HEART DISEASE PREDICTION SYSTEM USING MACHINE LEARNING AND STREAMLIT

Abstract

Heart disease is one of the leading causes of death worldwide. Early detection plays a crucial role in reducing mortality. This project implements a machine learning–based Heart Disease Prediction System that analyzes patient medical attributes and predicts the likelihood of heart disease. A Random Forest Classifier is used to build the predictive model and is deployed as a web application using Streamlit Community Cloud, providing real-time predictions with probability scores.

1. Introduction

Heart disease diagnosis traditionally requires multiple medical tests and expert evaluation. Many cases go undetected in early stages. With advancements in machine learning, predictive systems can analyze historical medical data and assist in early risk detection. This project aims to develop an automated and accessible heart disease prediction system.

1.1 Problem Statement

Manual diagnosis is time-consuming and depends heavily on expert availability. There is a need for an automated system that can predict heart disease risk using commonly available medical parameters.

1.2 Objectives

• To build a machine learning model for heart disease prediction.
• To analyze and visualize model performance.
• To deploy the model as a live web application.
• To provide an interactive user interface for predictions.

2. Dataset Description

The dataset used in this project is obtained from Kaggle. It contains patient medical records with attributes such as age, gender, blood pressure, cholesterol levels, ECG results, and other clinical parameters. The target variable indicates the presence or absence of heart disease.

3. System Architecture

The system architecture includes data preprocessing, model training, evaluation, and deployment. The trained model and scaler are saved using pickle and loaded into the Streamlit application for real-time prediction, ensuring efficiency and scalability.

4. Methodology

Data preprocessing involves removing duplicate records and splitting the dataset into training and testing sets. Feature scaling is performed using StandardScaler. A Random Forest Classifier is trained and evaluated using accuracy, classification report, ROC curve, and ROC-AUC score.

5. Model Evaluation

The model achieved good predictive performance. Evaluation metrics such as accuracy, classification report, ROC curve, and ROC-AUC score demonstrate the model’s reliability in distinguishing between high-risk and low-risk patients.

6. Feature Importance Analysis

Random Forest provides feature importance scores that help identify the most influential medical attributes contributing to predictions. This improves transparency and interpretability of the model.

7. Web Application Design

The Streamlit-based web application includes dataset preview, model performance metrics, ROC curve visualization, feature importance charts, and a user input module that allows users to enter medical parameters and receive instant predictions.

8. Deployment

The application is deployed using Streamlit Community Cloud. Deployment involves pushing code and trained model files to GitHub and configuring Streamlit Cloud to host the application publicly.

9. Technologies Used

Python, NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn, Streamlit, GitHub, and Streamlit Community Cloud.

10. Advantages

- Fast and accurate predictions.
- User-friendly web interface.
- No local installation required.
- Scalable and accessible deployment.

11. Limitations

- Depends on dataset quality.
- Not a substitute for professional medical diagnosis.
- Limited to predefined features.

12. Future Enhancements

- Integration with real-time medical data.
- Use of advanced deep learning models.
- Mobile-friendly application.

13. Conclusion

This project successfully demonstrates the application of machine learning and web technologies in healthcare prediction. By combining Random Forest classification with Streamlit deployment, the system provides a practical and accessible solution for early heart disease risk assessment while emphasizing responsible use.

14. References

- Kaggle Heart Disease Dataset.
- Scikit-learn Documentation.
- Streamlit Documentation.