Heart Disease Risk Prediction: Logistic Regression vs. Random Forest

A Comparison of Logistic Regression & Random Forest

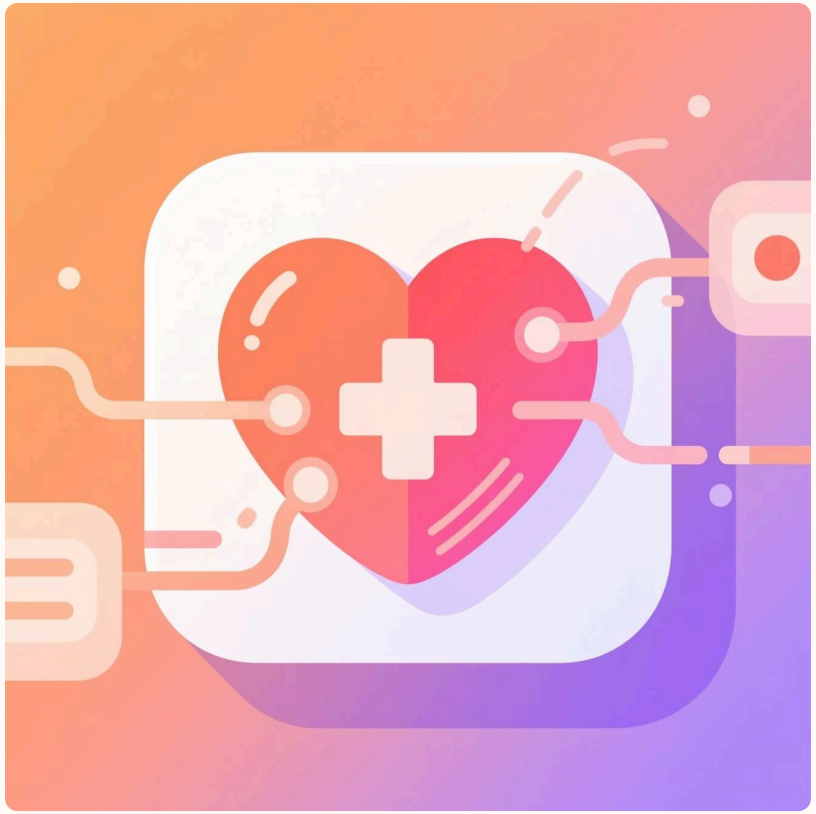
Dataset: Kaggle Heart Disease

Problem Statement & ML Objective

Our primary objective is to accurately predict binary heart disease risk (0/1) in patients. This involves generating a probability score $P(HeartDisease = 1|features)$ rather than a simple classification.

Why probabilities matter in medical decision-making:

- Quantifying uncertainty for clinicians.
- Facilitating stratified treatment plans.
- Enabling early intervention based on risk thresholds.



> Evaluation Objective: Accuracy + Calibrated Risk + Interpretability

Linear Regression: A Brief Overview

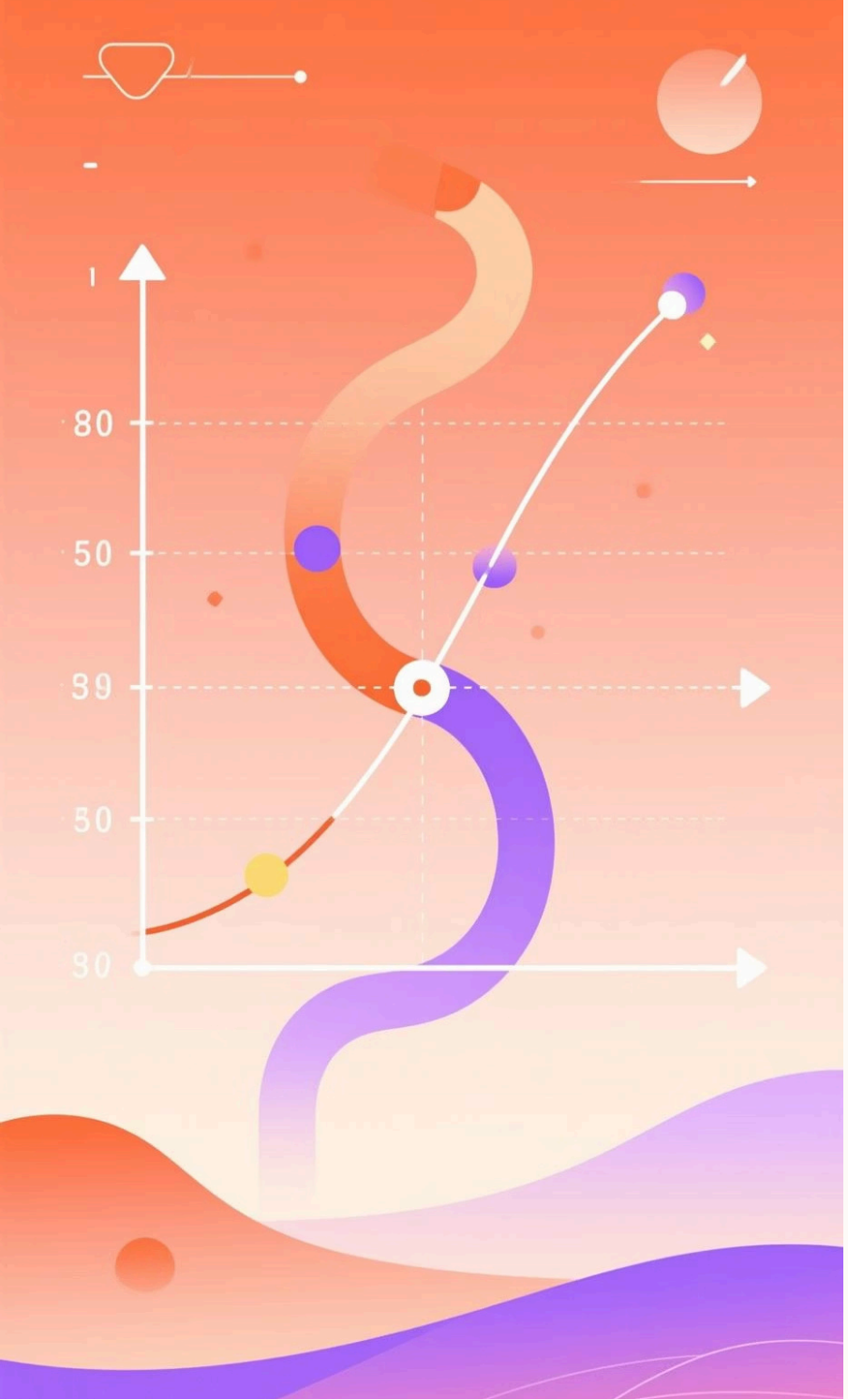
Linear Regression, while foundational, is primarily designed for continuous outcome prediction. Its core equation models a linear relationship between features and target:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

The model minimises the Mean Squared Error (MSE) to find optimal coefficients:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Why it fails for binary classification: Linear regression outputs real values $(-\infty, \infty)$, which cannot be directly interpreted as probabilities $[0, 1]$. It also lacks a natural decision boundary for classification.



Logistic Regression: Core Idea

Sigmoid Transformation

The logistic function, or sigmoid, transforms linear outputs into probabilities $[0, 1]$.

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Prediction Example

If $z = \beta_0 + \beta_1 x \dots \beta_n x$, then $P(Y = 1|X) = \sigma(z)$.

For example, $\sigma(0) = 0.5$, $\sigma(3) \approx 0.95$.

Decision Boundary

A threshold (e.g., 0.5) on the probability defines the classification decision boundary, separating classes.

Logistic Regression: Mathematics + Derivation

The logistic hypothesis maps any real-valued input to a probability:

$$h_\theta(x) = \sigma(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

The logit transformation connects probability p to a linear model:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \theta^T x$$

The Log-Likelihood function for N observations is maximised to find optimal θ:

$$L(\theta) = \sum_{i=1}^{N} [y_i \log(h_\theta(x_i)) + (1 - y_i) \log(1 - h_\theta(x_i))]$$

Minimising the negative Log-Likelihood (Cross-Entropy Loss):

$$J(\theta) = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(h_\theta(x_i)) + (1 - y_i) \log(1 - h_\theta(x_i))]$$

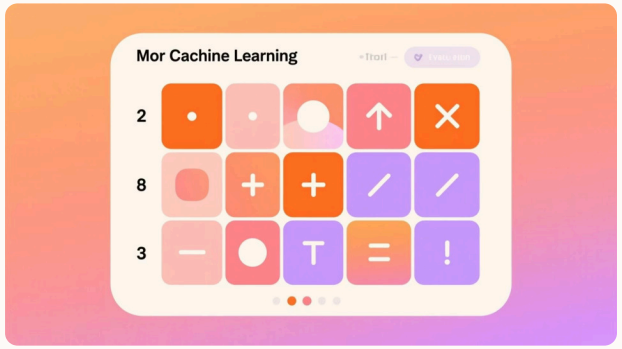
Gradient Descent Update Rule for each parameter θ_j:

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{N} (h_\theta(x_i) - y_i) x_{i,j}$$

Logistic Regression Metrics & Interpretation

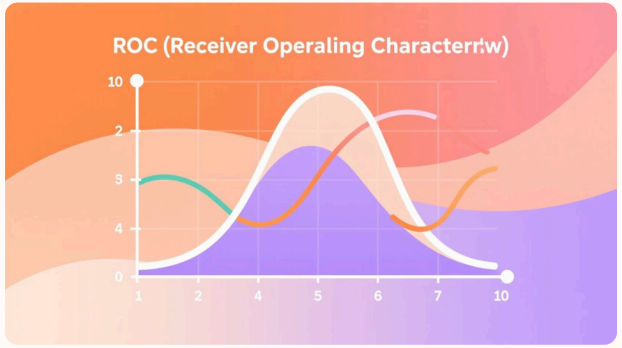
Confusion Matrix

Evaluates model performance by summarising true positives, true negatives, false positives, and false negatives.



Key Metrics

- **Precision:** $TP/(TP+FP)$
- **Recall:** $TP/(TP+FN)$
- **F1-score:** Harmonic mean of precision & recall.
- **AUC-ROC:** Probability that model ranks a random positive sample higher than a random negative sample.



Odds Ratio Interpretation

e^{β_j} represents the change in odds of heart disease for a one-unit increase in x_j, holding other features constant. Crucial for clinical interpretability.

Example: An Odds Ratio of 1.5 for "Smoking" means smokers have 1.5 times the odds of heart disease compared to non-smokers.

Random Forest: Concept & Architecture

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training. For classification tasks, the output is the class selected by most trees (majority vote).

1

Bagging (Bootstrap Aggregating)

Each tree is trained on a bootstrap sample (random sampling with replacement) from the original dataset.

$$\text{Bootstrap Sample } D_k \subset D$$

2

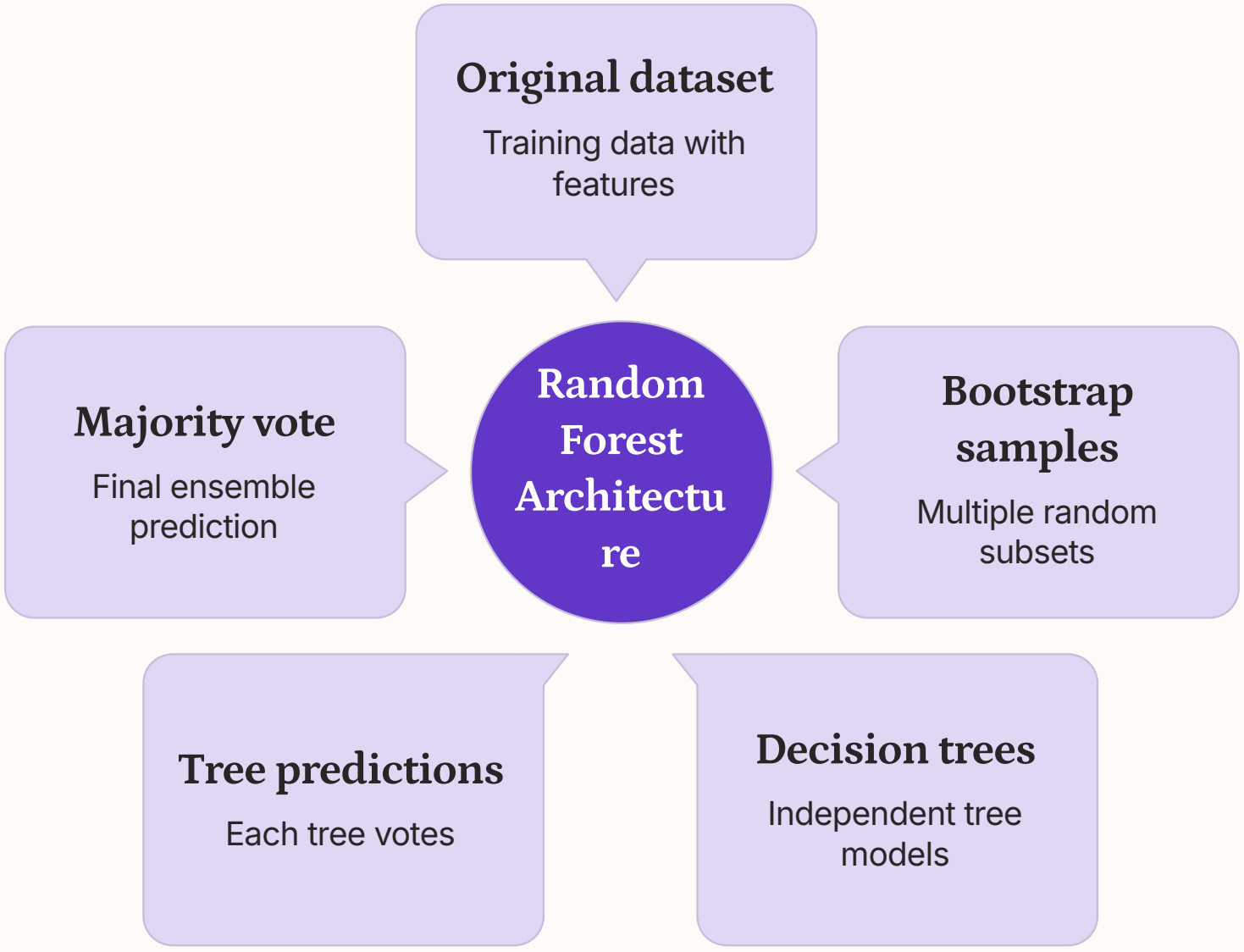
Feature Randomness

At each split in a decision tree, only a random subset of features is considered, decorrelating the trees.

3

Ensemble Prediction

Final prediction is derived by aggregating the predictions of individual trees, typically via majority voting for classification.



Random Forest Mathematics

Gini Impurity

Measures the probability of misclassifying a randomly chosen element in the dataset if it were randomly labeled according to the distribution of labels in the subset.

$$G = 1 - \sum_{k=1}^{K} p_k^2$$

Entropy

Measures the disorder or uncertainty in a dataset. A split aims to maximise Information Gain (reduction in entropy).

$$H = -\sum_{k=1}^{K} p_k \log_2(p_k)$$

Feature Selection

At each node, $\sqrt{N_{features}}$ random features are considered for splitting to decorrelate trees.

Aggregated Prediction (for classification):

$$\hat{y} = \text{mode}(\text{Tree}_1(x), \ldots, \text{Tree}_K(x))$$

Bias-Variance Tradeoff:

$$\text{Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

Out-of-Bag (OOB) Error: The error estimated by using samples not included in a tree's bootstrap. Provides an unbiased estimate of generalisation error.

Random Forest Strengths for Heart Disease Prediction

Non-linearity Handling

Effectively captures complex, non-linear relationships between medical features and heart disease risk, which linear models often miss.

High Accuracy

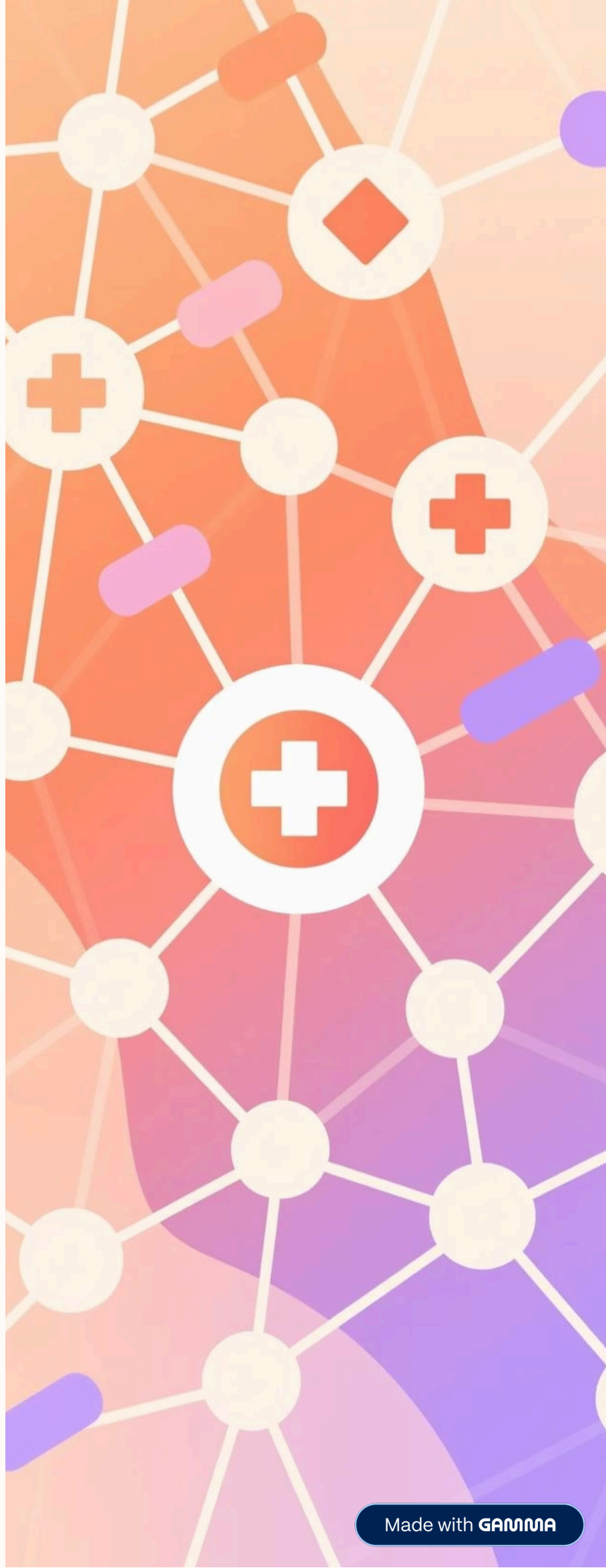
Generally achieves higher predictive accuracy on complex medical datasets due to ensemble averaging, reducing overfitting compared to single decision trees.

Feature Importance

Provides quantifiable feature importance scores, highlighting which physiological parameters are most influential in heart disease prediction for clinical insights.

Robustness

Resilient to noise and missing values, making it suitable for real-world clinical data that often contain imperfections.



Final Model Choice: A Hybrid Approach

Primary Model: Logistic Regression

- **Interpretable Coefficients:** Direct clinical interpretation of feature impact (odds ratios).
- **Probability Output:** Provides well-calibrated risk probabilities essential for patient counselling.
- **Suitable for Clinical Documentation:** Transparent and explainable, aligning with regulatory requirements and medical best practices.

Secondary Model: Random Forest

- **Validates Predictions:** Acts as a robust second opinion, confirming or highlighting discrepancies in Logistic Regression's output.
- **Captures Nonlinearities:** Identifies complex interactions and patterns missed by the linear assumptions of Logistic Regression.
- **Higher Accuracy:** Utilised for scenarios demanding peak predictive performance and identifying subtle risk factors.

Clinical Decision

Final decisions using probabilities

Random Forest

Validation layer providing feedback

Logistic Regression

Core model with interpretable odds