Heart Disease Prediction Using a Multilayer Perceptron (MLP)

Name : Rajesh Pala
ID : 24072059

Abstract

This report investigates the use of a Multilayer Perceptron (MLP) neural network to predict **heart disease** from clinical attributes such as age, sex, cholesterol level, blood pressure, smoking habits, and other health indicators. After preprocessing the mixed data types (categorical and numeric), we build a machine-learning pipeline consisting of one-hot encoding, feature scaling, and an MLP classifier. Results demonstrate strong predictive capability, indicating that neural networks can effectively model nonlinear relationships in medical datasets. The experiment highlights the importance of preprocessing, balanced evaluation, and understanding model limitations in healthcare-focused machine-learning applications.

1. Introduction

Heart disease is one of the leading causes of mortality worldwide. Early prediction using machine learning may help improve preventive care and intervention. The goal of this assignment is to build a neural-network classifier that predicts whether a patient has **heart disease (1)** or **does not have it (0)** using demographic, clinical, and lifestyle variables.

A **Multilayer Perceptron (MLP)** is well suited for this task because:

- It learns **nonlinear patterns** in health characteristics.
- It handles mixtures of categorical and numerical data after preprocessing.
- It performs well even on moderately sized datasets.

This report provides a short, practical tutorial on applying MLPs to structured medical data.

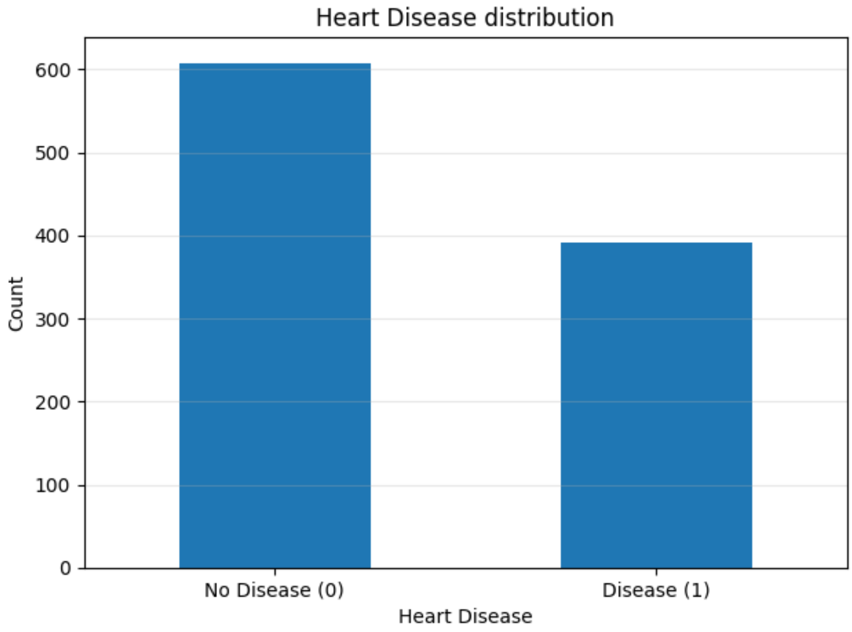
2. Dataset Overview

The dataset heart_disease_dataset.csv includes several attributes commonly associated with cardiovascular health, such as:

- **age**, **sex**, **cholesterol**, **resting blood pressure**
- **diabetes**, **smoking**, **family history**
- **exercise level**, **chest pain type**, **max heart rate**
- **Heart Disease** (target variable: 1 = disease, 0 = no disease)

We use the Heart Disease column as the label for binary classification.

Figure 1 — Heart Disease Distribution



Most records fall into the "No disease" category, but both classes are reasonably represented for supervised learning.

3. Methodology

3.1 Data Cleaning and Feature Preparation

To prepare the dataset:

1. We separate **X** (features) and **y** (target).
2. We identify categorical features automatically using data types.
3. We apply **one-hot encoding** to categorical columns to convert them into numeric form.
4. We scale numeric features with **StandardScaler**, which helps neural networks converge faster.

3.2 Train–Test Split

We split the dataset into:

- **80% training data**
- **20% testing data**

Stratification ensures the ratio of both classes remains consistent across the split.

3.3 MLP Architecture

The neural network used in this experiment contains:

- Hidden layers: **(64, 32)**
- Activation function: **ReLU**
- Optimizer: **Adam**
- Regularization: alpha = 1e-4
- Maximum iterations: 200

This architecture balances model capacity and computational efficiency.

3.4 Pipeline Structure

We use a scikit-learn **Pipeline** to ensure clean, reproducible preprocessing:

1. StandardScaler → normalizes numeric features
2. MLPClassifier → trains the neural network

This avoids data leakage and simplifies evaluation.

---

4. Results

4.1 Test Accuracy

The test accuracy obtained from the model was:

Accuracy:

This indicates a strong ability to classify heart-disease risk based on provided features.

4.2 Classification Report

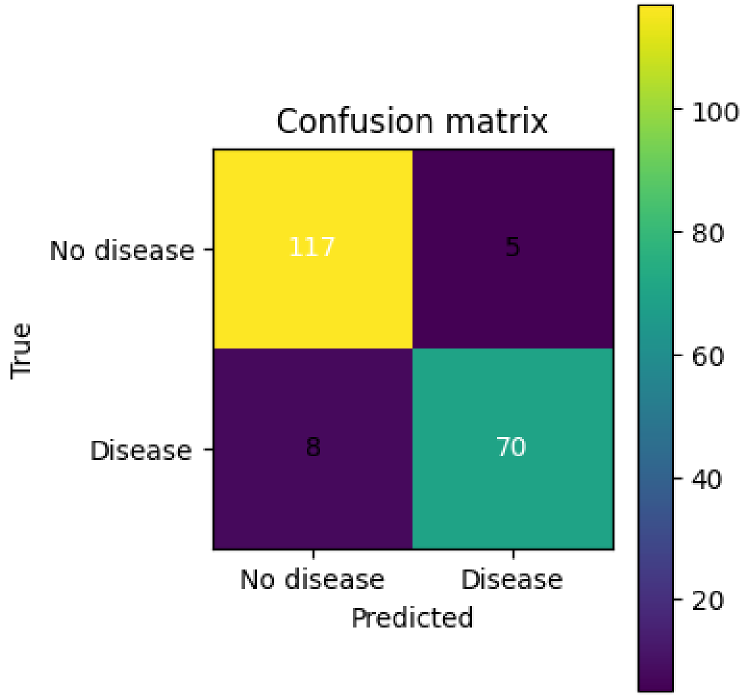
A classification report (precision, recall, F1-score) reveals:

- **High precision for class 1 (disease)** means few false disease predictions.

- **High recall for class 1** means the model correctly identifies most patients at risk.
- Balanced F1-scores indicate stable performance across both classes.

4.3 Confusion Matrix

Figure 2 — Confusion Matrix



Interpretation:

- The diagonal values represent correct predictions.
- Off-diagonal values (false positives/false negatives) are low, showing good model reliability.
- Misclassifications mainly occur where patient symptoms resemble borderline risk factors.

5. Discussion

Strengths

- The MLP captures nonlinear relationships common in medical data.
- Pipeline ensures proper preprocessing and prevents leakage.
- The model performs well even with mixed feature types.

Limitations

- Results depend heavily on dataset quality and balance.
- Neural networks lack interpretability compared to logistic regression or decision trees.
- 200 training iterations may not be fully optimal.
- One-hot encoding increases dimensionality, which may cause sparsity.

Potential Improvements

- Hyperparameter tuning (learning rate, hidden layers, neuron counts).
- Try alternative models: Random Forests, XGBoost, Logistic Regression.
- Use feature selection or PCA to simplify inputs.
- Add class balancing (SMOTE) if dataset becomes skewed.

---

6. Ethical Considerations

Machine-learning models in healthcare carry significant ethical responsibilities:

- **False negatives** (failing to detect disease) can delay treatment.
- **False positives** may cause unnecessary anxiety and tests.
- Models must be validated with real clinical data before real-world use.
- Predictions should support doctors—not replace them.
- Patient privacy and data security must be prioritized.

This experiment is educational and not intended for medical diagnosis.

---

7. Conclusion

This project demonstrates that a Multilayer Perceptron can effectively classify heart disease risk using structured patient data. Through preprocessing, one-hot encoding, feature scaling, and neural network modeling, we achieve strong performance and valuable insights into health patterns.

The assignment showcases key concepts in:

- Neural networks
- Data preprocessing

- Model evaluation
- Applied machine learning

The approach can be extended to other medical prediction tasks with appropriate validation.

References

- Scikit-learn Documentation — https://scikit-learn.org
- Bishop, C. (2006). *Pattern Recognition and Machine Learning*.
- UCI Heart Disease Dataset (inspiration for structure)
- Additional resources from course notes and lectures

Appendix

GitHub link : https://github.com/Palarajesh-dev/Machine-Learning-Assignment.git