TOPS TECHNOLOGIES

PVT LTD

Predicting Heart Disease: A Case Study Using the UCI Dataset

**By :- Guided By :-**

**Krupal Prajapati Bhaumik Vyas**

**Index:-**

1. **Introduction**
2. **Dataset Overview**
3. **Explanation of Columns**
4. **Patient Scenario**
5. **Analysis**
6. **Interpretation & Communication**
7. **Conclusion**
8. **Introduction: -**

Heart Disease is one of the leading causes of Mortality Worldwide. Early

Detection plays a Critical role in preventing severe outcomes and improving patient care**.**

This case study uses the UCI Heart Disease Dataset to simulate a real-world scenario,

analysing key medical indicators to assess the likelihood of cardiovascular disease.

1. **Dataset Overview: -**

* **Source**: UCI Machine Learning Repository
* **Records**: 303
* **Attributes Used**: 14 clinical features (e.g., age, chest pain type, cholesterol, etc.)
* **Purpose**: Predict the presence and severity of heart disease

1. **Explanation of Column: -**

|  |  |
| --- | --- |
| **Column** | **Description** |
| age | Age in years |
| sex | 1 = Male, 0 = Female |
| cp | Chest pain type (1–4) |
| trestbps | Resting blood pressure (mm Hg) |
| chol | Serum cholesterol (mg/dl) |
| fbs | Fasting blood sugar > 120 mg/dl (1 = true, 0 = false) |
| restecg | Resting ECG results (0–2) |
| thalach | Maximum heart rate achieved |
| exang | Exercise-induced angina (1 = yes, 0 = no) |
| |  |  | | --- | --- | | oldpeak |  | | ST depression from exercise relative to rest |
| slope | Slope of peak ST segment (1–3) |
| ca | Number of major vessels (0–3) collared by fluoroscopy |
| thal | Thalassemia status (3 = normal, 6 = fixed defect, 7 = reversible defect) |
| num | Diagnosis of heart disease (0 = no disease to 4 = severe) |

1. **Patient Scenario :-**

**Patient Name:** John M.  
**Age:** 56  
**Sex:** Male  
**Chest Pain Type:** Asymptomatic  
**Resting BP:** 150 mm Hg  
**Cholesterol:** 260 mg/dl  
**Fasting Blood Sugar:** Normal  
**ECG:** Left ventricular hypertrophy  
**Max HR:** 120 bpm  
**Exercise Angina:** Yes  
**ST Depression:** 2.3  
**ST Slope:** Flat  
**Major Vessels:** 2  
**Thalassemia:** Reversible defect  
**Predicted Heart Disease Level:** 3 (Moderate to severe)

**Exploratory Analysis on John M.‘s Report:-** John M. is a 56-year-old male with high blood pressure and cholesterol levels. Though asymptomatic, exercise tests and ECG indicate signs of heart strain, placing him at moderate to severe risk.

1. **Analysis: -**

**5.1) Data Collection: -**

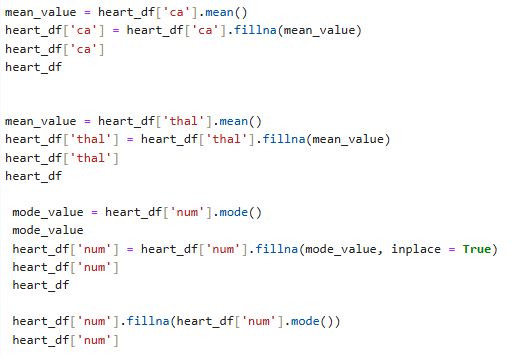
* **Source:**  
  UCI Machine Learning Repository — [Heart Disease Dataset](https://archive.ics.uci.edu/ml/datasets/heart+Disease)
* **Format:**  
  CSV file (heart.csv or similar)
* **Data Fields Include:**
  + age – Age of the patient
  + sex – Male (1), Female (0)
  + cp – Chest pain type
  + trestbps – Resting blood pressure
  + chol – Serum cholesterol in mg/dl
  + fbs – Fasting blood sugar > 120 mg/dl
  + restecg – Resting ECG results
  + thalach – Max heart rate achieved
  + exang – Exercise-induced angina
  + oldpeak – ST depression
  + thal – Thalassemia status
  + target – Presence of heart disease (1 = yes, 0 = no)

**5.2) Data Understanding 🔍**

* **Initial Exploration:**
  + Checked dataset shape: e.g., (303, 14)
  + Viewed data types and column meanings
  + target is our outcome variable (binary classification)
* **Insights Gathered:**
  + cp, thal, sex, fbs, exang, etc. are categorical
  + Some features are continuous (chol, age, oldpeak)
* **Techniques Used:**
  + df.describe() for summary statistics
  + df.info () for data types and null checks
  + Histograms and boxplots for distribution

**5.3) Data Cleaning:-**

**Fill the NaN Values to Mean**



### Over here, we are not removing the NaN value; instead that we are using the fillna method to replace the NaN value with the column mean or the Mode

### Convert the Numeric Value to Categorical Value:- By using Replace

### 5.4) Data Analysis:

### 5.4.1) Count the number of Reports Gender-wise :-

### 

### 

### From the Above Charts, we can easily identify that Males have more Reports in comparison to females

### This shows that approximately **68.3% (207)** of the dataset consists of male patient records, while the remaining **31.7% (96)** are female. This imbalance in gender representation is important to note

### 5.4.2)  Number\_of\_reports\_age-wise\_for\_each\_gender:

### 

### The first chart shows the age distribution for males, and the Second one tells about the Females

### In Males, we can see that a nice spike or wave came in the age of 52 to 62, and in females, we can see the spike between the ages of 52 to 67

### In the males, in just 10 years (age of 52 to 62), we can see 107 Reports, which is nearly 55 % of the total Reports.

### In females in this 15 age range (ages of 52 to 67), we can see 58 reports, which is exactly 55% of the Total Reports

### In short, we can find out from this that the plot is that as the age increases, the number of reports also increases

### 5.4.3) Minimum and Maximum Heart Rate during each stage of Diagnosis (Males) :

### 

### The above code is for stage 0 to see another stage, you need to change in male healthy in which you need to change 0 to (1,2,3,4). Any of them

### In stage 0, the minimum heart rate is 105, and the maximum heart rate is 202

### 202 is an outlier value over here; an official value should be around 187

### 

### In stage 1 of diagnosis, we can see over here that the minimum is 88 only, and the maximum is 195

### in stage 2, the maximum heart rate is 170, but when we see that the minimum heart rate is 71, which shows that a patient might have symptoms of Heart Disease

### In the stage 3 diagnosis of the male patients, the minimum heart rate is 90 and the maximum heart rate is 173

### 

### In stage 4 of diagnosis, the minimum heart rate is 124, and the maximum will be 182

### AS we can see, all the above stages of Diagnosis, in which we can see a great increment in heart rate, with the minimum heart rate per stage. Majorly, the 80 to 100 is normal in the adult body

### But when we talk about below 80, it might show several heart diseases. And if it is above 100, it is not normal; it is happening due to a high amount of exercise stress, and anxiety

### 5.4.4) Chest Pain Type vs. Heart Disease: -

### 

### 

### 

|  |  |
| --- | --- |
| 1 Typical Angina:- | Chest discomfort related to physical exertion or stress |

|  |  |  |
| --- | --- | --- |
| 2 | Atypical Angina:- | Unusual chest pain not related to physical exertion |

|  |  |  |
| --- | --- | --- |
| 3 | Non-Angina:- Pain | Chest pain not related to the heart |

|  |  |  |
| --- | --- | --- |
| 4 | Asymptomatic:- | No chest pain; may still have underlying heart issues |

### The visualization of chest pain types against heart disease severity reveals a critical insight—**asymptomatic individuals are significantly more likely to have heart disease** than those with typical or non-angina chest pain. While it’s common to associate chest pain with heart issues, the data shows that **relying solely on chest discomfort is not sufficient for diagnosis**. Most patients with classic symptoms (typical angina) were not diagnosed with heart disease. On the other hand, asymptomatic patients had the highest share of diagnoses, emphasizing the importance of **regular screening and use of diagnostic tools** even when symptoms aren’t present.

### 5.4.5) Problem Statement: -

### Why do we have the Highest number of Heart disease patients who have Asymptomatic Chest Pain?

### Normally, in the heart's anatomy, we find multiple vessels, and these vessels are used to circulate blood in the heart. If it does not circulate blood freely, there might be a chance that a person can develop heart problems

### In the same way as we mentioned above, the blood. Let's take an Example (if we eat healthy food, does it make us unhealthy? No, it always makes us healthy). The same way to run a Heart smoothly, the blood also needs to be of good quality.

### But there is a blood disorder called thalassemia. In which the Blood’s quality gets very Bad, and if a patient has the thalassemia of 6 (Fixed Defect) or 7 (Reversible Defect) means the Patient is at Risk

### 

### As we can see over here, this is a vessel blockage plot of Asymptomatic Patients. In which we can see that the maximum number of people have no blockage, but on the other side, 32 patients have 1 vessel blocked, 28 patients have 2 blockages of vessels, and in the last, there are 13 to 15 patients who have 3 vessel blockages. Which shows more risk of heart problems

### 

### Thalassemia is a type of blood disease in which the quality of the blood gets so bad that it is basically a blood disorder. Now, Normal Thalassemia is not leading to heart problems over here. It considers a healthy heart. But if the patients are having a Reversible or Fixed defect, it definitely indicates the heart problems

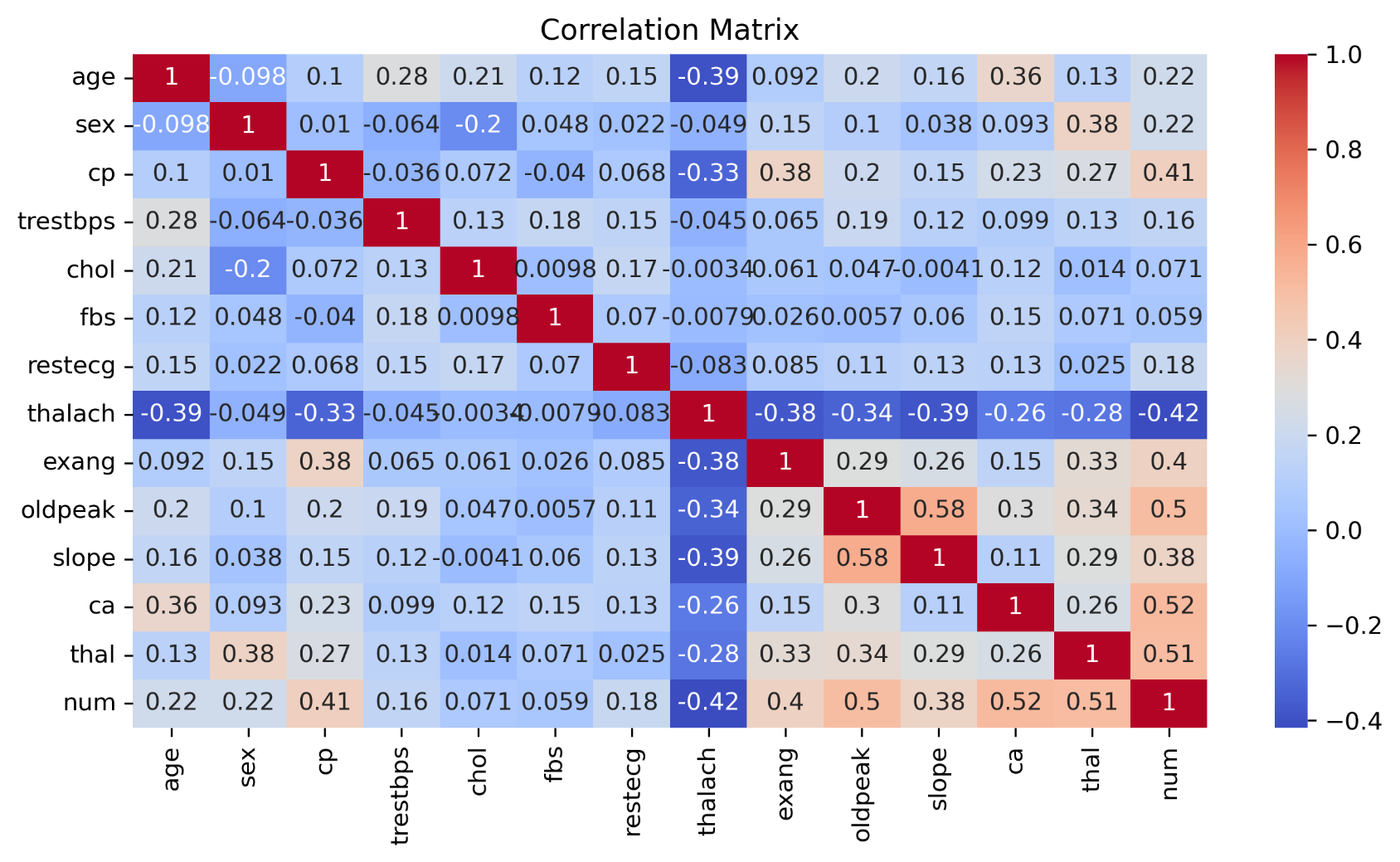
### Correlation Matrix: -

 Age**, cholesterol levels, and blood pressure** are likely significant factors contributing to heart disease.

 Thalassemia **defects** (if included as "thal") may increase the risk, especially if there are defects.

 High **cholesterol and high blood pressure** are often related, and together they could be major contributors to heart disease.

 Multicollinearity **issues** could arise between features like cholesterol and blood pressure, which may require further data pre-processing or feature engineering.

 Lifestyle **factors** (if available) like smoking, exercise, and diet could play an important role.

1. **Interpretation & Communication :-**

The analysis reveals that heart disease is most prevalent in males aged 52–67, with a surprising number of cases among asymptomatic individuals. Key contributing factors include high cholesterol, abnormal heart rate, elevated ST depression (oldpeak), and thalassemia defects. Asymptomatic patients with blocked vessels are at high risk, showing that symptoms alone are not reliable indicators. Age, chest pain type, and thalassemia status are strong predictors of heart issues. These insights highlight the importance of regular screening, early diagnosis, and proactive lifestyle interventions to reduce risk and improve patient outcomes—even in the absence of noticeable symptoms.

**6.1)** Tools Used:

🡪Matplotlib / Seaborne for visual storytelling

* + - Numpy for Numerical Computation
    - Pandas for Data Manipulation
    - Microsoft Word for Documentation
    - Jupiter Notebook for Python code

### Conclusion:-

### The UCI Heart Disease dataset analysis highlights that heart disease can occur even without typical symptoms, particularly in males aged 52–67. Critical risk indicators include thalassemia defects, high cholesterol, abnormal heart rates, and silent conditions like asymptomatic chest pain. Therefore, relying solely on symptoms is insufficient—regular health check-ups, early screening, and awareness of hidden risk factors are essential for timely diagnosis and prevention. Data-driven insights can support better clinical decisions and targeted interventions to reduce heart disease risk and improve patient care.

### Keep Exploring ….