Assignment

Dataset Overview

Domain: Healthcare & Medical Diagnosis

Focus: Heart disease prediction and clinical risk assessment

Content: Medical attributes of patients including age, gender, chest pain type, cholesterol levels, blood pressure, maximum heart rate, ECG results, exercise-induced angina, ST depression, and other clinical measurements.

Primary Use: Healthcare professionals and data analysts utilize this data to identify risk patterns and early symptoms of heart disease, enabling early diagnosis and preventive healthcare strategies.

Business Objectives

Business Goal	Description
Reduce Mortality	Identify high-risk patients early to provide preventive treatment and interventions
Cost Reduction	Lower hospital readmission rates and reduce expensive emergency operations
Treatment Planning	Determine which clinical factors contribute most significantly to heart attacks
Patient Segmentation	Group patients for customized disease management programs
Risk Scoring Model	Predict heart disease likelihood for new patients using machine learning algorithms

Data Analysis Assignment

Instructions: Use only this dataset to solve all questions. All questions are scenario-oriented and require combining NumPy + Pandas concepts.

1. Find the total number of patients and how many of them have heart disease
2. Calculate the average age of patients using NumPy and compare it with Pandas .mean() result. Are both same?
3. Which gender (Male/Female) has higher average cholesterol levels? (Use groupby + mean)
4. Use NumPy to extract all patients above 60 years of age and count how many of them have heart disease
5. Which chest pain type is most commonly observed? Also compute its percentage in the dataset
6. Find the patient with the maximum resting blood pressure and display their complete record
7. Using Pandas + NumPy, detect outliers in the cholesterol column (IQR method). How many outliers exist?
8. Group patients by age category (e.g., <40, 40–55, 55+) and calculate the heart disease rate in each group
9. Calculate correlation among features (cholesterol, blood pressure, max heart rate, age) and identify the strongest relationship
10. Using NumPy conditional indexing, extract all male patients who have heart disease and cholesterol > 250
11. Find the top 3 most influential risk factors using .corr() sorted highest to lowest
12. Group by chest pain type and compute the average max_heart_rate for each type
13. Using NumPy where(), create a new column "cholesterol_status":
    a. HIGH if cholesterol > 240 else NORMAL
14. Among patients with HIGH cholesterol, what percentage have heart disease?
15. Group by gender and calculate mean blood pressure, cholesterol, and max heart rate (all in one result)
16. Find the age range (min–max) of patients using NumPy and match with Pandas
17. Check whether the dataset has missing values. If yes, replace numerical NaN with column median using NumPy
18. Using value_counts(), identify the combination of (gender + chest pain type) that occurs most frequently
19. Segment patients into BMI-based risk (use NumPy cut if BMI column exists) and compute avg heart disease rate per segment
20. Create a NumPy array of cholesterol values and find percentile values (25th, 50th, 75th). Interpret the results