# **Heart Disease Dataset Analysis**

**Name**:- Jasneet Kaur Chhabra, Krishna Uprit, Tanishq Vyas, Saishreyas Vaval

**Batch** - 2A

## **Abstract**

This document provides an in-depth analysis of a heart disease dataset to identify patterns and risk factors associated with heart disease. By evaluating the dataset, we explore how predictive analytics can enhance early diagnosis and prevention strategies. We will be using Classification techniques.

Dataset Availability Statement– This dataset was taken from the following Website:- <https://www.kaggle.com/datasets/oktayrdeki/heart-disease>

## **Introduction**

### 

### **Background**

Heart disease is one of the leading causes of mortality worldwide. Early detection and diagnosis of heart disease can significantly improve patient outcomes and reduce healthcare costs. The dataset we are analyzing contains multiple health indicators such as blood pressure, cholesterol levels, smoking habits, exercise routines, and genetic predispositions.

Machine learning provides an efficient way to identify patterns and relationships in this dataset to classify individuals into high-risk or low-risk categories. By using classification models, we aim to improve early diagnosis and preventive healthcare strategies.

### 

### **Objective/Purpose**

The primary objective is to predict whether a person has heart disease (Yes/No) based on these features.

**Scope**

This study focuses on structured data containing demographic, lifestyle, and medical history variables. Only numerical and categorical features relevant to heart disease prediction are considered.

## **Problem Statement**

Heart disease is one of the leading causes of mortality worldwide. Early detection and diagnosis of heart disease can significantly improve patient outcomes and reduce healthcare costs. The dataset we are analyzing contains multiple health indicators such as blood pressure, cholesterol levels, smoking habits, exercise routines, and genetic predispositions. The primary objective is to predict whether a person has heart disease (Yes/No) based on these features. Heart disease prediction remains a significant challenge due to the complexity of risk factors. The key problem is to determine the best machine learning approach—classification for accurate heart disease prediction. Understanding the relationship between various risk factors and heart disease is crucial. Features such as cholesterol levels, blood pressure, and lifestyle habits may contribute significantly.

### **Identification of Problem**

The dataset consists of both categorical and numerical attributes, making it necessary to preprocess and analyze data before model selection.

The problem we are solving is a binary classification problem, where the goal is to classify individuals into two categories:

* Positive (Yes): The person has heart disease.
* Negative (No): The person does not have heart disease.

**Challenges in the problem**

* Imbalanced Data: If there are significantly more 'No' than 'Yes' cases, we may need resampling techniques.
* Feature Selection: Identifying the most important predictors among various health indicators.
* Handling Missing Values: Missing data in key variables could impact model accuracy.
* Model Generalization: Ensuring the model does not overfit but generalizes well on new patient data.

### **Hypothesis Formulation**

To develop an effective classification model, we formulate the following hypotheses based on medical research and common risk factors of heart disease:

**Null Hypothesis (H0)**

There is no significant relationship between the given health indicators and the likelihood of having heart disease.

**Alternative Hypothesis (H1)**

Certain health indicators significantly influence the likelihood of heart disease. Some potential relationships include:

1. Higher Blood Pressure increases the risk of heart disease.

2. Elevated Cholesterol Levels correlate with higher heart disease incidence.

3. Smoking and Alcohol Consumption contribute to a higher risk of heart disease.

4. Obesity (High BMI) increases the probability of heart disease.

5. Low Physical Activity (Exercise Habits) is a risk factor.

6. Family History of Heart Disease is a strong predictor.

7. Diabetes is positively correlated with heart disease risk.

8. Lower Sleep Hours and High-Stress Levels may be associated with heart disease.

9. Higher CRP (C-reactive protein) and Homocysteine Levels are indicators of increased heart disease risk.

Through data exploration and classification modeling, we will test these hypotheses and determine which factors have the most significant impact on heart disease prediction.

* H1: Machine learning models can accurately predict heart disease risk.
* H2: Classification models will outperform regression models in predicting heart disease status.

## **Methodology**

### **Search Strategy**

The dataset was explored using statistical and machine learning techniques to identify trends and correlations.

### **Selection Criteria**

* Inclusion: Relevant features such as demographic data, health metrics, and lifestyle choices.
* Exclusion: Unnecessary or highly missing data fields.

### **Data Extraction**

Data preprocessing steps include handling missing values, encoding categorical variables, and normalizing numerical features.

## **Main Body (Literature Review & Analysis)**

### **Thematic Organization**

The analysis is structured around:

* Feature importance in predicting heart disease
* Challenges in data preprocessing and handling missing values

### **Critical Analysis**

The dataset presents both numerical and categorical variables, requiring careful preprocessing.

**Key Findings**

* Blood pressure, cholesterol levels, and diabetes status are significant predictors of heart disease.
* Classification models outperform regression models for heart disease prediction.
* Random Forest and Gradient Boosting models provide the best accuracy.

### **Limitations of Existing Studies**

Many studies lack diverse datasets, which limits model generalization. Additionally, handling missing values remains a challenge.

## **Discussion**

### **Trends & Patterns**

* Patients with high cholesterol and hypertension exhibit a higher likelihood of heart disease.
* Lifestyle factors such as smoking and lack of exercise significantly influence heart disease risk.

### **Theoretical & Practical Implications**

Accurate predictive models can aid in early diagnosis, helping medical professionals intervene sooner.

### **Future Directions**

* Exploring deep learning models for improved predictions.
* Incorporating time-series data for tracking disease progression.

## **Conclusion**

### **Summary**

The analysis confirms that classification models, particularly Random Forest and Gradient Boosting, are best suited for heart disease prediction.

### **Contribution**

This study highlights key risk factors and demonstrates the effectiveness of classification models in predicting heart disease.

### **Final Thoughts**

Machine learning has the potential to revolutionize heart disease prediction, allowing for proactive healthcare interventions. Future research should focus on refining models using larger and more diverse datasets.