. What type of data bias is this?
(b) How can such bias affect ML model predictions?
(c) Suggest one way to reduce the impact of data bias.

Q9.
(a) If you have only 50 data points, would you use a very complex model or a simple one? Why?
(b) What danger arises if the model has too many parameters compared to the dataset size?
(c) Why do simpler models often generalize better on small datasets?

Part B – Lab Questions

Q1.

(a) Create a 1D NumPy array with values 1 to 20.
(b) Extract all prime numbers from it.
(c) Compute the mean and variance of the extracted primes.

Q2.

(a) Create a 4×4 NumPy array with numbers 1 to 16.
(b) Extract the 2×2 bottom-left sub-matrix.
(c) Compute the determinant of the sub-matrix.

Q3.

Create a DataFrame with 5 students and marks in 3 subjects.

(b) Add a column for total and average marks.

(c) Identify the topper and print their name with average.

Q4.

(a) Simulate 1000 coin tosses using NumPy (1=Head, 0=Tail).
(b) Count frequency of heads and tails.
(c) Estimate probability of heads. Is it close to 0.5? Why/Why not?

Q5.

(a) Create a DataFrame of employees with columns: ID, Name, Salary.
(b) Add a Bonus column = 10% of Salary.
(c) Display employees with salary above average.

Q6.

(a) Create a 3×3 NumPy array with values 1 to 9.
(b) Find its transpose and inverse.
(c) Verify that $A \times A^{-1} \approx I$.

Q7.

(a) Generate random daily temperatures (30 values, range 20–40°C).
(b) Find the hottest and coldest day.
(c) Compute mean, median, and standard deviation of temperatures.

Q8.

(a) Create a Pandas Series of marks for 8 students.
(b) Replace all marks below 40 with "Fail".
(c) Count how many students passed.

Q9.

(a) Generate 500 random integers between 1 and 6 (simulate dice rolls).
(b) Count how many times each face appears.
(c) Compute relative frequencies and compare with theoretical 1/6.

Q10.

(a) Create a DataFrame with 6 products (Name, Quantity, Price).
(b) Add a column "Total = Quantity × Price".
(c) Find which product generated maximum sales revenue.

Q11.

(a) Create a DataFrame with some missing values (NaN).
(b) Fill missing values with column mean.
(c) Drop rows where more than 1 value is missing.

Q12.

(a) Create a NumPy array with integers from 1 to 30.
(b) Reshape it into a 5×6 matrix.
(c) Extract all even numbers from the matrix and compute their average.

Q13.

(a) Create a DataFrame of 6 students with columns: Name, Age, Marks.
(b) Select only students who scored above the overall average marks.
(c) Display names of students younger than 20 whose marks are above 60.

Q14.

(a) Create a Pandas DataFrame with 5 employees having columns: Name, Tasks_Completed, Hours_Worked.
(b) Add a new column Efficiency = Tasks_Completed / Hours_Worked.
(c) Identify the employee with the **highest efficiency** and print their name and value.

Part C – Case Study

Case Study 1 – Student Performance Analytics

You are given data of 100 students containing their marks in **Math, Science, and English**.
(a) Using Pandas, calculate the average marks for each student and identify the top 5 performers.
(b) Compute subject-wise average and standard deviation; comment which subject shows the highest variation.
(c) Draw a boxplot for each subject and comment on outliers.

Case Study 2 – Sales Data Exploration

A company records sales transactions in a dataset with columns: **Product, Quantity, Price, Region**.
(a) Add a new column Total = Quantity × Price and compute overall revenue.
(b) Group data by region and identify which region contributes the most sales.
(c) Plot a histogram of total sales per product and discuss which products are underperforming.

Case Study 3 – Predicting Fuel Efficiency (Auto MPG Dataset)

You are analyzing the **Auto MPG dataset (UCI Repository)** with features: mpg, cylinders, displacement, horsepower, weight, acceleration, model year, origin.
(a) Compute the correlation of mpg with each numeric feature and discuss which factors influence fuel efficiency most.
(b) Plot a scatter plot of weight vs mpg; describe the trend you observe.