Interpreting Random Forests on a Heart Disease Dataset

Student name: Prabhu Sekhar Reddy Settipalli
Student ID: 23073211
GitHub: https://github.com/prabhu124-sep/Interpreting-Random-Forests-on-a-Heart-Disease-Dataset
Dataset: *Heart Failure Prediction* (heart.csv, 918 patients)

---

1. Introduction

Machine learning models are increasingly used in healthcare to estimate a patient's risk of disease based on routine clinical measurements. Tree-based ensemble methods such as Random Forests are popular because they handle mixed data types well and often achieve strong predictive performance with limited tuning. However, clinicians and patients need more than an accurate "black-box" prediction: they also need to understand which factors drive the model's decisions.

This tutorial shows how to interpret a Random Forest classifier trained to predict heart disease using a real-world heart failure dataset. The goals are:

To train a reasonably accurate Random Forest model for binary heart-disease prediction using scikit-learn.

To explain, at a beginner-friendly level, what a Random Forest is and how it works.

- To examine **global feature importance** to see which clinical variables the model relies on most.
- To use **partial dependence plots (PDPs)** to visualise how individual features and feature pairs influence the predicted probability of heart disease.
- To discuss the limitations and ethical implications of using such models in healthcare.

For accessibility, all plots in the accompanying notebook use a colour-blind-friendly palette and clear text descriptions so that readers with common forms of colour-vision deficiency and screen-reader users can follow the analysis.

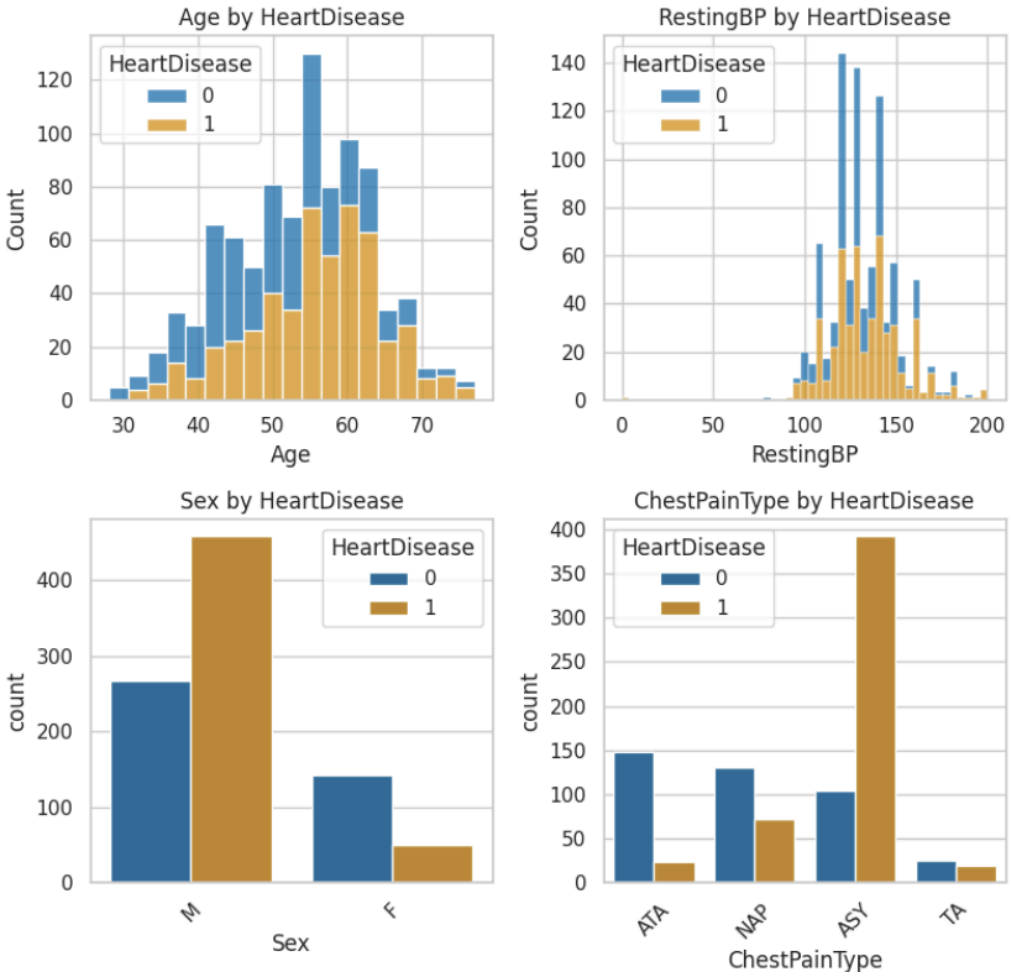
2. Dataset: Heart Failure Prediction

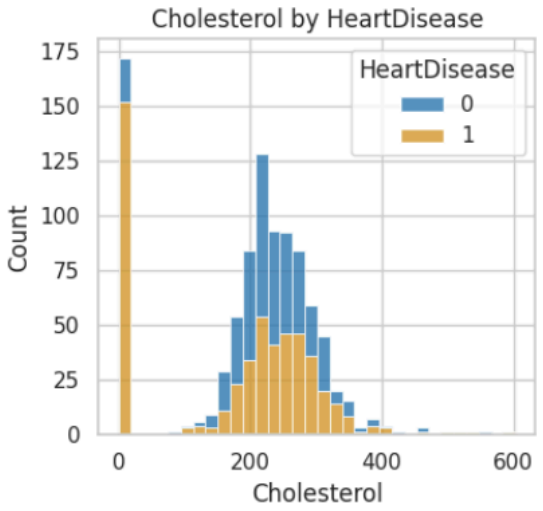
The tutorial uses the *Heart Failure Prediction* dataset from Kaggle, derived from a heart-disease study and containing 918 patients. Each row corresponds to one patient with 11 input features:

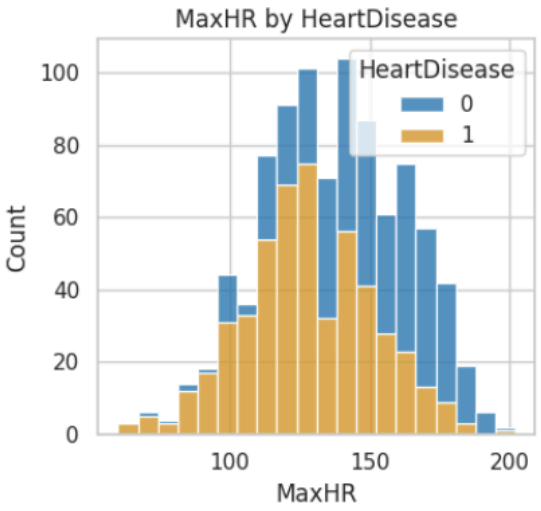
- Numeric: age (Age), resting blood pressure (RestingBP), serum cholesterol (Cholesterol), maximum heart rate during exercise (MaxHR), and ST depression (Oldpeak).
- Categorical: sex (Sex), chest pain type (ChestPainType), fasting blood sugar (FastingBS), resting ECG (RestingECG), exercise-induced angina (ExerciseAngina), and ST slope (ST_Slope).

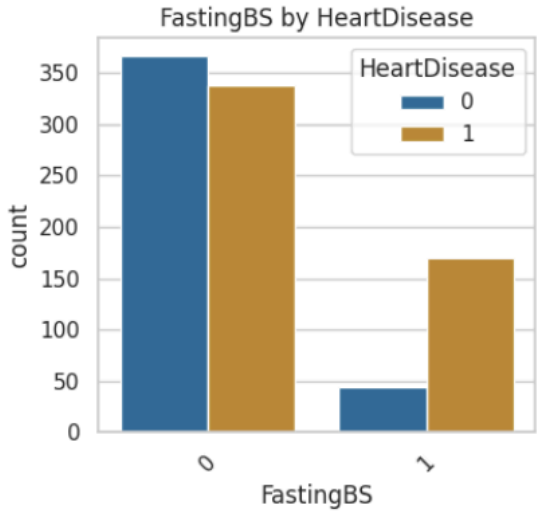
The target variable HeartDisease indicates presence (1) or absence (0) of heart disease, with about 55% positive and 45% negative cases, so the classes are reasonably balanced. The data are split into a training set (80%, 734 patients) and a test set (20%, 184 patients) using stratified sampling to preserve this class balance.

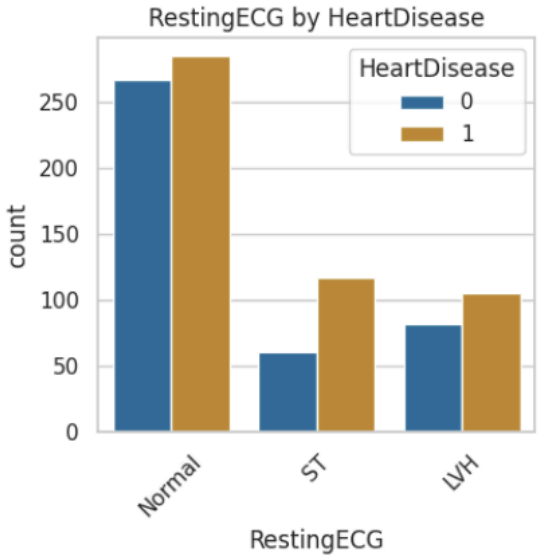
Figure 1 – Feature distributions by class

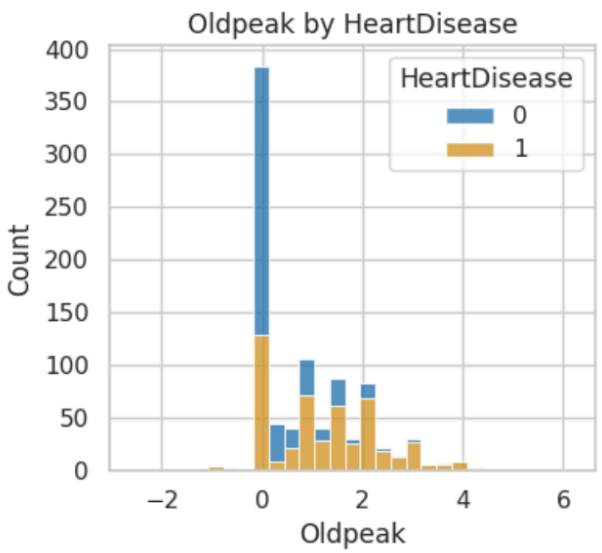


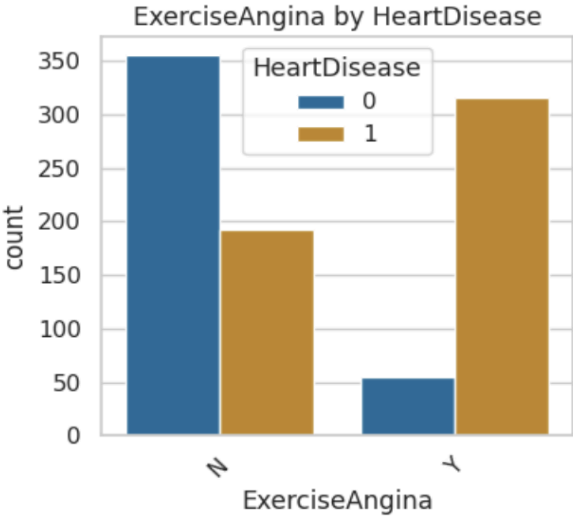












These exploratory plots show that patients with heart disease tend to be slightly older, have lower maximum heart rate, and more often present with asymptomatic chest pain and exercise-induced

angina than patients without heart disease. Such visible differences suggest that a classifier should be able to learn useful patterns from the data.

---

3. What Is a Random Forest?

3.1 Decision trees in plain language

A **decision tree** is a flow-chart model that asks a sequence of yes/no questions about the input features and then outputs a prediction at the bottom ("leaf"). For heart disease, a path might look like:

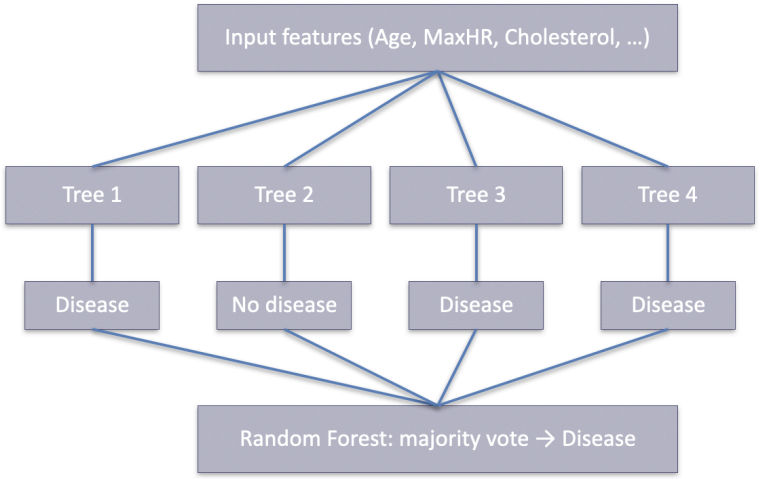
- Is age > 55?
- If yes, is MaxHR < 140?
- If yes, predict "disease"; otherwise, predict "no disease".

The tree chooses these splits to best separate patients with and without heart disease in the training data, usually by minimising an impurity measure such as Gini impurity.

3.2 From one tree to a forest

A **Random Forest** builds many decision trees instead of just one and combines their predictions. Two kinds of randomness make the trees different:
a) **Bagging (bootstrap aggregating):** Each tree is trained on a random sample of patients drawn with replacement from the training data.
b) **Random feature selection:** At each split, a tree only considers a random subset of features instead of all features.
For classification, each tree votes for a class, and the forest predicts the majority vote across all trees. This averaging reduces variance and makes the model more robust than a single deep tree, which might overfit the training data.

Figure 2 – Conceptual diagram of a Random Forest



---

4. Training and Evaluating the Model

The implementation uses a scikit-learn pipeline that:

- Passes numeric features through unchanged.
- One-hot encodes categorical features using OneHotEncoder(handle_unknown="ignore").
- Trains a RandomForestClassifier with 200 trees and default depth.

A pipeline keeps preprocessing and model fitting in a single object and reduces the risk of accidentally leaking information from the test set into the training process.

After fitting on the training set, predictions are evaluated on the held-out test set using a confusion matrix. For a binary classifier, the four entries are:

- **True Positive (TP):** predicted disease and truly has disease.
- **True Negative (TN):** predicted no disease and truly no disease.
- **False Positive (FP):** predicted disease but truly healthy (false alarm).
- **False Negative (FN):** predicted no disease but truly has disease (missed case).

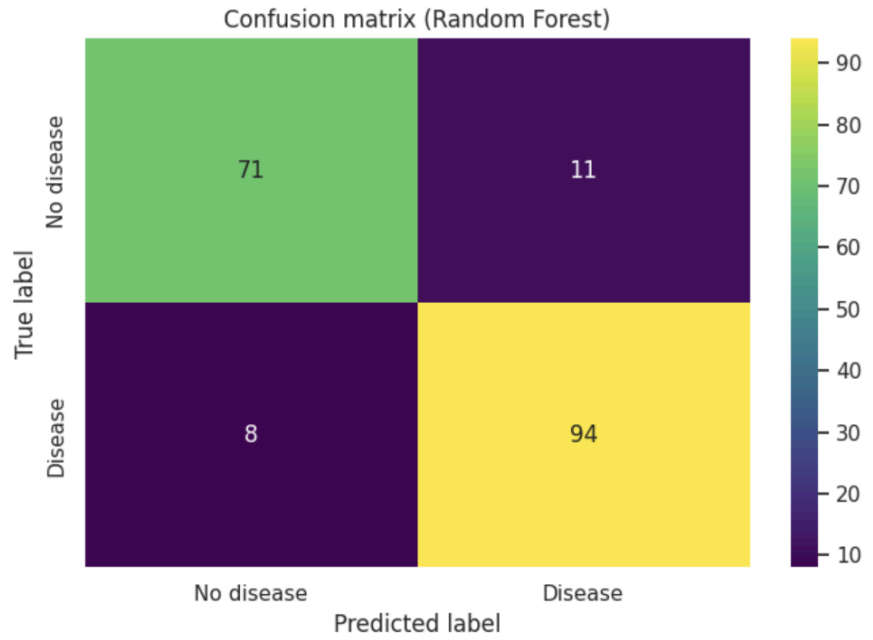
From these counts, common metrics are:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$

$$\text{FPR} = \frac{FP}{FP + TN}, \quad \text{FNR} = \frac{FN}{TP + FN}$$

Figure 3 – Confusion matrix

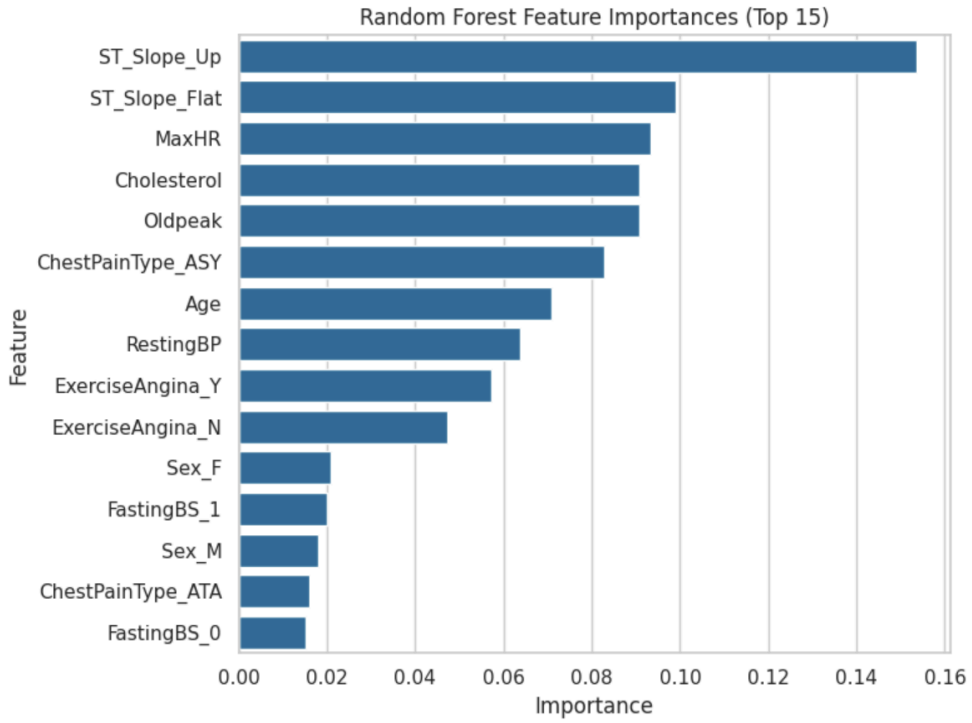


In this model, TP = 94, TN = 71, FP = 11, and FN = 8, giving a test accuracy of about 0.90. Precision and recall for both classes are around 0.87–0.92, meaning the model performs similarly well at identifying diseased and non-diseased patients, although the false negatives remain clinically important.

5. Global Feature Importance

Tree-based models can provide global feature importance based on the average reduction in impurity (for example, Gini impurity) that results from splitting on each feature across all trees. After preprocessing, the Random Forest sees 21 features (five numeric and 16 one-hot-encoded categories), and the top 15 importance scores are plotted.

Figure 4 – Top 15 feature importances



The most important features are the ST segment slope categories (ST_Slope_Up and ST_Slope_Flat), maximum heart rate (MaxHR), serum cholesterol, ST depression (Oldpeak), asymptomatic chest pain type (ChestPainType_ASY), age, resting blood pressure, and exercise-induced angina. Clinically, this is reassuring: ST-segment changes on the ECG, exercise capacity, and chest-pain characteristics are well-known markers of underlying coronary disease.

Impurity-based feature importance has limitations: it can over-emphasise variables with many possible split points and can be unstable when predictors are strongly correlated, so it should be treated as an approximate global summary rather than a definitive ranking of causal effects.

---

6. Partial Dependence: How Features Affect Risk

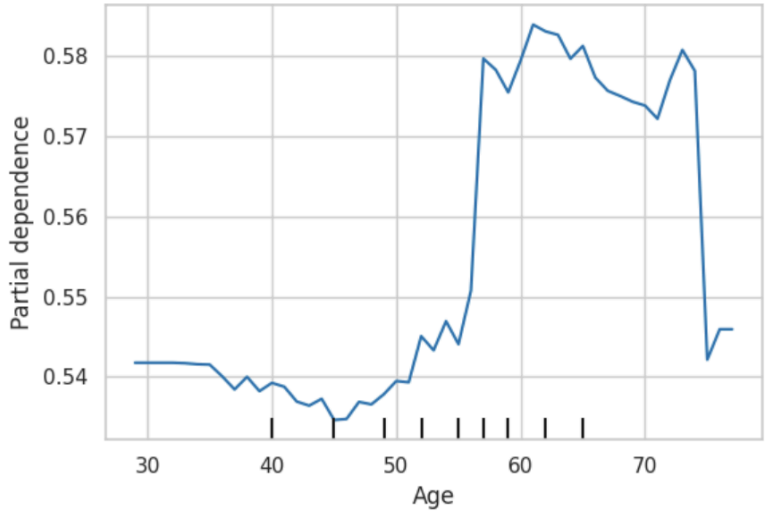
Partial dependence plots estimate the average model prediction as a function of one (or a few) features, marginalising over the others. For a feature x_j , the partial dependence for class-1 probability can be written as

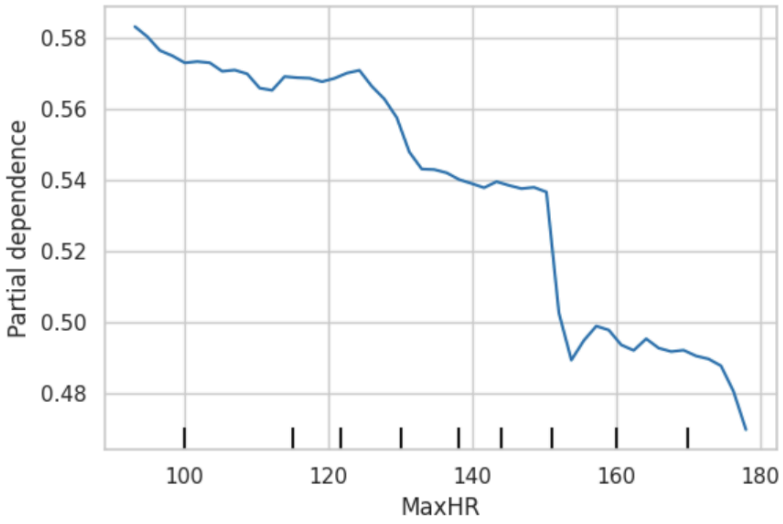
$$PD(xj) = Ex - j[f(x - j, xj)],$$

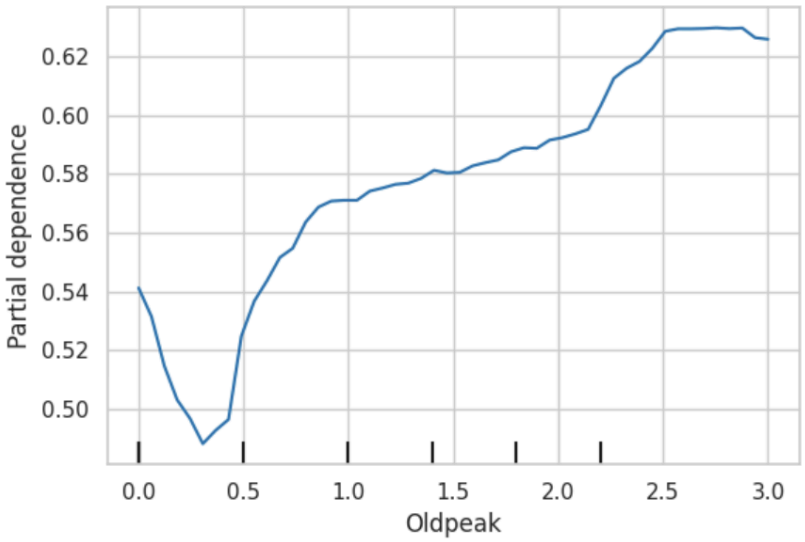
where x_j are all other features and fff is the model's prediction function. Scikit-learn approximates this expectation by averaging predictions over the training set for a grid of feature values.

Three clinically important features are examined: Age, MaxHR, and Oldpeak.

Figure 5 – 1D PDPs for Age, MaxHR, and Oldpeak







- **Age:** The partial dependence curve is slightly lower for patients in their late 30s and early 40s and rises for patients in their late 50s and 60s, indicating that the model associates older age

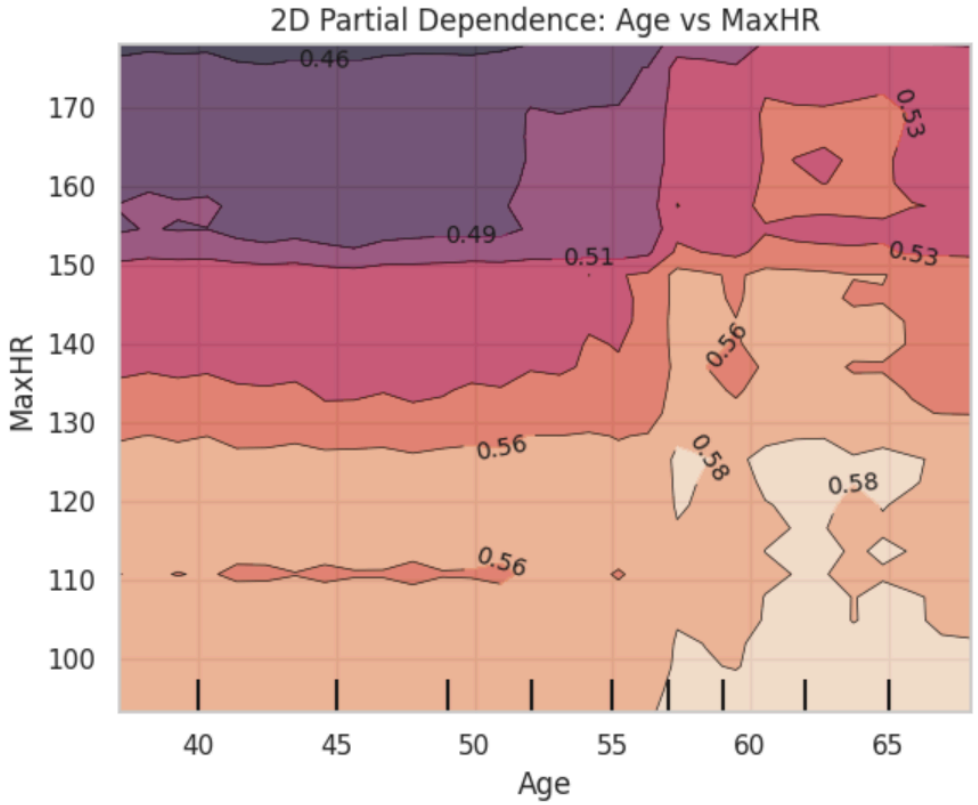
with higher risk.

- **MaxHR:** As maximum heart rate increases from roughly 100 to 170 beats per minute, the predicted probability of heart disease steadily decreases, consistent with the idea that good exercise capacity is protective.
- **Oldpeak:** At very low ST depression values the predicted risk is lower, but as Oldpeak increases beyond about 0.5–1.0, the estimated risk rises, reflecting the link between larger ST-segment depression and more severe myocardial ischemia.

These plots make the model's behaviour more transparent for beginners: rather than just seeing a probability, they can see how specific clinical measurements shift that probability up or down.

To examine interactions, a 2D partial dependence plot is generated for Age and MaxHR.

Figure 6 – 2D PDP for Age vs MaxHR



The 2D plot reveals that the highest predicted risk occurs in regions where patients are older and achieve relatively low maximum heart rates during exercise, while younger patients with high maximum heart rates lie in lower-risk regions. This illustrates how the Random Forest combines multiple signals instead of relying on any single feature.

---

7. Limitations, Ethics, and Accessibility

Despite good performance, this Random Forest is still only an approximation and makes clinically important errors, including some false negatives where patients with heart disease are predicted as low risk. In practice, a model like this should be used only as a **decision-support tool**, complementing but not replacing clinical judgement.

The interpretability tools used here also have limitations. Impurity-based feature importance can be biased toward variables with many categories or splits and may not behave well when predictors are highly correlated. Partial dependence plots assume that the feature being varied is independent of the others, which is often false in medical data, so PDP curves should be read as approximate descriptions of model behaviour, not causal effects.

From an ethical standpoint, medical ML systems must address bias, privacy, and transparency. Training data may underrepresent certain demographic groups, leading to systematically worse performance for those patients if deployed without fairness checks. At the same time, explainability methods like feature importance and PDPs support the ethical requirement of explainability by helping clinicians and patients understand and contest model outputs.

The notebook and figures are designed with accessibility in mind: a colour-blind-friendly palette is used, fonts are large, and key graphics are accompanied by descriptive text so that screen-reader users can grasp the essential message even without viewing the image directly.

---

8. Conclusion

This tutorial demonstrated how a Random Forest classifier can be trained on a real heart-disease dataset and then opened up using feature importance and partial dependence plots to provide clinically meaningful explanations. The model achieved around 90% accuracy on the test set, relied heavily on ECG-derived and exercise-related features, and exhibited intuitively reasonable relationships between age, maximum heart rate, ST depression, and predicted risk.

For students and practitioners, the key lesson is that interpretable tools can turn a tree-based ensemble from an opaque “black box” into a model whose behaviour can be inspected, questioned, and related back to domain knowledge. At the same time, the limitations and ethical issues discussed here emphasise that explainability is only one component of responsible AI in healthcare, which also requires rigorous validation, fairness analysis, and ongoing human oversight before any clinical deployment.

References

1. Kaggle. (2021). *Heart failure prediction dataset*. Retrieved December 9, 2025, from https://www.kaggle.com/datasets/fedesoriano/heart-failure-predictionkaggle
2. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. (RandomForestClassifier documentation: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html)[2]
3. Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). Retrieved from https://christophm.github.io/interpretable-ml-book/christophm.github
4. Rasheed, K., Qureshi, R., Qamar, A. M., et al. (2022). Explainable, trustworthy, and ethical machine learning for healthcare. *Computers in Biology and Medicine, 145*, 105403. https://doi.org/10.1016/j.compbiomed.2022.105403sciencedirect
5. Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science, 4*, 123–144. https://doi.org/10.1146/annurev-biodatasci-092820-114757pmc.ncbi.nlm.nih
6. Hoche, M., et al. (2025). What makes clinical machine learning fair? A practical ethics framework. *PLOS Digital Health, 4*(3), e0000728. https://doi.org/10.1371/journal.pdig.0000728journals.plos