Heart Disease Risk Assessment System

Final Project Report

1. Introduction

This project presents a **machine learning–based Heart Disease Risk Assessment System** that predicts the likelihood of heart disease using clinical patient data.

The system demonstrates the complete data science workflow, including data preprocessing, exploratory data analysis, feature selection, model training, hyperparameter tuning, and deployment through an interactive Streamlit application.

2. Dataset Description

The dataset used in this project was obtained from the **UCI Machine Learning Repository**. It contains clinical attributes related to heart health, including demographic, physiological, and diagnostic features.

3. Data Preprocessing & Exploratory Data Analysis

Data preprocessing was a crucial step to ensure model reliability and performance.

- Invalid values represented by '?' were replaced with missing values (NaN).
- The dataset was found to be largely complete, with only a small number of missing values in the ca and thal features.
- Numerical features were analyzed using **histograms** to observe distributions and skewness.
- A **correlation heatmap** was used to study relationships between variables and the target feature.
- **Boxplots** were applied to detect potential outliers.

A reusable preprocessing pipeline was built using **Scikit-learn**, where numerical features were imputed using the median and scaled using StandardScaler, while categorical features were imputed using the most frequent value and encoded using OneHotEncoder.

The output of this stage was a clean and transformed dataset ready for modeling.

4. Feature Selection

Feature selection was performed to identify the most influential attributes contributing to heart disease prediction.

A **Random Forest–based feature importance method** was used due to its ability to handle non-linear relationships and provide reliable importance scores for tabular medical data.

The analysis showed that features such as **ST slope**, **chest pain type**, and **maximum heart rate** were among the most significant predictors. Based on this analysis, a reduced dataset containing only the most relevant features was created and used for subsequent model training.

5. Model Training & Evaluation

Multiple supervised learning models were trained and evaluated to compare their predictive performance:

- Logistic Regression
- Random Forest
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)

Logistic Regression served as a baseline model. Random Forest demonstrated the best overall performance, offering strong accuracy and robustness. SVM achieved competitive results but required higher computational cost, while KNN was simpler to implement but less accurate.

Model evaluation was conducted using accuracy scores and confusion matrices to assess classification performance.

6. Hyperparameter Tuning

To further improve performance, **RandomizedSearchCV** was applied to all models for hyperparameter optimization.

After tuning, Random Forest achieved the highest accuracy, outperforming the other models. The tuning process provided valuable insights into model behavior and ensured a fair and optimized comparison across algorithms.

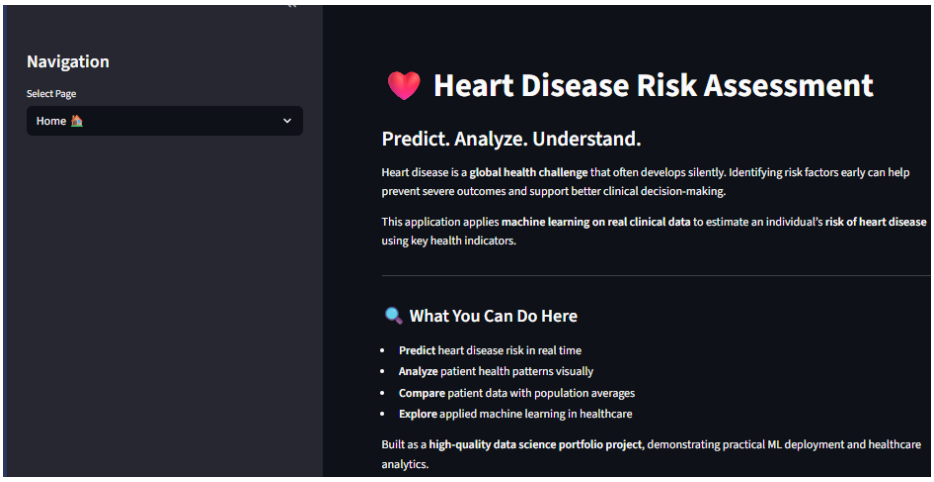
As a result, **Random Forest was selected as the final model** for deployment.

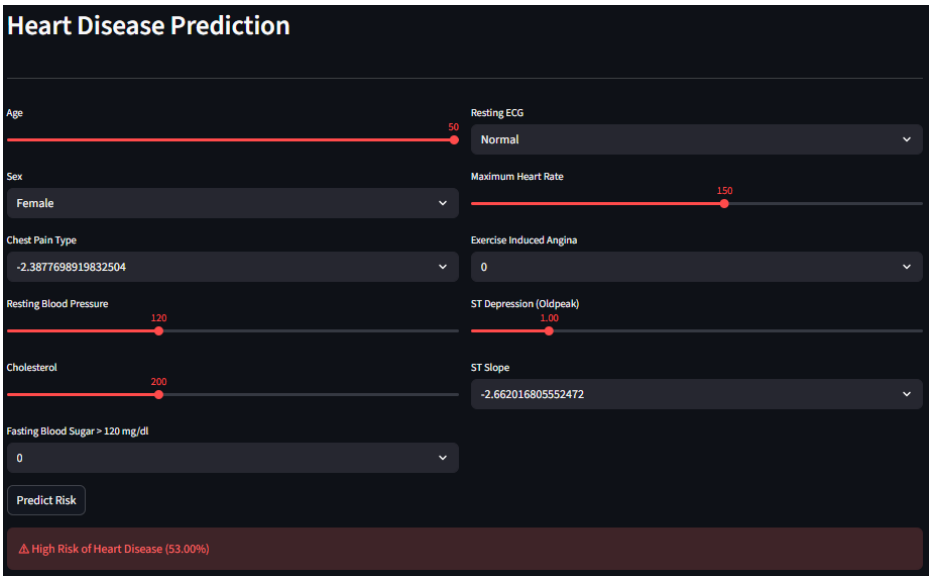
7. Deployment

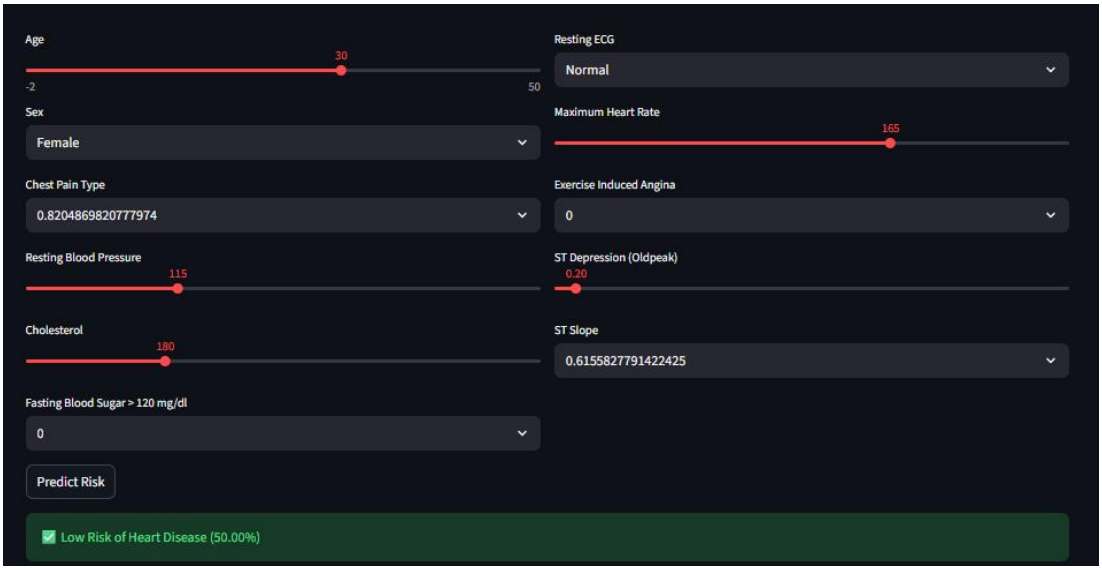
The final trained model was deployed using **Streamlit**, enabling users to:

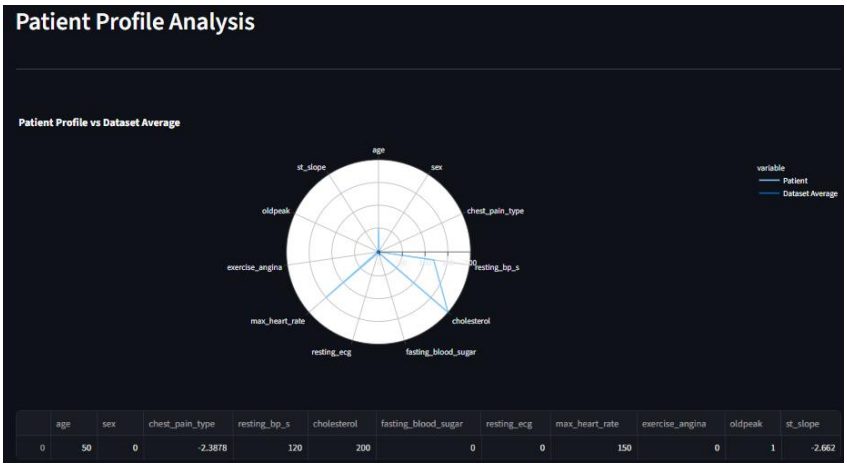
- Input patient health parameters
- Receive real-time heart disease risk predictions
- Visualize patient data in comparison to dataset averages

The application provides an interactive and user-friendly interface, demonstrating the practical application of machine learning in healthcare.









8. Conclusion

This project successfully demonstrates an end-to-end machine learning pipeline applied to a real-world healthcare problem. Through effective preprocessing, feature selection, model comparison, and hyperparameter tuning, a robust prediction system was developed and deployed.