Heart Disease Prediction Using Machine Learning

1. Introduction

Heart disease is one of the leading causes of death worldwide. Early detection and prediction of heart disease can help reduce health risks and improve patient care. With the rapid growth of data and machine learning techniques, predictive models can be developed to analyze medical data and support decision-making.

This project aims to build a **machine learning–based heart disease prediction system** that classifies whether a person is likely to have heart disease based on selected health-related features. The project also focuses on deploying the trained model using **Streamlit**, making it accessible through a web interface.

2. Objective of the Project

The main objectives of this project are:

- To identify whether the given problem is a **classification problem**
- To preprocess and clean the dataset
- To apply feature selection and feature scaling
- To train and evaluate machine learning models
- To deploy the final model using a Streamlit web application

3. Dataset Description

The dataset used in this project is a **Heart Disease dataset** containing patient health information. It includes both numerical and categorical features such as age, blood pressure, cholesterol level, BMI, sleep hours, and lifestyle-related attributes.

The target variable is **Heart Disease Status**, which indicates whether a patient has heart disease (True) or not (False). Since the output variable has two classes, this problem is identified as a **binary classification problem**.

4. Data Preprocessing

Before applying machine learning models, the dataset was carefully preprocessed.

First, the dataset was explored using functions like head(), info(), and describe() to understand the structure and data types. Missing values were identified using null checks. Necessary cleaning steps were performed to ensure data consistency.

Feature selection was then carried out to select the most relevant numerical features for prediction. This was done to simplify the model and make deployment easier. The selected features were:

- Age
- Blood Pressure
- Cholesterol Level
- BMI
- Sleep Hours

Since the selected features were numerical, no feature encoding was required. The target variable was encoded using **Label Encoding**.

5. Feature Scaling

Feature scaling was applied using **StandardScaler** to normalize the numerical features. Scaling ensures that all features contribute equally to the model and improves the performance of algorithms such as Logistic Regression.

6. Machine Learning Models

Two machine learning models were trained and evaluated:

6.1 Logistic Regression

Logistic Regression was used as a baseline model for binary classification. It is simple, interpretable, and effective for linear decision boundaries.

6.2 Random Forest Classifier

Random Forest, an ensemble learning method, was used to improve prediction performance. To handle class imbalance, the parameter class_weight="balanced" was applied. This ensures that both classes are treated fairly during training.

7. Model Evaluation

The models were evaluated using the following metrics:

- Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix

Although both models achieved similar accuracy, Random Forest was chosen as the final model because it handled class imbalance better and provided more reliable predictions.

8. Model Saving

The final trained model and the fitted scaler were saved using the **pickle** library. Saving the model allows it to be reused during deployment without retraining.

9. Deployment Using Streamlit

The trained model was deployed using **Streamlit**, a Python framework for building interactive web applications. The Streamlit app allows users to enter health parameters such as age, blood pressure, cholesterol level, BMI, and sleep hours.

The app processes user input, applies scaling, and predicts whether heart disease is detected. The application was deployed on **Streamlit Cloud**, making it publicly accessible via a web link.

10. Conclusion

This project demonstrates a complete machine learning pipeline, starting from data preprocessing and feature selection to model training, evaluation, and deployment. The Streamlit-based interface makes the model easy to use and understand.

The project highlights how machine learning can be applied in the healthcare domain to support early detection of heart disease and improve awareness.

11. Future Scope

The project can be enhanced by:

- Adding more clinical features

- Using advanced models such as XGBoost
- Improving model performance using hyperparameter tuning
- Deploying the model using APIs or mobile applications

12. Tools & Technologies Used

- Python
- Pandas, NumPy
- Scikit-learn
- Streamlit
- GitHub