PERSONALIZED HEALTHCARE RECOMMENDATION SYSTEM

Internship Project | Unified Mentor Pvt. Ltd.

Developed by: Mohd Isaar

Role: Data Analyst Intern

==========

1. INTRODUCTION

The Personalized Healthcare Recommendation System aims to leverage machine learning techniques to predict the likelihood of heart disease based on multiple clinical parameters.

This project uses a cleaned, merged dataset containing various cardiovascular health indicators, enabling predictive insights that can assist in early diagnosis and lifestyle recommendations.

The system integrates data preprocessing, exploratory data analysis, model training, evaluation, and deployment through a Streamlit web application for end-user interaction.

---

2. OBJECTIVES

• Analyze and interpret key features affecting cardiovascular health.

• Build a reliable machine learning model to predict the presence of heart disease.

• Provide a user-friendly, interactive prediction interface using Streamlit.

• Deliver actionable, data-driven healthcare insights and recommendations.

---

3. DATASET OVERVIEW

Dataset Name: cleaned_merged_heart_dataset.csv

Features:

age, sex, cp, trestbps, chol, fbs, restecg, thalachh, exang, oldpeak, slope, ca, thal, target

Target Variable:

1 = Heart disease present

0 = No heart disease

Each row represents an individual patient's health record, covering vital signs, test results, and other risk indicators.

---

4. TOOLS AND TECHNOLOGIES

Programming Language: Python

Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, joblib

Web Framework: Streamlit

Visualization Tools: Matplotlib, Seaborn

Model Storage: Joblib

IDE: Jupyter Notebook

---

5. STEP-BY-STEP IMPLEMENTATION

STEP 1 — Importing Libraries and Checking Few Rows

The necessary Python libraries were imported for data manipulation, visualization, and modeling.

Initial inspection of the dataset was done using df.head() and df.info() to verify data structure and integrity.

STEP 2 — Dataset Overview

The dataset includes several cardiovascular health indicators. Each record corresponds to a patient, with a binary target column indicating disease presence.

STEP 3 — Exploratory Data Analysis (EDA)

Objective: Identify important trends and correlations within the dataset.

Visualizations included:

- Target Distribution: Class balance between patients with and without heart disease.

- Correlation Heatmap: Highlights inter-feature relationships and key influencing factors.

Additional Visuals:

- Distribution of Age vs Target

- Boxplot of Cholesterol vs Target

- Barplot of Chest Pain Type vs Target

- Lineplot of Maximum Heart Rate vs Age

- Pairplot for key numerical attributes

STEP 4 — Data Preprocessing

Performed several key operations:

- Handling missing values and data inconsistencies.

- Feature scaling using StandardScaler.

- Separation of X (features) and y (target).

- Data split into training (80%) and testing (20%) sets.

STEP 5 — Model Selection and Training

Initially, multiple models were tested, including:

- Logistic Regression
- Random Forest Classifier
- Gradient Boosting

However, due to version incompatibility between Streamlit and scikit-learn, ensemble models faced deployment issues.

Therefore, the Logistic Regression model was finalized as the most stable and compatible choice, achieving approximately 72% accuracy after parameter tuning.

STEP 6 — Model Evaluation

The Logistic Regression model was evaluated on the test dataset using:

- Accuracy Score: 72%

- Precision, Recall, and F1-Score
- ROC-AUC Curve Visualization

The model provided a balanced trade-off between interpretability and predictive power, making it well-suited for deployment.

STEP 7 — Model Deployment Preparation

The finalized Logistic Regression model and its scaler were serialized using joblib.dump() for deployment in the Streamlit web application.

STEP 8 — Streamlit Application Setup

An interactive Streamlit app was developed allowing users to:

- Input health parameters manually.
- Instantly receive heart disease predictions.
- View personalized recommendations for preventive actions.

Application File: app.py

---

9. CONCLUSION

This project successfully delivers an end-to-end machine learning solution for predicting heart disease likelihood based on patient attributes.

The final model—Logistic Regression—was chosen for its compatibility, stability, and interpretability, offering a smooth integration with Streamlit's real-time interface.

Through structured EDA, robust preprocessing, and interactive deployment, this project demonstrates a practical, data-driven healthcare system that can support early diagnosis and personalized lifestyle recommendations.

Future enhancements could include:

- Integration of live patient monitoring data.
- Deployment using cloud-based services (AWS, Azure).
- Expanding predictions to other diseases or health risks.

10. PROJECT SUMMARY TABLE

Step	Phase	Description
1	Importing Libraries	Imported pandas, numpy, matplotlib, seaborn, sklearn
2	Data Overview	Loaded and understood dataset features
3	EDA	Visualized relationships between variables
4	Preprocessing	Scaled features and split data
5	Model Training	Compared models and finalized Logistic Regression
6	Evaluation	Assessed accuracy and performance metrics
7	Deployment	Saved model and integrated with Streamlit
8	Streamlit Setup	Built user interface for real-time prediction
9	Conclusion	Final summary and future enhancements

FINAL DETAILS

Final Accuracy: 72%

Model Used: Logistic Regression

Tools Used: Python, Streamlit, Scikit-learn, Pandas, Seaborn

Developer: Mohd Isaar — Data Analyst Intern, Unified Mentor Pvt. Ltd.

==========