Jio Institute

PROJECT REPORT

UCI Heart Disease

Under guidance of:

Prof. MV Yugandhar

T.A. Nikita Divay

Project members:

Mr. Abhijit Zende

Mr. Saksham Agrawal

Mr. Sharvin Shinde

Table of contents:

1. Introduction:

Cardiovascular disease is one of the leading causes of mortality worldwide. Early diagnosis and understanding of the risk factors play a vital role in prevention. In this project, we performed an in-depth statistical analysis of the UCI Heart Disease dataset to extract meaningful insights, identify key associations, and evaluate significant differences among patient subgroups based on various health indicators.

2. Why Do We Need Statistical Analysis?

Statistical analysis is essential in healthcare datasets for the following reasons:

- To summarize and understand data distributions (e.g., age, cholesterol).
- To detect relationships between variables (e.g., age vs. heart rate).
- To validate hypotheses about disease indicators (e.g., chest pain type and disease presence).
- To guide medical decision-making through evidence-backed statistical findings.

3. Statistical Tests Performed

a) Descriptive Statistics

- Used to summarize data: mean, median, min, max for variables like age, cholesterol, and blood pressure.

b) Correlation Analysis

- Measured the strength and direction of relationships between numerical variables.
- Example: Negative correlation between age and maximum heart rate (older patients tend to have lower heart rates).

c) One-Sample T-Test

- **Objective:** Compare sample mean to a known value.
- **Example:** Is average cholesterol significantly different from 200 mg/dL?
- **Conclusion:** No significant difference ($p = 0.81$).

d) Two-Sample T-Test

- **Objective:** Compare means between two independent groups.
- **Example:** Max heart rate in patients with vs. without heart disease.
- **Conclusion:** Significant difference ($p < 0.001$).

e) ANOVA (Analysis of Variance)

- **Objective:** Compare means across more than two groups.
- **Example:** Cholesterol levels across different chest pain types.
- **Conclusion:** Significant differences observed ($p < 0.001$).

f) Simple Linear Regression

- **Objective:** Predict cholesterol based on age.
- **Conclusion:** Age negatively affects cholesterol slightly but model has low predictive power ($R^2 = 0.0074$).

g) Chi-Square Test of Independence

- **Objective:** Test for association between two categorical variables.
- **Example:** Chest pain type vs. heart disease presence.
- **Conclusion:** Strong association found ($p < 2.2e-16$).

4. Key Differences

Test Type	Use Case	Data Requirement	Example in Project
One-Sample T-Test	Compare mean to a fixed value	One numeric variable	Mean cholesterol vs. 200 mg/dL
Two-Sample T-Test	Compare means of two independent groups	One numeric + one binary categorical	Heart rate for disease vs. non-disease
ANOVA	Compare means of more than two groups	One numeric + one categorical (≥3 levels)	Cholesterol across chest pain types
Chi-Square Test	Test association between two categorical variables	Two categorical variables	Chest pain type vs. heart disease
Correlation	Measure linear relationship between two numerics	Two numeric variables	Age vs. Heart Rate
Linear Regression	Predict one variable based on another	One numeric dependent, one independent	Predict cholesterol from age

5. Conclusion:

This project demonstrated how statistical analysis can uncover significant relationships in health data. Key insights include:

- Patients with heart disease tend to have lower maximum heart rates.
- Chest pain type has a significant association with disease presence.
- Age shows weak but significant effects on cholesterol levels.

Statistical tools not only aid in interpretation but also provide evidence-based support for medical hypotheses. Such analysis can play a crucial role in the early detection and prevention of heart disease.

6. References and Source Code:

- UCI Heart Disease Dataset: https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data/
- Montgomery, D.C., & Runger, G.C. (2014). *Applied Statistics and Probability for Engineers*.
- Moore, D. S., McCabe, G. P., & Craig, B. A. (2017). *Introduction to the Practice of Statistics*.
- Online R documentation: https://www.rdocumentation.org/

Source Code:

https://github.com/Abhiz2411/HeartWiseAnalytics