Mini Project Documentation

Heart Disease Risk Analysis & Predictive Insights

Technology Stack: Python, NumPy, Pandas, Matplotlib
Target Audience: Data Engineering Corporate Batch

Project Overview

Objective

To analyze patient clinical data and identify key factors influencing heart disease using Exploratory Data Analysis (EDA) and statistical techniques.

The project focuses on:

- Business-oriented data interpretation
- Risk factor identification
- Visualization-driven insights
- Data-driven decision support

Business Context

Cardiovascular disease is one of the leading causes of mortality globally. Healthcare institutions require:

- Early risk identification
- Pattern recognition across demographics
- Data-driven clinical decision support

This project simulates a real-world healthcare analytics use case where we derive insights from patient clinical parameters.

Dataset Description

Feature	Description
age	Age of patient
sex	Gender (1 = Male, 0 = Female)
cp	Chest pain type
trestbps	Resting blood pressure
chol	Cholesterol level
fbs	Fasting blood sugar
restecg	Resting ECG results
thalach	Maximum heart rate
exang	Exercise induced angina
oldpeak	ST depression
slope	Slope of ST segment
ca	Number of major vessels
thal	Thalassemia
target	Disease presence (1 = Yes, 0 = No)

Project Architecture

Data Collection
↓
Data Cleaning (Pandas)

↓
Exploratory Data Analysis
↓
Statistical Analysis (NumPy)
↓
Visualization (Matplotlib)
↓
Insight Generation
↓
Business Recommendation

Project Phases

Phase 1: Data Understanding & Cleaning

Tasks:

- Load dataset
- Check null values
- Validate data types
- Detect outliers
- Perform summary statistics

Deliverables:

- Cleaned dataset
- Data profiling report

Phase 2: Exploratory Data Analysis (EDA)

A. Demographic Analysis

- Age distribution
- Gender-based disease comparison

B. Clinical Parameter Analysis

- Cholesterol vs Disease
- Blood Pressure vs Disease
- Heart Rate vs Disease

C. Risk Feature Analysis

- Chest pain type impact
- Exercise induced angina impact
- Major vessel count correlation

Phase 3: Statistical Insights

Using NumPy:

- Mean, Median, Standard Deviation
- Correlation Matrix
- Risk filtering conditions
- High-risk group segmentation

Phase 4: Visualization Strategy

Visuals Required:

- Histograms (Age, Cholesterol)
- Bar Charts (Gender vs Target)
- Boxplots (Heart Rate vs Target)
- Scatter Plots (Age vs Cholesterol)
- Correlation Heatmap (optional)

Advanced Analysis (Corporate-Level Thinking)

Risk Segmentation Model

Create risk categories:

- Low Risk
- Medium Risk
- High Risk

Using:

- Age
- Cholesterol
- Blood Pressure
- Oldpeak

Business Use:

Helps hospitals prioritize patient monitoring.

Expected Insights

Participants should be able to answer:

- Does cholesterol strongly impact heart disease?
- Is male population more vulnerable?
- Does exercise-induced angina significantly increase risk?
- Which feature has strongest correlation with disease?

Deliverables

Each participant/team must submit:

- Python Script / Jupyter Notebook
- 8–10 Visualizations
- Insight Summary Report (1–2 pages)

- Business Recommendation Slide (PPT – 3 slides)

Evaluation Criteria (Corporate Batch)

Criteria	Weightage
Code Quality	20%
Data Handling	20%
Visualization Clarity	20%
Insight Quality	25%
Business Interpretation	15%

Business Recommendations (Sample)

Based on analysis, recommendations may include:

- Prioritize patients with:
  - High cholesterol (>240)
  - Age > 50
  - High oldpeak
- Develop early screening programs for high-risk groups
- Use predictive scoring system in hospital systems

Learning Outcomes

After completion, participants will:

✓ Perform real-world healthcare data analysis
✓ Use Pandas for data manipulation
✓ Apply NumPy for statistical operations
✓ Build meaningful visualizations
✓ Translate analysis into business insights

✓ Think like a Data Analyst in enterprise setup