G H Patel College of Engineering & Technology

(The Charutar Vidhya Mandal (CVM) University)
Bakrol, Anand

COMPUTER ENGINEERING DEPARTMENT

Project Report
On
Heart Disease Prediction

Submitted By

Name of Student: Varad Chaudhari
Enrollment Number: 12202130501062

Name of Student: Vivek Vallabhan
Enrollment Number: 12202130501065

Artificial Intelligence & Machine Learning (202046702)
A.Y. 2024-25 EVEN TERM

1. Objective

The objective of this project is to develop a predictive model that can determine whether a patient is likely to have heart disease based on their medical features. The main goals include:

- Building an accurate classification model using machine learning techniques.
- Evaluating and comparing the performance of multiple algorithms.
- Providing probabilistic output to support clinical decision-making.
- Implementing the model in a scalable, reusable script for real-time prediction. Heart disease remains a leading cause of mortality worldwide, and this tool aims to assist healthcare professionals by offering quick and reliable decision support.

.

2. Dataset Used

The dataset used in this project is the **Heart Disease Dataset** sourced from the **UCI Machine Learning Repository**. It is a well-known dataset in the data science community for binary classification problems.

- **Total Samples**: 1025
- **Features**: 13 independent variables (clinical features)
- **Target**: 1 (Presence of heart disease), 0 (Absence of heart disease)

➤ Key Features:

- **Age**: Age of the patient
- **Sex**: Gender (1 = Male, 0 = Female)
- **cp**: Chest pain type (0–3)
- **trestbps**: Resting blood pressure (in mm Hg)
- **chol**: Serum cholesterol in mg/dl
- **fbs**: Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
- **restecg**: Resting electrocardiographic results (values 0,1,2)
- **thalach**: Maximum heart rate achieved
- **exang**: Exercise induced angina (1 = yes; 0 = no)
- **oldpeak**: ST depression induced by exercise relative to rest
- **slope**: Slope of the peak exercise ST segment
- **ca**: Number of major vessels colored by fluoroscopy (0–3)
- **thal**: 3 = normal; 6 = fixed defect; 7 = reversible defect

3. Model Chosen

To identify the best-performing model, multiple machine learning algorithms were evaluated:

- **Logistic Regression** – A baseline linear model for binary classification.
- **Random Forest Classifier** – An ensemble model with decision trees for robust predictions.
- **Support Vector Machine (SVM)** – A powerful classifier that finds the optimal hyperplane for separation.

After thorough evaluation, **SVM (Support Vector Machine)** was chosen as the final model due to its excellent balance between precision, recall, and overall accuracy.

4. Performance Metrics

Each model was evaluated using the following metrics:

- **Accuracy**: Overall correctness of the model.
- **Precision**: Percentage of correctly predicted positive cases.
- **Recall**: Ability to detect all actual positive cases.
- **F1 Score**: Harmonic mean of precision and recall.
- **ROC AUC Score**: Area under the Receiver Operating Characteristic curve.

Comparison Table:

Metric	Logistic Regression	Random Forest	SVM (Chosen)
Accuracy	80.98%	92.68%	**92.20%**
Precision	76.19%	89.47%	**90.83%**
Recall	91.43%	97.14%	**94.29%**
F1 Score	83.12%	93.15%	**92.52%**
ROC AUC Score	92.97%	97.61%	**97.71%**

5. Challenges & Learnings

Challenges:

- **Model Selection**: Choosing a model that generalized well without overfitting required multiple experiments.
- **Data Imbalance Consideration**: Even though the dataset was relatively balanced, careful monitoring of precision and recall was necessary to avoid bias.
- **Evaluation Beyond Accuracy**: Initially, accuracy appeared high for all models. A deeper dive into recall and AUC scores revealed actual model robustness.

Learnings:

- Understood the critical importance of **data preprocessing**, especially when working with medical data.
- Gained practical experience in **model evaluation using multiple metrics**, beyond just accuracy.
- Learned how to **compare multiple classifiers** and select one based on both performance and interpretability.
- Built a modular and reusable script (predict.py) to **simulate real-world predictions**.
- Understood the **end-to-end workflow of a ML project**, from data loading to deployment-ready model saving.