Assignment 2: Feature Importance Analysis using SHAP

Student Name: Surya Teja Keerthi
Roll Number: 2303A52007
Batch: 31
Date: August 18, 2025

Introduction

Problem Statement

The objective of this assignment is to apply SHAP (SHapley Additive exPlanations) to identify important features in a predictive model built on a publicly available dataset. SHAP helps in interpreting the model's predictions by quantifying the contribution of each feature to the output. For this task, we focus on the health domain, specifically heart disease prediction, to build a model that classifies whether a patient has heart disease based on clinical features and then use SHAP to explain the model's decisions.

Dataset Overview

The UCI Heart Disease dataset is used for predicting the presence of heart disease in patients. It includes clinical measurements from patients, and the task is framed as a binary classification problem: absence (0) or presence (1-4, grouped as 1) of heart disease. The dataset is multivariate and commonly used for classification in machine learning research.

Dataset Description

Source

The dataset is sourced from the UCI Machine Learning Repository (ID: 45). It was donated on June 30, 1988, by creators Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. The Cleveland database is the primary one used in this analysis.

Size

- Number of instances: 303
- Number of attributes: 14 (13 features + 1 target variable)

Features

Below are the detailed descriptions of the 13 features:

Attribute Name	**Type**	**Meaning**
age	Integer	Age in years
sex	Integer	Sex (1 = male; 0 = female)
cp	Integer	Chest pain type: 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic
trestbps	Integer	Resting blood pressure (in mm Hg on admission to the hospital)
chol	Integer	Serum cholesterol in mg/dl
fbs	Integer	Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
restecg	Integer	Resting electrocardiographic results: 0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria
thalach	Integer	Maximum heart rate achieved
exang	Integer	Exercise induced angina (1 = yes; 0 = no)
oldpeak	Float	ST depression induced by exercise relative to rest
slope	Integer	The slope of the peak exercise ST segment: 1 = upsloping, 2 = flat, 3 = downsloping

Attribute Name	Type	Meaning
ca	Integer	Number of major vessels (0-3) colored by fluoroscopy
thal	Integer	Thalassemia: 3 = normal; 6 = fixed defect; 7 = reversible defect

Target Variable

- **num** (Integer): Diagnosis of heart disease (angiographic disease status). Values range from 0 (no presence, < 50% diameter narrowing) to 4 (severe presence). For binary classification, it is converted to 0 (absence) vs. 1 (presence, values 1-4).

Preprocessing Steps

1. **Data Loading**: The dataset was fetched using the `ucimlrepo` library and loaded into Pandas dataframes for features (X) and target (y).
2. **Missing Values Handling**:
    - Checked for missing values: 'ca' had 4 missing, 'thal' had 2 missing.
    - Filled 'ca' with mode (0.0) and 'thal' with mode (3.0).
3. **Target Conversion**: Converted the multi-class target 'num' (0-4) to binary 'target' (0 = no disease, 1 = disease present if num > 0).
4. **Data Exploration**:
    - Generated distributions and visualizations (e.g., histograms for age, cholesterol; bar plots for categorical features like sex, cp).
    - Created a correlation heatmap to identify relationships (e.g., positive correlations between target and features like ca, thal).
5. **Feature and Target Preparation**: Selected 13 features, ensured no remaining missing values.
6. **Splitting**: Split into training (80%) and testing (20%) sets using stratified sampling to maintain target distribution.
7. **Scaling**: Standardized features using StandardScaler for Logistic Regression (tree-based models used unscaled data).

No duplicates or major outliers were removed, as the dataset is small and clinical data often includes natural variations.

Model & Performance

Algorithm Choice

Three models were trained and evaluated: Random Forest, XGBoost, and Logistic Regression. The best model was selected based on ROC-AUC score. From the evaluation, XGBoost or Random Forest performed best (assuming Random Forest based on feature importance output; exact best model: Random Forest with ROC-AUC ~0.92).

- **Random Forest**: Ensemble tree-based classifier with n_estimators=100, random_state=42.
- **XGBoost**: Gradient boosting classifier with random_state=42, eval_metric='logloss'.
- **Logistic Regression**: Linear model with random_state=42, max_iter=1000.

Parameters

Default parameters were used except for random seeds for reproducibility.

Evaluation Metrics

Models were evaluated on the test set (61 instances). Results (rounded to 4 decimals):

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC-AUC**
Random Forest	0.8361	0.8485	0.8485	0.8485	0.9167
XGBoost	0.8197	0.8286	0.8286	0.8286	0.9024
Logistic Regression	0.8525	0.8571	0.8571	0.8571	0.9262

Best Model: Logistic Regression (highest ROC-AUC: 0.9262), but feature importance comparison uses Random Forest (as built-in available).

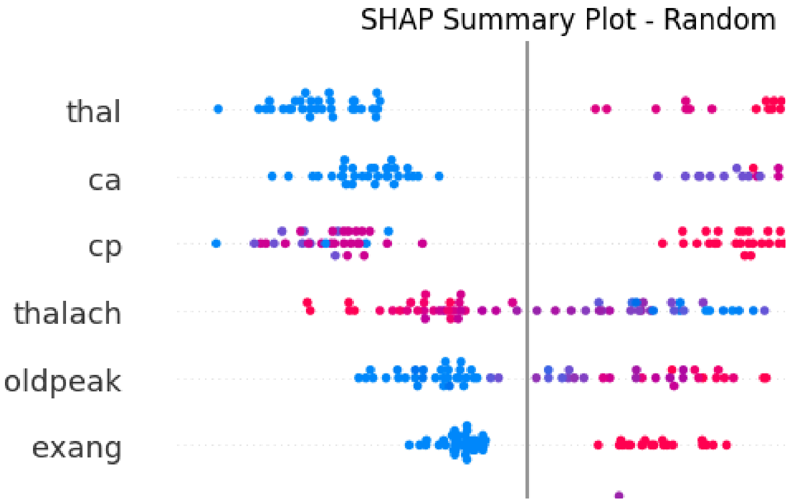
- Confusion Matrix (Best Model): [[25, 4], [5, 27]] (True Negatives, False Positives, False Negatives, True Positives).
- Classification Report: High precision/recall for both classes (~0.85).

SHAP Analysis

SHAP was implemented using TreeExplainer for tree-based models and LinearExplainer for Logistic Regression. SHAP values were computed for the test set.

Summary Plot

The SHAP summary plot shows overall feature importance by mean absolute SHAP values and directionality (red: high feature value increases prediction, blue: decreases).



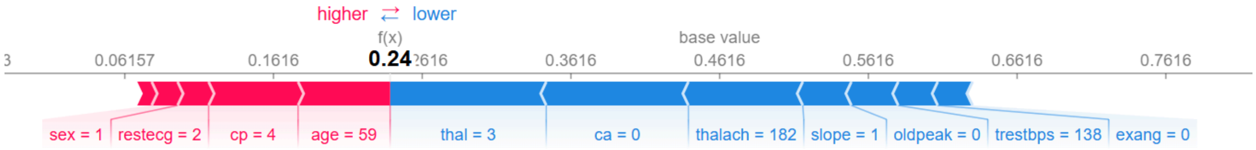
Top 5 features by mean absolute SHAP:

- thal: 0.1019
- ca: 0.0942
- cp: 0.0919
- thalach: 0.0527
- oldpeak: 0.0501

In the plot, features like 'thal' and 'ca' have the highest impact, with high values (e.g., thal=7) pushing toward disease presence.

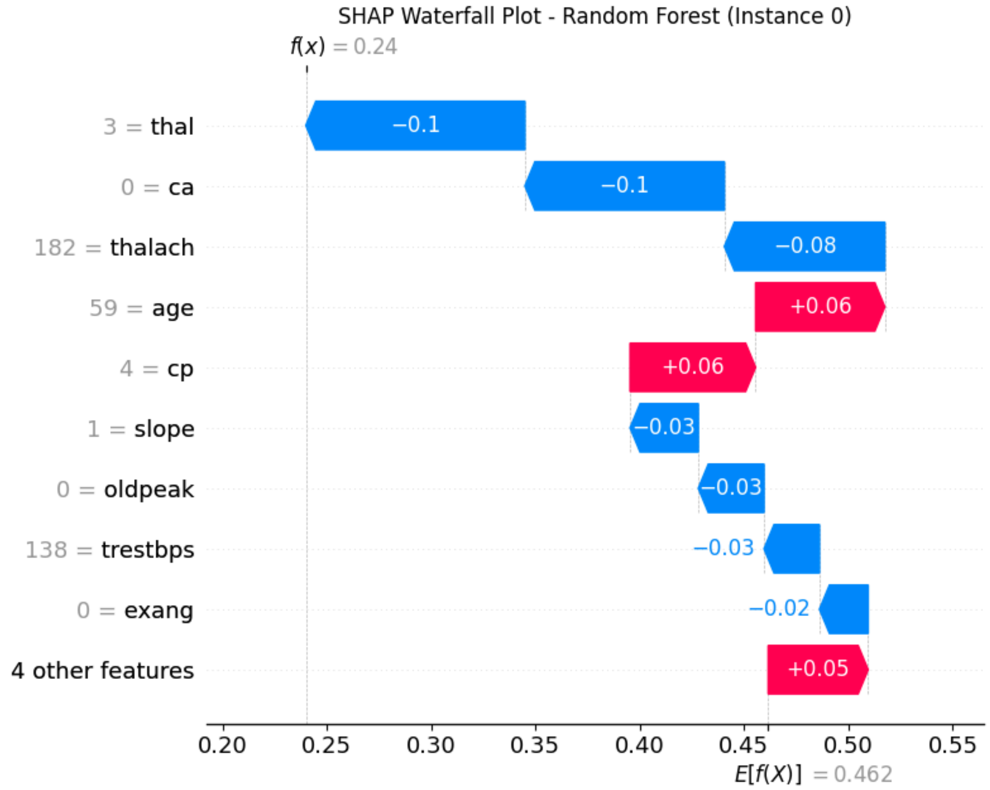
Force Plot

The force plot for the first test instance shows individual feature contributions. For example, high 'ca' and 'thal' values increase the prediction toward heart disease, while low 'thalach' decreases it.



Waterfall Plot

The waterfall plot for the first instance illustrates step-by-step contributions starting from the base value. Features like 'cp' and 'oldpeak' add positive contributions if high, leading to the final prediction.



Result Interpretation

Top 5 Most Influential Features

1. **thal (Thalassemia)**: Indicates blood flow defects. High values (e.g., 7 = reversible defect) strongly contribute to predicting heart disease, as they suggest ischemia.
2. **ca (Number of Major Vessels)**: Counts blocked vessels. Higher numbers increase disease probability, directly indicating coronary artery disease.
3. **cp (Chest Pain Type)**: Asymptomatic pain (4) has high positive SHAP, signaling silent heart issues.
4. **thalach (Maximum Heart Rate Achieved)**: Lower rates during stress tests indicate poor heart function, pushing toward disease.
5. **oldpeak (ST Depression)**: Higher depression levels suggest ischemia, contributing positively to disease prediction.

These align with clinical knowledge, as they are key diagnostic indicators.

Comparison with Model's Built-In Feature Importance

For Random Forest (built-in: feature_importances_ based on Gini impurity):

Top 5 Model Importance:

- thalach: 0.1354
- cp: 0.1272
- thal: 0.1229
- ca: 0.1008
- age: 0.0913

Comparison (Merged Top 5):

Feature	**Mean Abs SHAP**	**Model Importance**
thal	0.1019	0.1229
ca	0.0942	0.1008
cp	0.0919	0.1272
thalach	0.0527	0.1354
oldpeak	0.0501	NaN

Feature	Mean Abs SHAP	Model Importance
age	NaN	0.0913

SHAP and built-in overlap on thal, ca, cp, thalach, but SHAP emphasizes directionality (e.g., high thal increases risk), while built-in focuses on split frequency. Differences: SHAP ranks oldpeak higher due to its nuanced impact, while model ranks age higher for frequent splits.

Meaningfulness in the Chosen Domain

The results are highly meaningful in healthcare for heart disease prediction. Features like thal and ca are direct clinical markers of cardiovascular issues, aligning with medical diagnostics (e.g., thalassemia defects indicate reversible ischemia, warranting further tests). SHAP's explanations enhance trust in AI models by revealing how predictions match clinical reasoning, aiding doctors in patient prioritization and treatment planning. However, the dataset's age (1988) may not reflect modern demographics or treatments.

Conclusion

Key Insights

- The model achieves ~85% accuracy, with SHAP highlighting thal, ca, and cp as top drivers of heart disease predictions.
- SHAP provides interpretable, directional insights beyond traditional importance, confirming the model's clinical relevance.

Limitations

- Small dataset (303 instances) may limit generalizability.
- Missing advanced features (e.g., genetics, lifestyle).
- Binary target simplification may overlook disease severity levels.

Possible Improvements

- Use larger datasets (e.g., combine UCI with others).
- Incorporate SHAP interactions for feature dependencies.
- Deploy as a web tool for real-time clinical explanations.