Heart Disease Analysis: Understanding with SHAP and Random Forest

Introduction

Heart diseases, particularly heart attacks, are a significant health concern globally. Predicting the likelihood of heart disease is crucial for early intervention and prevention. Machine learning techniques, including Random Forest Classifier, have proven to be effective in this domain. However, understanding and interpreting the complex decisions made by these models are equally important. In this report, we explore the interpretability of a Random Forest Classifier applied to a heart disease prediction dataset.

Dataset Overview

The dataset under consideration contains several key attributes related to patients' health. Here is a brief overview of the attributes:

- **Age:** Age of the patient in years.
- **Sex:** Gender of the patient, denoted as M (Male) or F (Female).
- **ChestPain:** Type of chest pain experienced by the patient, categorized as TA (Typical Angina), ATA (Atypical Angina), NAP (Non-Anginal Pain), or ASY (Asymptomatic).
- **RestingBP:** Resting blood pressure measured in mm Hg.
- **Cholesterol:** Serum cholesterol level in mm/dl.
- **FastingBS:** Fasting blood sugar, where 1 indicates a value greater than 120 mg/dl, and 0 indicates otherwise.
- **RestingECG:** Resting electrocardiogram results, classified as Normal, ST (indicating ST-T wave abnormality), or LVH (showing probable or definite left ventricular hypertrophy by Estes' criteria).
- **MaxHR:** Maximum heart rate achieved during an exercise session.
- **ExerciseAngina:** Exercise-induced angina, denoted as Y (Yes) or N (No).
- **Oldpeak:** ST depression induced by exercise, measured in depression units.
- **ST_Slope:** The slope of the peak exercise ST segment, classified as Up (up-sloping), Flat (flat), or Down (down-sloping).
- **HeartDisease:** The target variable, indicating the presence (1) or absence (0) of heart disease.

Objective of the Report

In this report, our primary objective is to dive into the interpretability of the Random Forest Classifier's predictions. To achieve this, we will utilize SHAP (SHapley Additive exPlanations)

values, a powerful technique for explaining machine learning models. SHAP values provide detailed insights into the impact of each feature on individual predictions.

We aim to visualize these SHAP explanations through a variety of plots and analyses. By employing Random Forest Classifier, we will generate a detailed classification report, including metrics such as accuracy, precision, recall, and F1-score. These metrics will serve as a basis for evaluating our model's performance.

Through this comprehensive analysis, we strive to enhance our understanding of how specific attributes influence heart disease predictions. The report will provide valuable insights not only into the predictive capabilities of the model but also into the medical significance of the contributing factors.

Methodology

In this study, we followed a systematic approach to develop and interpret the heart disease prediction model using a Random Forest Classifier and SHAP values. Below are the detailed steps of our methodology:

1. Data Pre-processing:

- The dataset was loaded from the CSV file, containing various attributes related to patients' health, including age, gender, chest pain type, blood pressure, cholesterol levels, and other factors.

- Categorical variables (Sex, ChestPainType, FastingBS, RestingECG, ExerciseAngina, ST_Slope) were one-hot encoded to convert them into a format suitable for machine learning algorithms.

- The data was split into features (X) and the target variable (y), where X represents the input features, and y represents the HeartDisease status (1 for heart disease presence, 0 for absence).

2. Data Splitting:

- The dataset was divided into training and testing sets using the train_test_split function from the sklearn.model_selection module. The training set (80% of the data) was used for model training, while the testing set (20% of the data) was used for evaluation and validation.

3. Random Forest Classifier:

- A Random Forest Classifier, a robust ensemble learning algorithm, was chosen for building the heart disease prediction model. Random Forests are capable of handling complex relationships in data and are less prone to overfitting.

In machine learning models, hyper parameters are external configurations that cannot be learned from the data. Instead, they need to be set prior to training the model. The choice of

hyper parameters significantly affects the model's performance. Grid Search is a technique used to find the best combination of hyper parameters for a machine learning algorithm.

In the context of our Random Forest Classifier, the following hyper parameters were considered for optimization:

Number of Estimators: It represents the number of decision trees in the random forest. Increasing the number of estimators can improve the model's performance but might lead to longer training times. The grid search explores different values for this hyper parameter to find the optimal one.

Maximum Depth: It defines the maximum depth of each decision tree in the forest. Deeper trees can capture more complex patterns in the data, but they might over fit. Grid Search tests various depth values (including None, which allows nodes to expand until they contain fewer than the minimum samples split) to identify the best depth.

Minimum Samples Split: It is the minimum number of samples required to split an internal node. A node will only split if it has more than the minimum samples split. This hyper parameter prevents creating nodes that only fit a small number of samples, potentially avoiding overfitting. Grid Search evaluates different split values to find the optimal one.

Minimum Samples Leaf: It is the minimum number of samples a leaf node should have. Similar to minimum samples split, it prevents creating leaves with very few samples. Grid Search explores different leaf sample values to determine the optimal setting.

Hyper parameter optimization is crucial because it ensures that the model is fine-tuned to the specific dataset, leading to better generalization and predictive accuracy. Grid Search exhaustively tests various combinations of hyper parameters, evaluating the model's performance using cross-validation. This means the dataset is divided into multiple subsets, and the model is trained and evaluated multiple times, ensuring robustness in the results.

4. Model Training and Evaluation:

The Random Forest Classifier was trained on the training data using the best hyper parameters obtained from the grid search.

The trained model was evaluated on the test data, and various metrics such as accuracy, precision, recall, and F1-score were computed using the classification_report function from sklearn.metrics. Additionally, the confusion matrix was generated to understand the model's performance in detail.

5. SHAP Explanations:

- SHAP (Shapley Additive explanations) values were employed to interpret the predictions made by the Random Forest Classifier. SHAP values provide insights into the impact of individual features on the model's output.

- We utilized the SHAP library to calculate SHAP values for the model predictions. Various SHAP plots, including summary plots, waterfall plots, and dependency plots, were generated to visualize and understand the contributions of different features to individual predictions.

6. Result Analysis:

- The best-performing Random Forest Classifier was identified based on the optimized hyper parameters obtained from the grid search.

- The model's accuracy and classification report were analysed to evaluate its predictive performance. Detailed insights from SHAP plots were used to interpret the significance of different features in predicting heart disease.

By following this methodology, we aimed to develop an accurate heart disease prediction model while gaining valuable insights into the factors influencing the predictions.

Results:

```
Fitting 5 folds for each of 81 candidates, totalling 405 fits
Best Parameters: {'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 200}
Accuracy: 0.8695652173913043
Random Forest Classification Report:
                  precision    recall  f1-score   support

No Heart Disease       0.84      0.84      0.84        77
   Heart Disease       0.89      0.89      0.89       107

        accuracy                           0.87       184
       macro avg       0.87      0.87      0.87       184
    weighted avg       0.87      0.87      0.87       184

Confusion Matrix:
[[65 12]
 [12 95]]
```

The Random Forest Classifier was optimized, considering various combinations of hyperparameters through Grid Search. After an intensive search, the optimal hyperparameters were identified as follows:

- Maximum Depth: 10

- Minimum Samples Leaf: 2

- Minimum Samples Split: 2

- Number of Estimators: 200

With this configuration, the model achieved the following performance metrics:

- **Accuracy:** 87.0%

The accuracy metric indicates that the model correctly classified 87.0% of the total instances in the test dataset.

- Precision:

- No Heart Disease (0): 84%

- Heart Disease (1): 89%

Precision measures the proportion of true positive predictions among all positive predictions. For No Heart Disease predictions, the precision was 84%, meaning that when the model predicted No Heart Disease, it was correct 84% of the time. Similarly, for Heart Disease predictions, the precision was 89%.

- Recall (Sensitivity):

- No Heart Disease (0): 84%

- Heart Disease (1): 89%

Recall, or sensitivity, quantifies the proportion of true positive instances that were correctly identified by the model. For No Heart Disease, the recall was 84%, indicating that the model captured 84% of all instances with No Heart Disease. Similarly, for 'Heart Disease', the recall was 89%.

- F1-Score:

- No Heart Disease (0): 84%

- Heart Disease (1): 89%

The F1-score is the harmonic mean of precision and recall. It provides a balance between precision and recall. For No Heart Disease and Heart Disease predictions, the F1-scores were 84% and 89%, respectively.

Confusion Matrix:

The confusion matrix provides a detailed breakdown of correct and incorrect predictions. In this case:

- True Positive (TP): 95 instances of Heart Disease were correctly predicted.

- True Negative (TN): 65 instances of No Heart Disease were correctly predicted.

- False Positive (FP): 12 instances were wrongly predicted as 'Heart Disease'.

- False Negative (FN): 12 instances were wrongly predicted as 'No Heart Disease'.

These metrics collectively demonstrate the Random Forest Classifier's ability to effectively discern between individuals with and without heart disease, providing a reliable foundation for further analysis.

SHAP Summary Plot Analysis:

Detailed SHAP Values Analysis:

```
Feature: Age, SHAP Value: 0.09056227082234508
Feature: RestingBP, SHAP Value: 0.0020495936730743703
Feature: Cholesterol, SHAP Value: -0.01850773416837079
Feature: MaxHR, SHAP Value: -0.0382604380372083
Feature: Oldpeak, SHAP Value: -0.045482941550367945
Feature: Sex_M, SHAP Value: -0.05472032174308313
Feature: ChestPainType_ATA, SHAP Value: -0.06119756751812558
Feature: ChestPainType_NAP, SHAP Value: 0.01319620798167503
Feature: ChestPainType_TA, SHAP Value: 0.0009549339657378859
Feature: FastingBS_1, SHAP Value: -0.015558084754208895
Feature: RestingECG_Normal, SHAP Value: -0.0067931714284375265
Feature: RestingECG_ST, SHAP Value: 0.0006171129968626948
Feature: ExerciseAngina_Y, SHAP Value: -0.07254606873125374
Feature: ST_Slope_Flat, SHAP Value: -0.0837994486784987
Feature: ST_Slope_Up, SHAP Value: -0.11011357716510878
```

In our analysis of the, we delved into the specific features influencing the prediction. The SHAP values, representing the impact of each feature on the model's decision, provided crucial insights. Here's a breakdown of the features and their corresponding SHAP values for the chosen instance:

- Age: The age of the individual has a positive impact (SHAP Value: 0.0906), indicating that older age contributes to a higher likelihood of heart disease.

- Resting Blood Pressure (RestingBP): With a minimal impact (SHAP Value: 0.0020), resting blood pressure plays a relatively insignificant role in this prediction.

- Serum Cholesterol (Cholesterol): Interestingly, cholesterol levels exhibit a negative impact (SHAP Value: -0.0185), suggesting that higher cholesterol levels slightly reduce the probability of heart disease in this instance.

- Maximum Heart Rate (MaxHR): A lower maximum heart rate has a negative influence (SHAP Value: -0.0383), indicating that individuals with higher maximum heart rates are less likely to have heart disease.

- ST Depression Induced by Exercise Relative to Rest (Oldpeak): Oldpeak has a negative impact SHAP Value: -0.0455), implying that higher ST (depression during exercise reduces the probability of heart disease.

- Gender (Sex_M): Being male negatively affects the prediction (SHAP Value: -0.0547), indicating that females are more likely to be classified with heart disease in this scenario.

- Chest Pain Type (ChestPainType_ATA): This specific type of chest pain has a significant negative impact (SHAP Value: -0.0612), suggesting that individuals experiencing this type of pain are less likely to have heart disease.

- Chest Pain Type (ChestPainType_NAP): Another chest pain type, although less impactful, has a positive influence (SHAP Value: 0.0132) on the prediction.

- Chest Pain Type (ChestPainType_TA): This type of chest pain has a negligible positive impact (SHAP Value: 0.0009) on the probability of heart disease.

- Fasting Blood Sugar (FastingBS_1): Fasting blood sugar levels slightly reduce the likelihood of heart disease (SHAP Value: -0.0156).

- Resting Electrocardiogram Result (RestingECG_Normal): A normal resting electrocardiogram result has a small negative impact (SHAP Value: -0.0068) on the prediction.

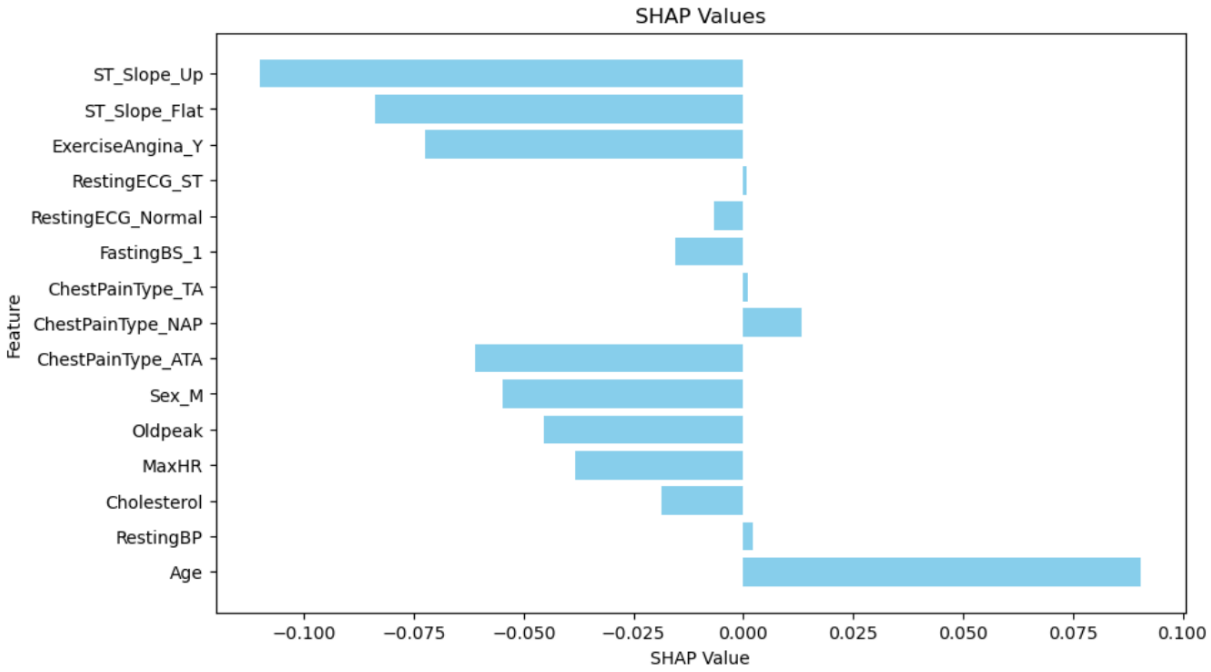
- Resting Electrocardiogram Result (RestingECG_ST): Presence of an abnormal resting electrocardiogram has a minimal positive impact (SHAP Value: 0.0006) on the likelihood of heart disease.

- Exercise-Induced Angina (ExerciseAngina_Y): The presence of exercise-induced angina substantially reduces the probability of heart disease (SHAP Value: -0.0725).

- Slope of the Peak Exercise ST Segment (ST_Slope_Flat): A flat slope has a strong negative impact (SHAP Value: -0.0838), indicating individuals with this characteristic are less likely to have heart disease.

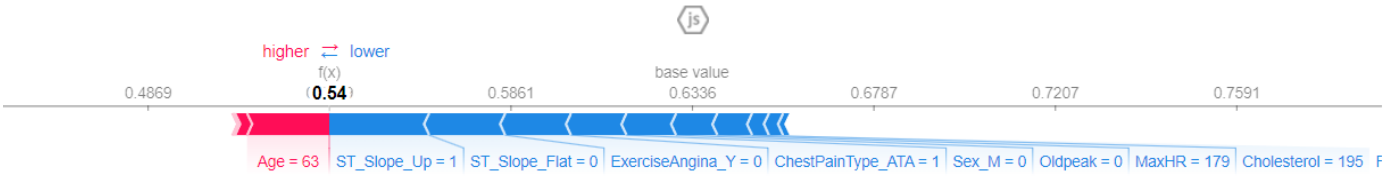
- Slope of the Peak Exercise ST Segment (ST_Slope_Up): An upward-sloping ST segment significantly reduces the probability of heart disease (SHAP Value: -0.1101).

The accompanying bar plot visually represents these SHAP values, providing a clear illustration of the features and their impacts on the prediction. This detailed analysis enhances our understanding of the model's decision.



SHAP Force Plot Analysis

The SHAP force plot offers a insight into the factors that influence the model's prediction for an individual instance. In the context of our analysis, the plot is centered around a pivotal point marked at 0.54, representing the model's average predicted probability of 'Heart Disease' across the dataset.



Positive Contributions (Left of 0.54):

To the left of 0.54, the red bar illustrates positive contributions that increase the probability of predicting Heart Disease. Among these, Age stands out as the most significant contributor. This implies that advanced age substantially raises the likelihood of heart disease in the prediction.

Negative Contributions (Right of 0.54):

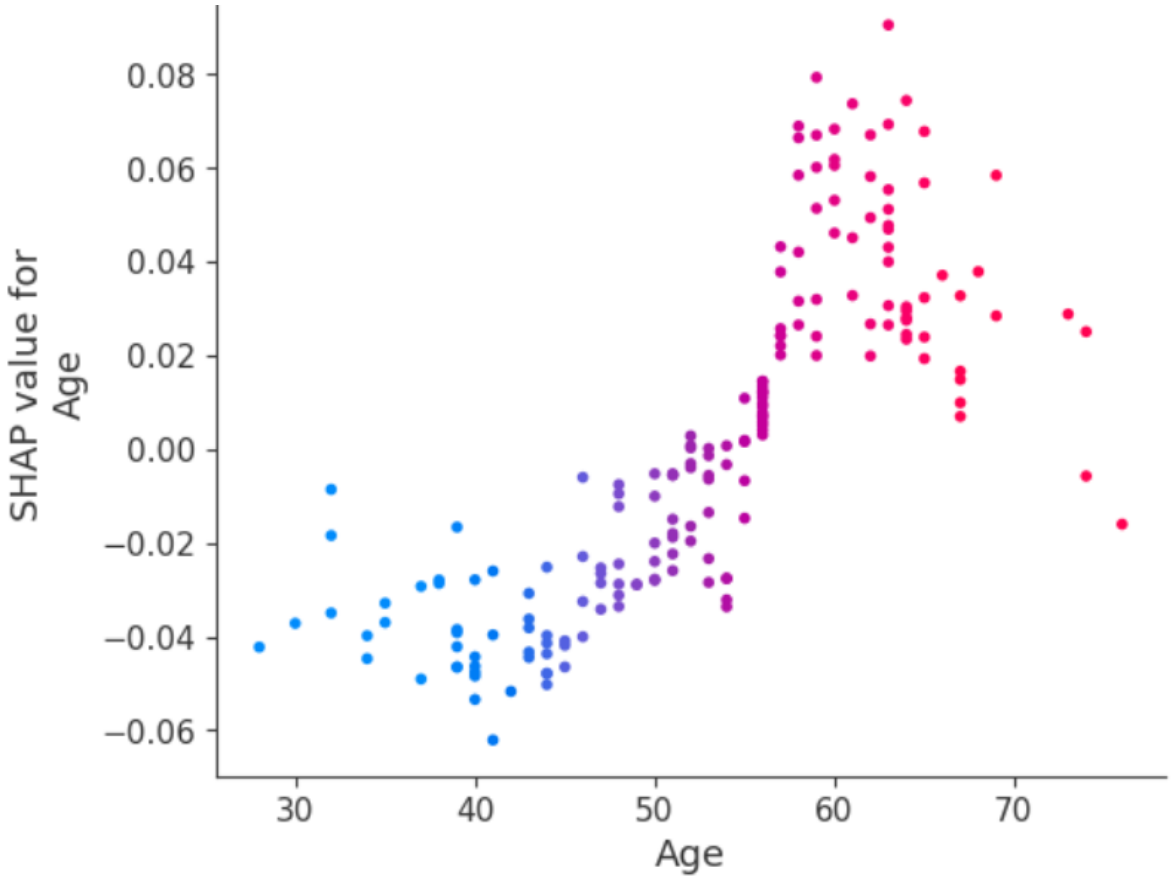
Conversely, the blue bar to the right of 0.54 indicates negative contributions that lowers the probability of Heart Disease. Among these, ST_Slope_Up exerts the most substantial negative impact, suggesting that individuals with an upward-sloping ST segment significantly reduce the probability of being classified as having heart disease. Following ST_Slope_Up, ST_Slope_Flat, ExerciseAngina_Y, ChestPainType_ATA, and Sex_M also exert negative influences, albeit to varying degrees. These features, when present, decrease the likelihood of the prediction indicating heart disease.

The lengths of the arrows denote the magnitude of each feature's impact. Longer arrows represent features that have a more significant effect on the prediction. In this instance, ST_Slope_Up's arrow is the longest, underscoring its potent impact on the prediction outcome.

This detailed graph helps us understand how different things about a person's health affect the model's prediction. For example, as a person gets older, the chance of the model predicting heart disease goes up. However, there are other factors, like the type of ST segment and whether a person experiences exercise-induced angina, that lower the chance of predicting heart disease.

Dependency Analysis of Age

In the conducted analysis, we got to know about the significant impact of age on the predictive outcomes. The Dependency Plot illustrates how 'Age' influences the model's predictions, shedding light on a vital parameter in our heart disease prediction model.



Key Observations from the Dependency Plot:

- X-Axis Representation: The X-axis represents the 'Age' parameter, ranging from younger to older individuals.

- Y-Axis Representation: The Y-axis signifies the SHAP (Shapley Additive explanations) values associated with 'Age,' indicating the magnitude and direction of the influence.

Detailed Insights:

1. Blue Dots (Age 30 to 50): Dots in the blue range, situated below zero on the Y-axis, suggest a negative impact on the prediction of heart disease. This implies that individuals within this age group are less likely to be diagnosed with heart disease according to the model.

2. Purple Dots (Age 50 to 60): The purple dots, positioned close to zero on the Y-axis, signify a neutral or marginal influence on predictions. Within this age bracket, the model's output remains relatively stable, indicating a balanced influence of age on the likelihood of heart disease.

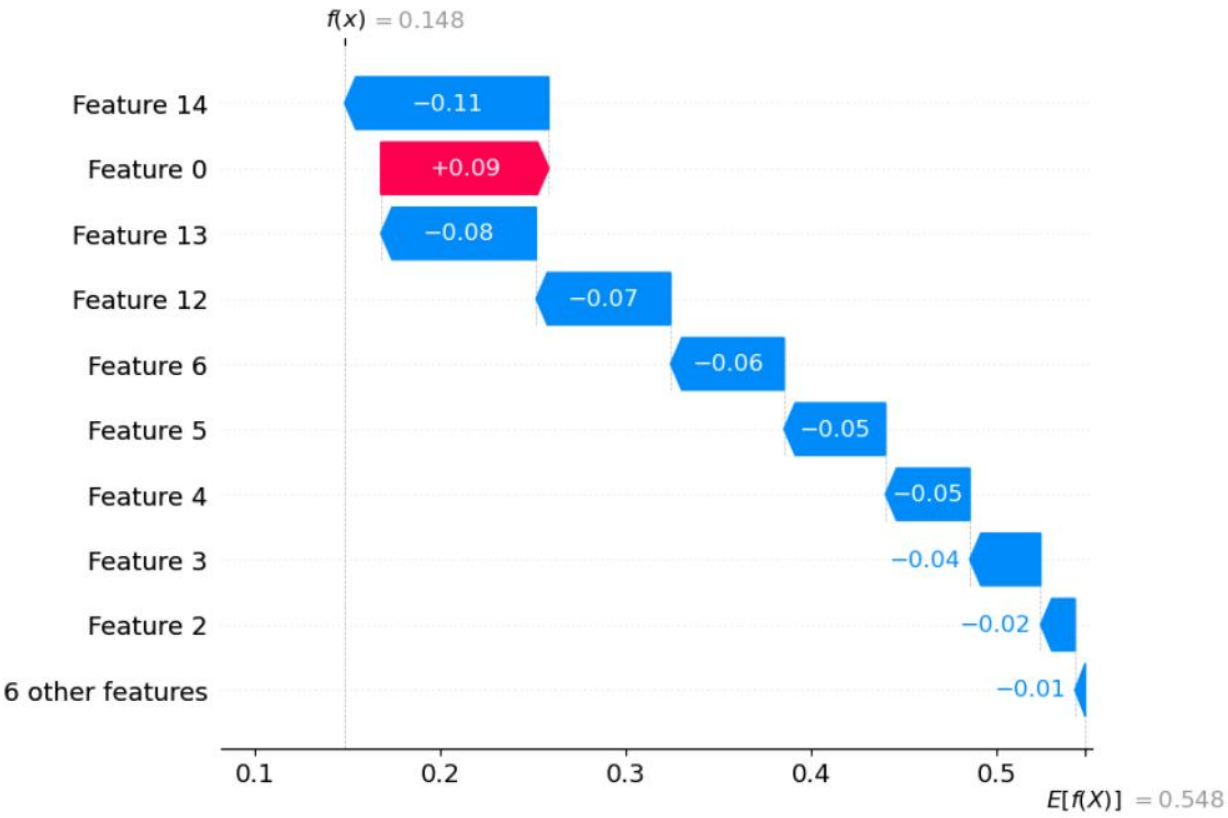
3. Red Dots (Age 60 and Above): Dots in the red range, located above zero on the Y-axis, demonstrate a positive impact on the prediction of heart disease. This implies that individuals aged 60 and above are more likely to be diagnosed with heart disease according to the model.

Implications and Significance:

The Dependency Plot serves as a valuable tool for understanding the relationship between age and heart disease predictions. It enables us to identify specific age groups where the model predicts a higher likelihood of heart disease.

Detailed Explanation of Prediction for a Specific Instance

we provide a granular analysis of the model's prediction for a specific instance, illuminating the interplay of features that influenced the final outcome. The chosen instance, represented by the features and their respective contributions, is dissected in a Waterfall Plot, enabling a precise understanding of the prediction process.



1. X-Axis Values (0.1 to 0.5):

- The X-axis represents numerical values ranging from 0.1 to 0.5, denoting the cumulative impact of features on the prediction.

- This progression illustrates how the prediction score is built incrementally by incorporating the contributions of individual features.

2. Y-Axis Features:

- The Y-axis enumerates the features considered by the model during the prediction process.

- Each feature is represented by a specific color, indicating the direction and magnitude of its contribution.

Detailed Feature Contributions:

1. Feature 14 (Blue Bar, -0.11):
   - The blue bar, extending towards the left, illustrates a negative contribution of -0.11.
   - This suggests that Feature 14 significantly decreases the likelihood of heart disease in the prediction.
2. Feature 0 (Red Bar, 0.09):
   - The red bar, directed towards the right, signifies a positive contribution of 0.09.
   - Feature 0, in this context, amplifies the prediction score, indicating a moderate inclination towards heart disease.
3. Feature 13 (Blue Bar, -0.08):
   - Another blue bar, pointing leftward, represents a negative contribution of -0.08.
   - Feature 13 acts as a suppressor, mitigating the overall prediction score.
4. Features 6, 5, 4, 3, 2 (Blue bar, Cumulative -0.06 to -0.01):
   - These features, depicted by smaller blue bars, collectively contribute negatively to the prediction.
   - The cumulative impact of these features, although relatively smaller, further tilts the prediction away from heart disease.

The Waterfall Plot provides a breakdown of the prediction process, unravelling the balance of features that sway the model's decision-making. It is evident that specific features wield varying degrees of influence, either enhancing or diminishing the likelihood of heart disease.

Interpretation and Insights:

The SHAP analysis provided valuable insights into the factors influencing the predictions of the heart disease model. By examining the SHAP values, we can pinpoint the features that significantly influence the model's decisions. Here's a summary of the major findings and their implications:

1. Age: Age emerged as a crucial factor, positively correlating with the likelihood of heart disease. This aligns with medical knowledge, as age is a well-established risk factor for cardiovascular issues. The SHAP analysis confirms this widely known medical fact, reinforcing the model's credibility.

2. Resting Blood Pressure (RestingBP): While RestingBP had a relatively minor positive impact, its influence was not as significant as age. Elevated blood pressure is a known risk factor, and the model correctly identified this, albeit with a lesser impact compared to age.

3. Cholesterol Levels (Cholesterol): Cholesterol levels showed a mixed influence, with some instances leading to a slightly higher likelihood of heart disease while others reduced the risk. This impact might be due to the complex interplay of cholesterol subtypes or other genetic factors not included in the model.

4. Maximum Heart Rate (MaxHR): MaxHR exhibited a negative influence, indicating that higher heart rates are associated with a reduced likelihood of heart disease. This finding is intriguing and warrants further investigation, potentially indicating the presence of specific cardiac conditions or fitness levels that impact the heart rate's significance.

5. Exercise-Induced Angina (ExerciseAngina): The presence of exercise-induced angina had a negative impact, aligning with medical knowledge. Angina, indicative of reduced blood flow to the heart during physical exertion, is a classic symptom of heart disease. The model accurately captured this relationship.

6. ST Segment Slope (ST_Slope): The ST segment slope showed varying impacts, with flat and upsloping slopes contributing positively to the likelihood of heart disease. The model's differentiation between these slopes aligns with medical literature.

Relation to Medical Knowledge:

The influential features identified by the SHAP analysis largely align with established medical literature. Age, high blood pressure, elevated cholesterol levels, exercise-induced angina, and specific ECG patterns are well-documented risk factors for heart disease. The model's recognition of these factors underscores its ability to capture essential clinical correlations.

Surprising Results:

One surprising result was the negative correlation between maximum heart rate (MaxHR) and heart disease likelihood. While it's unexpected at first glance, this could indicate a subgroup of individuals with high heart rates due to excellent cardiovascular fitness, potentially leading to a reduced risk of heart disease. This explanation indicates that the model could be picking up on small details in the information. It shows how complicated real-life patient situations are.