Final Internship Report

Project Title: Advanced Data Analysis on Health and Demographic Data to Identify Common Traits Leading to Heart Disease

Submitted by: Afiya Abbasi

Program: BBA – Business Analytics

Roll Number: MS23MS303003

Organization: Practo

Submission Date: 26 July 2025

1. Introduction

Cardiovascular diseases (CVDs) continue to be the leading cause of mortality worldwide. Early detection of risk factors plays a critical role in effective prevention and timely treatment. The objective of this internship project was to analyze a health and demographic dataset to uncover common traits among individuals diagnosed with heart disease. The analysis focused on statistical and visual exploration of variables including age, BMI, blood pressure, cholesterol, glucose levels, and lifestyle factors such as smoking and physical activity.

2. Objective

The primary goal of this project was to identify and understand the most prevalent health and demographic characteristics that contribute to heart disease. Using exploratory data analysis (EDA) and correlation insights, the findings were aimed at assisting healthcare practitioners and platforms like Practo in supporting proactive healthcare and personalized preventive strategies.

3. Dataset Overview

Dataset Source: Kaggle – Cardiovascular Disease Dataset

File Used: cardio_train.csv

Total Records: 70,000

Number of Features: 13

Target Variable: cardio (1 = diagnosed with heart disease, 0 = not diagnosed)

Key Variables in the Dataset

Feature	Description
age	Age in days (converted to years)
gender	1 = Female, 2 = Male
height, weight	Used to calculate Body Mass Index (BMI)
ap_hi, ap_lo	Systolic and diastolic blood pressure
cholesterol, gluc	1 = Normal, 2 = Above Normal, 3 = Well Above Normal
smoke, alco, active	Lifestyle indicators
cardio	Target variable

4. Data Cleaning and Preprocessing

- Removed the id column as it offered no analytical value.
- Converted age from days to years for better interpretability.
- Computed **BMI** using the formula:
  BMI = weight (kg) / (height (m))^2
- Transformed categorical variables such as cholesterol and cardio into descriptive labels.
- Visualized and examined outliers in key variables such as BMI and blood pressure for better accuracy in modeling.

5. Exploratory Data Analysis (EDA)

a. Heart Disease Distribution by Smoking Status

- **Observation:** Heart disease was observed in both smokers and non-smokers, with a slightly higher proportion among smokers.
- **Inference:** Smoking is linked to an increased risk, although it may not be a dominant factor in isolation.

b. Physical Activity Distribution

- **Observation:** Approximately 80% of participants reported being physically active.
- **Insight:** A significant number of physically active individuals were also diagnosed with heart disease, suggesting other influencing factors.

c. Heart Disease vs Physical Activity

- **Visual Analysis:** A higher incidence of heart disease was noted among inactive individuals.
- **Conclusion:** Physical inactivity is a contributing risk factor and supports the importance of regular physical activity.

d. BMI Distribution

- **Observation:** The majority of BMI values ranged between 20–35.
- **Further Analysis:** Higher BMI values were frequently associated with heart disease.
- **Inference:** Overweight and obesity are strong contributors to cardiovascular risk.

6. Correlation and Feature Analysis

a. Correlation Heatmap

- **Strong Correlations:**
  - Weight and BMI (positive correlation)
  - Systolic and diastolic blood pressure
- **Moderate Correlations:**
  - Age and heart disease
  - Cholesterol levels and heart disease
  - Glucose levels and heart disease
- **Inference:** Multiple health indicators such as high blood pressure, elevated BMI, and increased cholesterol levels are associated with higher risk of heart disease.

b. Systolic vs Diastolic Blood Pressure

- **Observation:** Heart disease cases clustered around systolic values above 140 mmHg.

- **Conclusion:** High blood pressure is a significant risk factor for heart disease.

c. Heart Disease Rate by Age

- **Observation:** The proportion of individuals with heart disease increased steadily with age.
- **Inference:** Age is a non-modifiable but strong risk factor for cardiovascular conditions.

d. BMI Distribution by Heart Disease

- **Analysis:** Violin plots showed individuals with heart disease tend to have higher BMI levels.
- **Conclusion:** BMI remains a critical screening metric for early risk identification.

e. Pair Plot Analysis

- Revealed multidimensional relationships among features such as age, BMI, systolic blood pressure, and glucose levels.
- **Insight:** Identified overlapping clusters indicating zones of increased risk.

7. Key Insights and Patterns

Trait	Observation	Risk Association
Age	Higher risk observed in older individuals	Strong
BMI	Increased BMI linked to heart disease	Moderate to Strong
Blood Pressure	Elevated systolic and diastolic values common	Strong
Cholesterol	Above-normal cholesterol more prevalent in cases	Moderate
Glucose	Higher glucose levels correlated with disease	Moderate
Smoking	Slightly higher prevalence among smokers	Weak to Moderate
Physical Activity	Inactive individuals show higher risk	Moderate

8. Limitations

- **Self-reported lifestyle data** may introduce reporting bias (e.g., smoking, physical activity).

- **Lack of data** on nutrition, genetics, mental health, and sleep—important contributors to cardiovascular risk.

- **Cross-sectional dataset** limits the ability to establish cause-effect relationships.

9. Recommendations for Practo

- Encourage regular monitoring of BMI, blood pressure, and cholesterol through the Practo platform.
- Implement preventive health risk calculators based on this analysis to personalize user care.
- Introduce AI-driven alerts for individuals with combinations of high-risk traits (e.g., high BMI, elevated BP, inactivity).
- Run awareness campaigns promoting healthy lifestyles—targeting smoking cessation, weight management, and physical activity.

10. Conclusion

This project effectively identified critical patterns and risk traits associated with heart disease through in-depth data analysis. Key variables such as age, BMI, blood pressure, and cholesterol levels emerged as strong indicators of cardiovascular risk. Health-tech platforms like Practo can utilize these insights to enhance user engagement, promote preventive care, and reduce the overall burden of heart disease through data-driven interventions.