University of KURDISTAN Hewlêr

University of Kurdistan – Hewlêr
College of Engineering

Project By

Mohammed Raheem

Najmadin Kawa

Module Leader

Prof. Ibrahim Ismael

End-to-End Machine Learning Project with GUI, GitHub Copilot, and Clean Code Practices

Introduction:

The project's Heart Disease Prediction Dataset, which used from Kaggle, includes crucial medical data that we can utilized to make a prediction if someone has heart disease or not. Because of its medical significance and well-structured nature, this dataset is useful specifically for machine learning programs in the healthcare word.

Dataset Overview

Source: Kaggle (Heart Disease Prediction Dataset by Utkarsh Singh)

- Size: 270 observations (patients)
- Features: 13 medical parameters plus 1 target variable
- Target Variable: Heart disease presence (binary classification)
- Data Quality: No missing values, clean dataset ready for analysis

Key Features Description

The dataset holds the following 13 health parameters that help as input keys for the machine learning:

1.Age: Patient's age in years (range: 29-77 years)

2.Sex: Gender (0 = Female, 1 = Male)

3.Chest Pain Type: Type of chest pain experienced (1-4 scale)

• Typical Angina

• Atypical Angina

• Non-Anginal Pain

• Asymptomatic

4.Resting Blood Pressure: Blood pressure at rest (mm Hg)

5.Serum Cholesterol: Cholesterol level in blood (mg/dl)

6.Fasting Blood Sugar: Whether fasting blood sugar > 120 mg/dl (0 = No, 1 = Yes)

7.Resting ECG Results: Electrocardiographic results at rest (0-2 scale)

• Normal

• ST-T Wave Abnormality

• Left Ventricular Hypertrophy

8.Maximum Heart Rate: Maximum heart rate achieved during exercise

9.Exercise Induced Angina: Whether exercise induces angina (0 = No, 1 = Yes)

10.ST Depression (Oldpeak): ST depression induced by exercise relative to rest

11.ST Segment Slope: Slope of the peak exercise ST segment (1-3 scale)

• Upsloping

• Flat

• Downsloping

12.Major Vessels: Number of major vessels colored by fluoroscopy (0-3)

13.Thalassemia: Blood disorder type (3, 6, 7)

• Normal

• Fixed Defect

• Reversible Defect

Target Variable Distribution

The target variable shows heart disease incidence with the following delivery:

• No Heart Disease (Class 0): 150 patients (55.6%)

• Heart Disease (Class 1): 120 patients (44.4%)

This distribution shows a stable dataset, which is amazing when it comes to machine learning model which balances the dataset and make it more accurate.

Clinical Significance

This dataset represents real-world medical scenarios where healthcare professionals need to assess heart disease risk based on multiple clinical indicators. The features included are commonly measured in routine cardiac examinations, making the model practically applicable in clinical settings.

Model Reliability

The model has accuracy of 85% and balanced the performance. The high accuracy for No Heart Disease is 92% with reducing false output.

Conclusion

This test well highlighted and use of machine learning in healthcare by predicting cardiac disease. An effective tool for assessing cardiovascular risk is produced by combining a well-curated dataset, a strong Logistic Regression model, and an intuitive user interface. This method is appropriate for screening applications because to its 85% accuracy, and clinical acceptability is guaranteed by the model's interpretable nature. Best practices in machine learning development, such as appropriate data preprocessing, model validation, and user interface design, are demonstrated in this project.