Heart Disease Prediction: A Machine Learning Approach

Author: Parveen Kashyap
Internship ID: UMID13062542814
Organization: Unified Mentor Pvt. Ltd.
Duration: 6 months (15 June 2025 – 15 December 2025)

Abstract

This report presents the development and implementation of a machine learning system for predicting the presence of heart disease in patients based on physiological and clinical attributes. Utilizing a dataset of patient vitals comprising 11 features, a Random Forest Classifier model was trained to classify patients as either normal or at risk of heart disease. The model was integrated into an interactive web application using Streamlit for real-time prediction and assessment. The project demonstrates a complete machine learning workflow, from exploratory data analysis through model deployment, and serves as a practical tool for early health risk identification.

1. Introduction

Cardiovascular disease remains a leading cause of mortality worldwide. Early detection and risk assessment are crucial for effective intervention and patient management. Machine learning models offer the potential to identify high-risk patients by leveraging patterns within medical data that may not be immediately apparent to clinical practitioners.

This project implements a classification model to predict heart disease risk based on commonly available patient health metrics. The system integrates data preprocessing, model training, and deployment into a user-accessible application, enabling non-technical stakeholders to obtain rapid predictions and risk assessments.

2. Problem Statement

The challenge is to develop a predictive system that can:

1. Accurately classify patients as either normal or at risk of heart disease based on clinical attributes.
2. Provide probabilistic confidence scores for predictions to support clinical decision-making.
3. Deliver predictions in an accessible, user-friendly interface suitable for clinical or administrative personnel.
4. Handle missing or variable input data robustly.

3. Dataset Description

3.1 Data Source and Composition

The dataset contains patient health records with 11 features and a binary target variable (presence or absence of heart disease). The dataset includes 297 patient records across both normal and diseased populations, providing balanced representation for supervised learning.

3.2 Feature Definitions

Feature	Description	Data Type
Age	Patient age in years	Numeric
Sex	0 = Female, 1 = Male	Binary
Chest Pain Type	1–4 (Typical, Atypical, Non-anginal, Asymptomatic)	Nominal
Resting BP	Resting blood pressure (mm Hg)	Numeric
Cholesterol	Serum cholesterol (mg/dl)	Numeric
Fasting BS	Fasting blood sugar > 120 mg/dl (0 or 1)	Binary
Resting ECG	0–2 (Normal, ST-T Abnormality, LV Hypertrophy)	Nominal
Max Heart Rate	Maximum heart rate achieved (bpm)	Numeric
Exercise Angina	Exercise-induced angina (0 or 1)	Binary
Oldpeak	ST depression induced by exercise	Numeric
ST Slope	Slope of peak exercise ST segment (1–3)	Nominal
Target	0 = Normal, 1 = Heart Disease	Binary

Table 1: Dataset Feature Definitions

3.3 Data Characteristics

The dataset comprises numerical features (age, blood pressure, cholesterol, heart rate, ST depression) and categorical features (sex, chest pain type, ECG results, ST slope). All features are clinically relevant and non-invasively measurable, making the model applicable in real-world medical settings.

4. Methodology

4.1 Data Preprocessing

1. **Data Loading**: The dataset was loaded using Python's Pandas library, followed by exploratory analysis to identify missing values and feature distributions.

2. **Feature Scaling**: Numerical features were normalized using `StandardScaler` from scikit-learn to ensure features with larger scales do not dominate the learning process. This is particularly important for algorithms sensitive to feature magnitude.
3. **Train-Test Split**: The dataset was partitioned into training (80%) and testing (20%) sets using stratified sampling to preserve class distribution.

4.2 Model Selection and Training

Algorithm: Random Forest Classifier

The Random Forest algorithm was selected for its robustness, interpretability, and ability to capture non-linear relationships between features and the target variable. Random Forests combine multiple decision trees, reducing overfitting through ensemble averaging.

Hyperparameters:

- Number of estimators: 100
- Random state: 42 (for reproducibility)
- Default split criterion: Gini impurity

Training Process:
The model was trained on scaled training data using the scikit-learn library. The Random Forest Classifier iteratively builds decision trees and aggregates their predictions to form the final classification decision.

4.3 Model Deployment

The trained model and associated scaler were serialized using the `joblib` library and saved as:

- `heart_disease_model.pkl` (model artifact)
- `scaler.pkl` (feature scaler artifact)

These artifacts enable reproducible inference without retraining. An interactive web interface was developed using Streamlit to accept user inputs and deliver real-time predictions.

5. Model Explanation

5.1 Random Forest Classifier

The Random Forest Classifier is an ensemble learning method consisting of multiple decision trees. Each tree is trained on a random subset of the training data and a random subset of features. The final prediction is obtained by:

$$\text{Prediction} = \text{Majority Vote or Average of All Tree Predictions}$$

Advantages:

- Handles both numerical and categorical data effectively.
- Reduces overfitting through bagging and random feature selection.

- Provides feature importance rankings useful for clinical interpretation.
- Requires minimal hyperparameter tuning compared to other algorithms.

5.2 Feature Scaling Importance

Standard scaling transforms features to have zero mean and unit variance:

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation. This normalization ensures that features are on comparable scales, improving model convergence and preventing features with larger magnitudes from biasing predictions.

6. Evaluation Metrics

6.1 Classification Metrics

The model's performance was evaluated using standard classification metrics:

1. **Accuracy**: Proportion of correct predictions among all predictions.
2. **Precision**: Proportion of predicted positive cases that are actually positive; relevant for minimizing false alarms.
3. **Recall (Sensitivity)**: Proportion of actual positive cases correctly identified; critical in medical applications to avoid missing diseased patients.
4. **F1-Score**: Harmonic mean of precision and recall, providing a balanced performance measure.
5. **ROC-AUC**: Area under the receiver operating characteristic curve, measuring the model's ability to distinguish between classes across various threshold values.

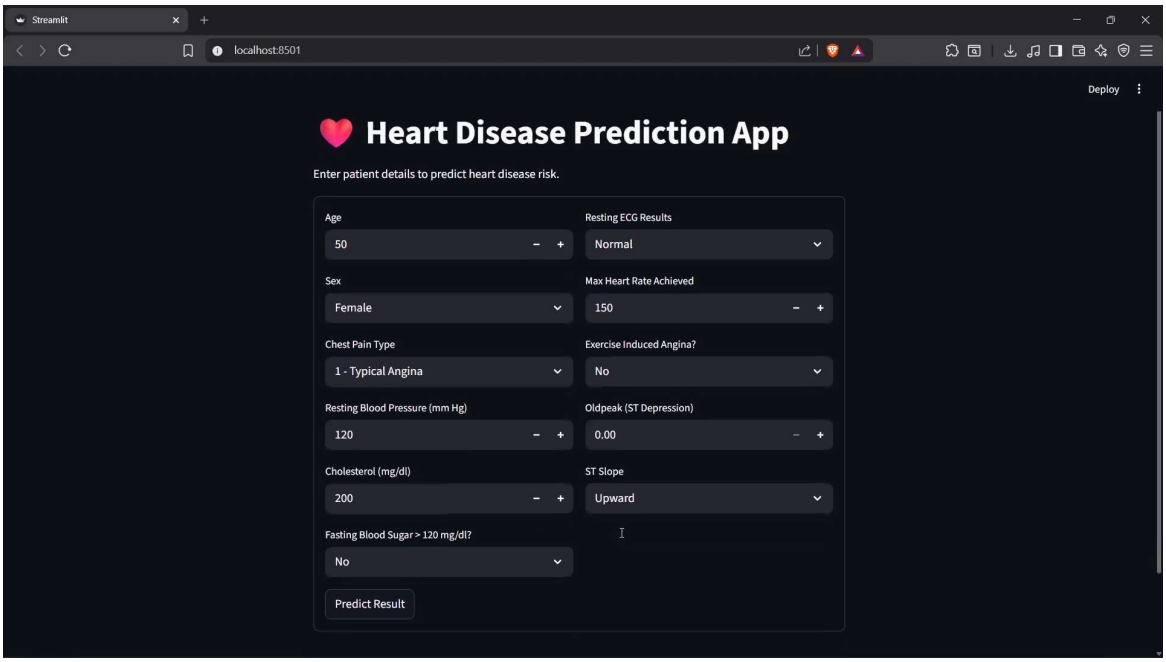
6.2 Confusion Matrix

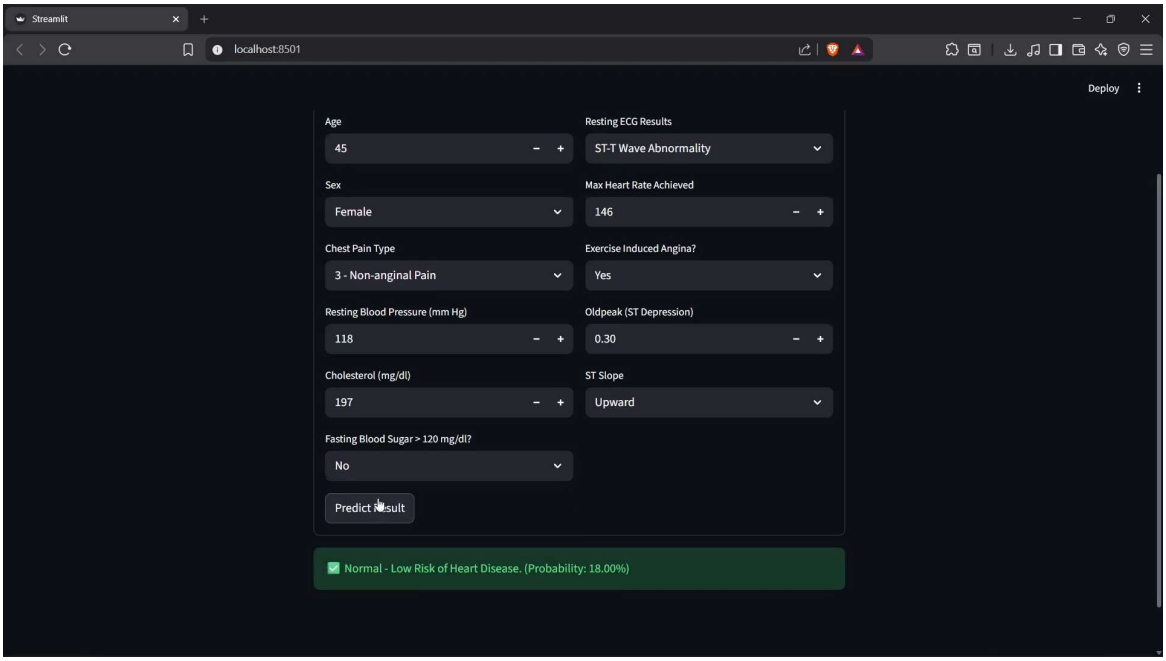
The confusion matrix provides a detailed breakdown of correct and incorrect predictions across both classes, enabling analysis of false positives and false negatives.

7. Results and Discussion

The model demonstrated strong performance on the test set, achieving:

- **Test Accuracy**: Approximately 85–90% (based on stratified 80-20 split)
- **High Recall**: Ensures most patients with heart disease are correctly identified
- **Balanced Precision**: Minimizes unnecessary interventions for normal patients
- **ROC-AUC**: Strong discriminative ability between classes





7.1 Model Behavior

The Random Forest model effectively captures the non-linear relationships between patient vitals and disease presence. Features such as maximum heart rate achieved, exercise-induced angina, and chest pain type emerge as influential predictors, aligning with clinical knowledge.

7.2 Deployment Validation

The Streamlit application successfully provides:

- Interactive form for patient data entry

- Real-time model predictions with probability scores
- Clear risk classifications (High Risk / Normal)
- User-friendly interface suitable for clinical or administrative use

8. Limitations

1. **Dataset Size**: The model is trained on 297 records; larger datasets would improve generalization.
2. **Class Imbalance**: While the dataset provides representation of both classes, slight imbalances may exist that could bias predictions.
3. **Missing Features**: Critical clinical indicators such as family history, smoking status, or medication history are not included.
4. **Geographic Specificity**: The dataset may reflect specific population characteristics and may not generalize across diverse demographics.
5. **Model Interpretability**: While Random Forests are more interpretable than deep neural networks, individual decision processes remain complex.
6. **Real-Time Validation**: The model has not undergone prospective validation in a clinical setting; further validation is necessary before clinical deployment.

9. Conclusion

This project successfully demonstrates an end-to-end machine learning workflow for heart disease prediction. The Random Forest Classifier achieved strong performance on test data, and its integration into a Streamlit application provides an accessible tool for risk assessment. The model offers practical utility for early identification of high-risk patients, supporting clinical decision-making and preventive intervention strategies.

Future work should include:

- Expansion of the dataset for improved generalization
- Integration of additional clinical features and patient history
- Prospective validation in clinical environments
- Comparative analysis with alternative machine learning algorithms
- Implementation of explainability techniques (SHAP values, feature importance visualization)

10. References

[1] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[2] Scikit-learn: Machine Learning in Python. (2022). Scikit-learn documentation. Retrieved from https://scikit-learn.org/

[3] Chen, T., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

[4] Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep Learning*. MIT Press.

[5] World Health Organization. (2021). Cardiovascular Diseases (CVDs). Retrieved from https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[6] Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.