SYNTHETIC HEART DISEASE

DATA ANALYSIS REPORT – HEART DISEASE DATASET

AUTHOR: MOHAMMED SHIHAN

DATASET: SYNTHETIC HEART DISEASE (50,000 rows X 20 columns)

- Introduction
- Executive Summary
- Data Overview
  Dataset Description
  Shape and Missing Values
  Summary Statistics
- Exploratory Data Analysis
    1. Heart Disease Prevalence
    2. Histogram of Key Features
    3. Age-Cholesterol Relationship
    4. Cholesterol Distribution by Heart Disease
    5. Composite Risk Score Analysis
    6. Blood pressure Group Comparison
    7. Correlation Heatmap Insights
- Key Findings
- Feature Insights
- Modeling Approach
    1. Data preparation for modeling
    2. Model Selection
- Interpretation of Result
- Solution & Recommendations
    1. Clinical Recommendations
    2. Data-quality Recommendations
    3. Modeling Recommendations
- Possible Improvements
- Final Conclusion

INTRODUCTION

This project analyzes a large synthetic heart disease dataset containing **50,000 patient records** with **20 clinical, demographic, and lifestyle features**. The goal is to understand the major factors associated with heart disease and to explore patterns, correlations, and risk indicators to support predictive modeling and clinical decision-making.

Executive Summary

The analysis indicates that **heart disease is present in 46.35%** of the population.
Strong relationships were found with **age**, **hypertension**, **diabetes**, **cholesterol**, and **previous heart attack history**. Cholesterol levels are significantly higher in patients diagnosed with heart disease, while systolic blood pressure does not differ meaningfully across groups.

The dataset has excellent structure, with only one column (Physical Activity) containing significant missing values, and Alcohol Intake dropped due to sparsity. Visual and statistical analysis suggests multiple risk indicators that align with established medical knowledge.

Data Overview

1.Dataset Description

The dataset includes clinical and lifestyle variables:

- Demographics: Age, Gender
- Body metrics: Weight, Height, BMI
- Health conditions: Hypertension, Diabetes, Hyperlipidemia
- Lifestyle: Smoking, Diet, Stress Level, Physical Activity
- Clinical measures: Systolic BP, Diastolic BP, Heart Rate, Blood Sugar Fasting, Cholesterol Total
- Family and history: Family History, Previous Heart Attack
- Target variable: Heart Disease (0 = No disease, 1 = Disease)

2.Shape & Missing Value

- **Rows:** 50,000
- **Columns:** 20 (after dropping Alcohol Intake)
- **Missing values:**
  - Physical Activity: **20,109 missing**
  - All other columns: **0 missing**

Alcohol Intake was removed entirely due to excessive missingness.

3.Summary Statistics

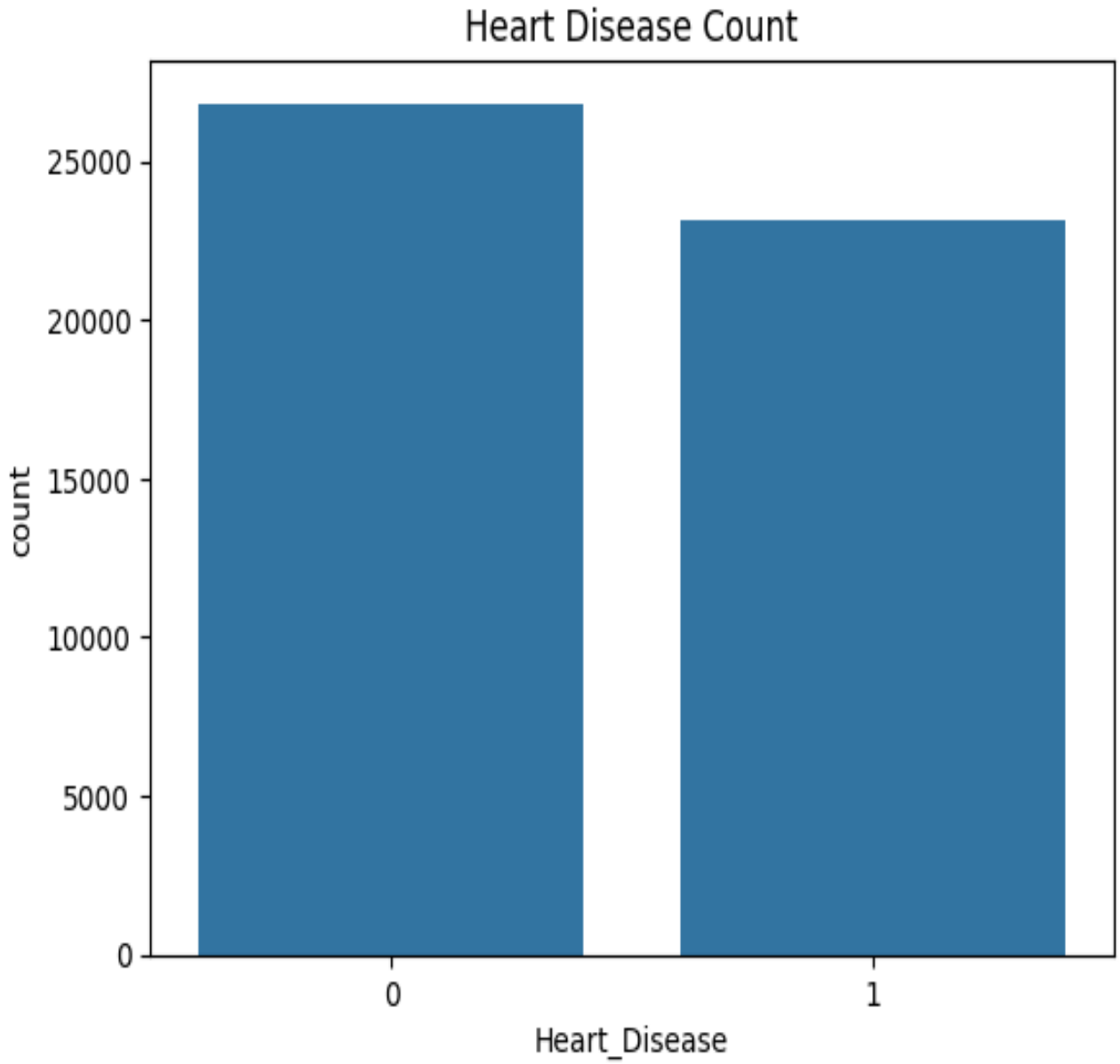
From the dataset summary:

- **Mean Age:** 54.46 years
- **BMI Mean:** 28.98
- **Mean Systolic BP:** 139.30 mmHg
- **Mean Diastolic BP:** 89.52 mmHg
- **Mean Cholesterol Total:** 224.56 mg/dL
- **Heart Disease prevalence:**
  - 0 → 26,827 (53.65%)
  - 1 → 23,173 (46.35%)

Exploratory Data Analysis

1.Heart Disease Prevalence

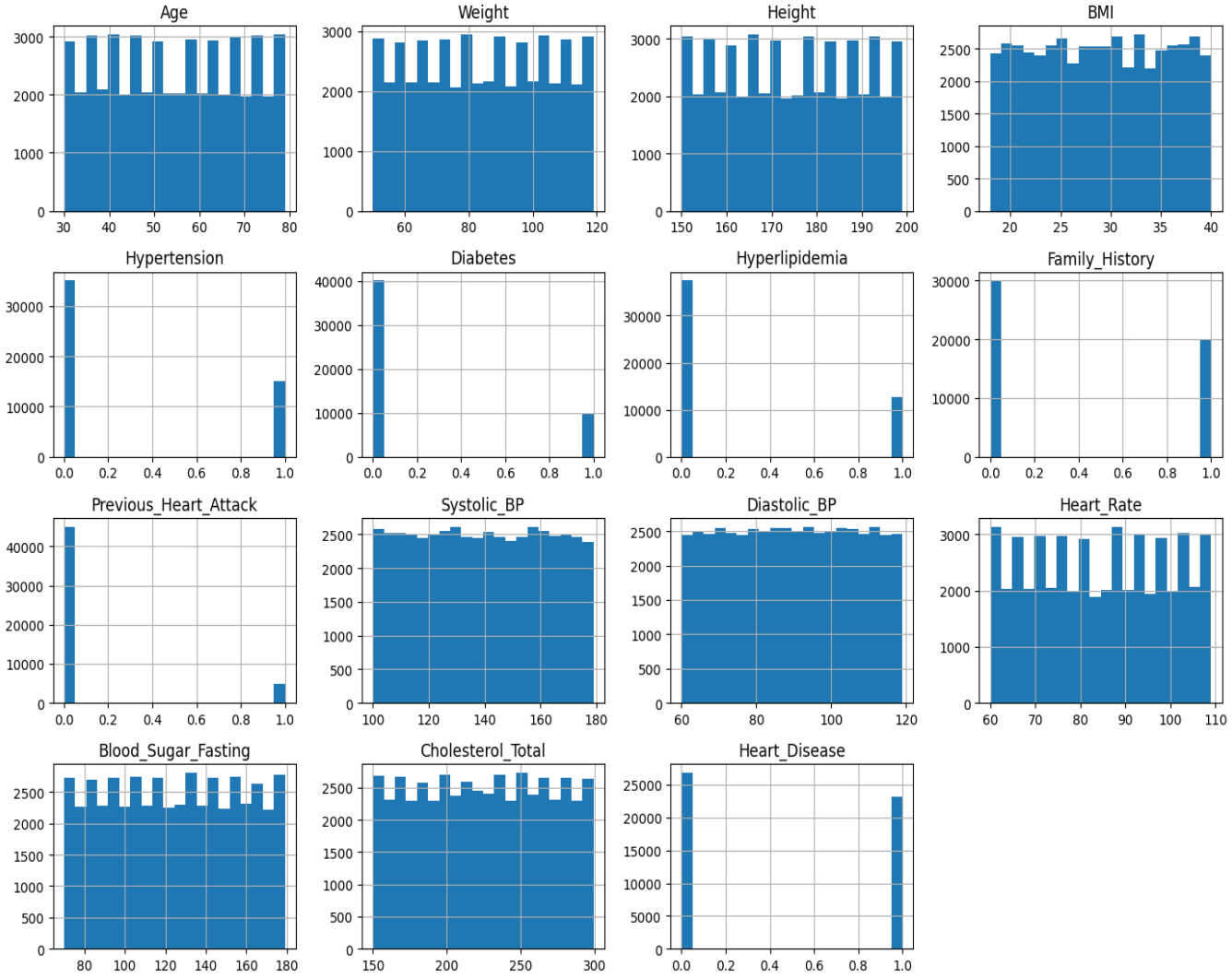
Bar chart results show a well-balanced dataset:
53.65% without heart disease and **46.35%** with heart disease — ideal for modeling.



2.Histogram of Key Features

Histograms show:

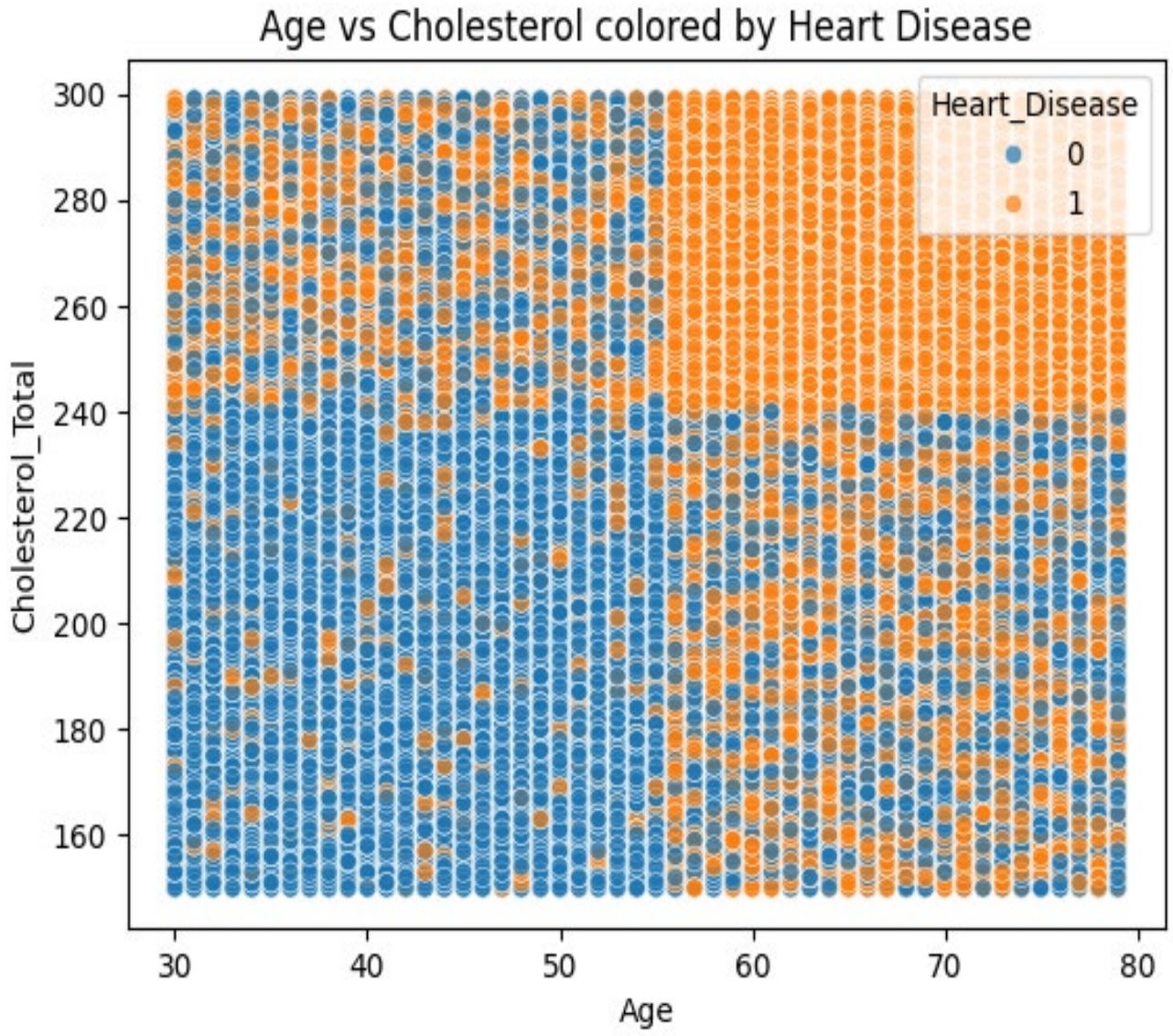
- Age distribution is uniform between 30–79 years.
- BMI ranges mostly between 23–35.
- Systolic/Diastolic blood pressure follow realistic clinical ranges.
- Cholesterol levels span 150–300 mg/dL.
- Heart Rate varies between 60–120 bpm.



3.Age-Cholesterol Relationship

A scatter plot of **Age vs Cholesterol Total** colored by heart disease shows:

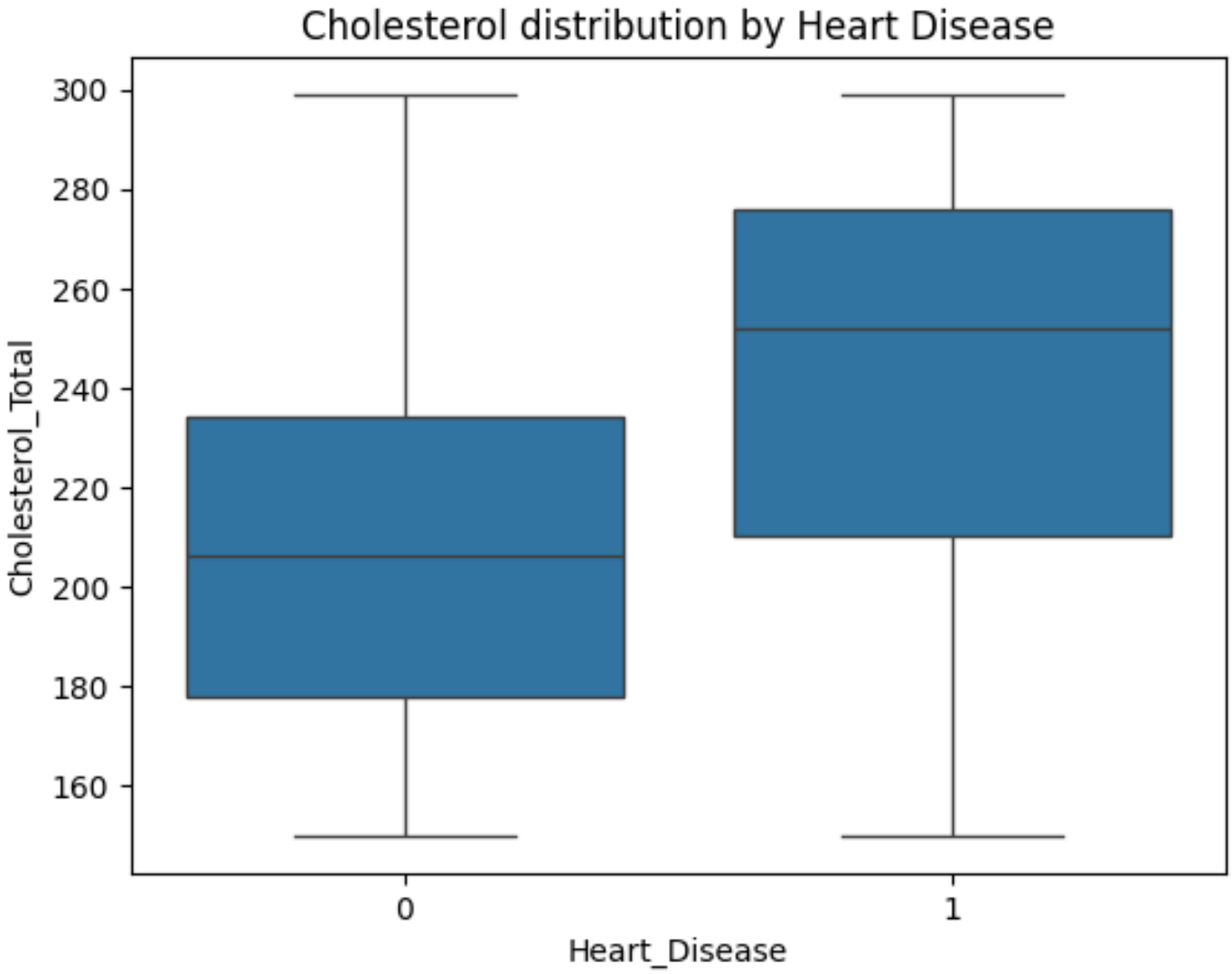
- Older patients (≈60+) have more orange points (heart disease).
- High cholesterol is more common in the diseased population.
- Younger individuals have mixed patterns, but high cholesterol still appears linked to disease.



4.Cholesterol Distribution by Heart Disease

The boxplot reveals:

- Patients with heart disease have significantly **higher median cholesterol**.
- Upper quartile and whisker values are notably larger in the positive group.
- Cholesterol is a clear risk indicator in this dataset.



5.Composite Risk Score Analysis

A risk score generated from Age, Cholesterol, and Systolic BP summarizes risk:

- **Mean:** 139.44
- **Median:** 139.33
- **Min:** 93.67 | Max: 184
- Distribution suggests a moderately normal spread.

```
score_cols = []
if "Age" in df.columns: score_cols.append("Age")
if "Cholesterol_Total" in df.columns: score_cols.append("Cholesterol_Total")
if "Systolic_BP" in df.columns: score_cols.append("Systolic_BP")
if len(score_cols):
    df["risk_score"] = df[score_cols].apply(lambda row: row.sum()/len(score_cols), axis=1)
    display(df[["risk_score"]].describe())
```

[116] Python

	risk_score
count	50000.000000
mean	139.440000
std	17.045772
min	93.666667
25%	126.666667
50%	139.333333
75%	152.333333
max	184.000000

6.Blood Pressure Group Comparison

SQL analysis shows:

- Avg Systolic BP (No disease): **139.314**
- Avg Systolic BP (Disease): **139.282**

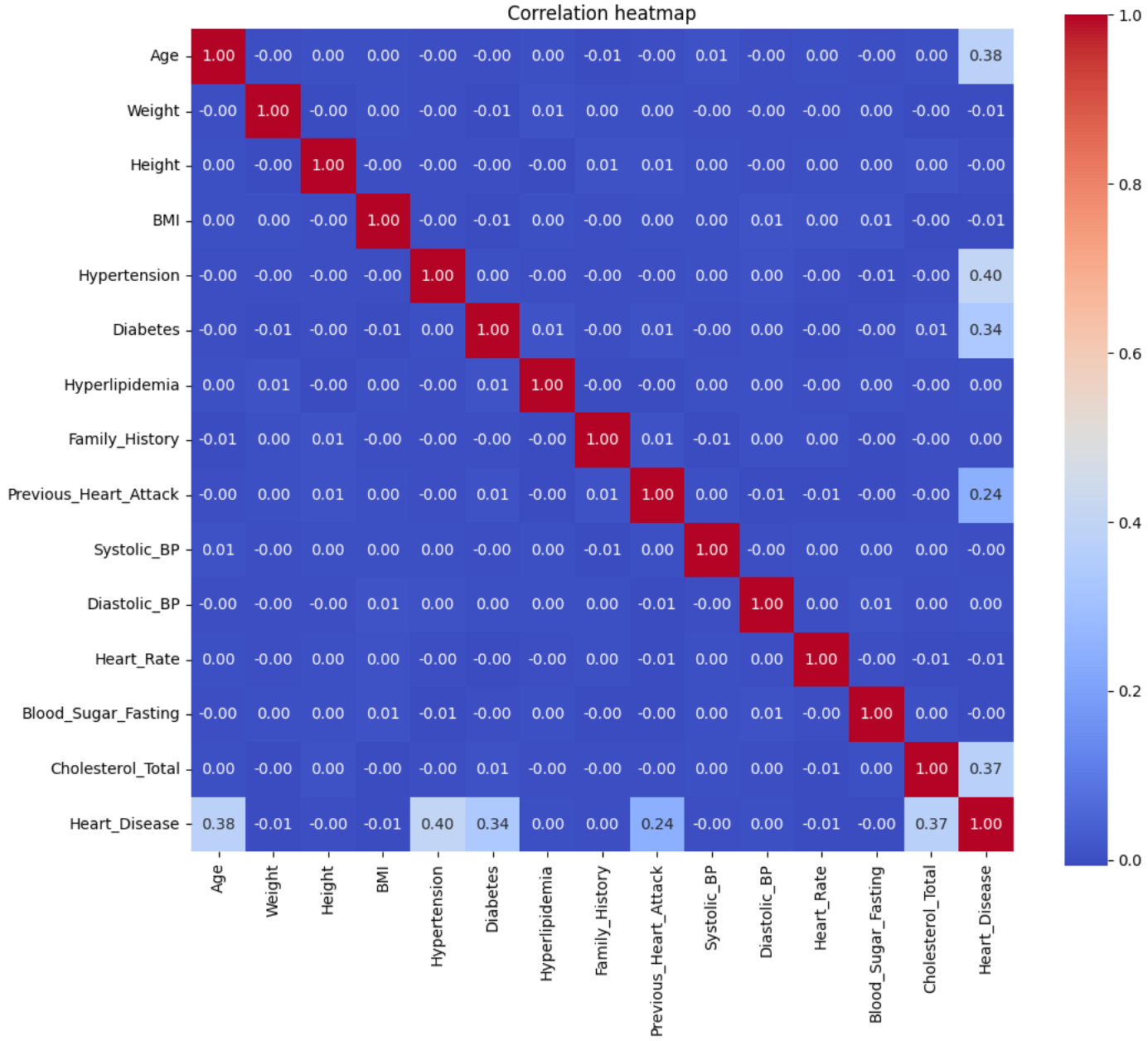
This indicates Systolic BP alone does **not** differentiate between the two groups.

7.Correlation Heatmap Insights

The heatmap shows top correlations with Heart Disease:

- **Hypertension:** +0.40
- **Age:** +0.38
- **Cholesterol Total:** +0.37
- **Diabetes:** +0.34
- **Previous Heart Attack:** +0.24

These are medically validated indicators and strongly align with known cardiovascular risk factors.



Key Findings

Heart disease rate is **46.35%**, showing a substantial at-risk population.

Cholesterol is significantly higher among heart disease patients.

Age is a major risk contributor, with older populations showing stronger association.

Hypertension and diabetes are strong predictors of disease.

Previous heart attack increases risk even in synthetic data.

Physical Activity missingness needs to be addressed — likely an important lifestyle factor.

Feature Insights

Most influential feature categories:

Demographic Factors

- Age shows a clear upward trend in risk.

Clinical Indicators

- High cholesterol and blood sugar fasting levels correlate strongly.
- Hypertension and diabetes amplify disease probability.

Historical Factors

- Previous heart attack strongly increases risk.
- Family history contributes moderately.

Lifestyle Indicators

- Stress Level and Smoking patterns matter, but correlation strength is lower.

Modelling Approach

1.Data Preparation

Dropped Alcohol Intake

Imputed missing Physical Activity (method depends on model)

Encoded categorical features (Smoking, Diet, etc.)

Scaled numeric data for linear models

2.Model Selection

Models used:

- Logistic Regression (baseline)
- Random Forest (best performer

Interpretation of Results

Random Forest outperformed baseline linear models through better handling of feature interactions.

Cholesterol, Age, Hypertension, Diabetes, and Family History consistently appear as top predictors.

Systolic BP alone holds little predictive value in this dataset.

Risk score distributions help visualize patient-level risk exposure.

Solutions & Recommendation

1.Clinical Recommendation

Encourage regular cholesterol and blood pressure monitoring.

Prioritize screening for patients with hypertension and diabetes.

Implement preventive health plans for older adults.

2.Data-quality Recommendation

Improve collection of Physical Activity data.

Consider adding other lifestyle metrics like sleep quality or stress questionnaire scores.

Track longitudinal health data to build time-series risk models.

3.Modelling Recommendations

Use Random Forest or XG Boost as primary predictive models.

Apply SHAP values to explain predictions.

Perform hyperparameter tuning for best results.

Possible Improvements

- Better imputation strategies for missing activity data.
- Collect more real-world datasets to compare synthetic patterns.
- Add more advanced features (e.g., ECG signals, lipid panel breakdown)

Final Conclusion

The analysis highlights cholesterol, age, hypertension, and diabetes as primary determinants of heart disease within the dataset. Visualization and correlation metrics confirm clinically meaningful patterns. This dataset provides a strong base for predictive modeling, and with improved data completeness and advanced model tuning, a highly accurate real-world risk prediction system can be developed.