Case Study: Clinical Decision Support for Early Heart Disease Detection

1. Problem Statement and Objectives

The high mortality rate associated with cardiovascular diseases is often exacerbated by late-stage diagnosis in primary care settings. Clinical experts face challenges in identifying subtle patterns within standard patient records that indicate early-stage heart disease.

Objectives:

- To implement the **EFEM-HDP ensemble framework** as a secondary screening tool.
- To provide doctors with a high-confidence probability score for risk assessment.
- To achieve a diagnostic accuracy of at least 85% on clinical benchmark data.

2. Data Preprocessing

The system was trained and evaluated using the Cleveland Heart Disease dataset from the UCI Machine Learning Repository. The dataset consists of 303 patient records with 13 clinical attributes and one binary target variable indicating the presence or absence of heart disease.

To ensure the model receives clean, high-quality data, the following steps were performed on the dataset:

- **Data Cleaning:** Missing values were identified and addressed using **mean imputation** to maintain dataset integrity.
- **Feature Scaling:** Since clinical values like cholesterol and age are on different scales, we applied **StandardScaler** to ensure no single feature dominated the model weights.
- **Dimensionality Reduction: PCA** was used to transform the data, reducing noise while keeping the most significant clinical patterns.
- **Feature Selection: RFE** was utilized to select only the most discriminative clinical attributes, improving the speed of diagnosis.

3. Model Selection and Development

The **EFEM-HDP** model was developed using a **Soft Voting Classifier** approach. This was selected over single-classifier models (like a single Decision Tree) because it aggregates the strengths of multiple algorithms:

- **Logistic Regression:** Provides a stable linear baseline.
- **Support Vector Machine (SVM):** Efficiently handles high-dimensional clinical features.
- **Random Forest:** Captures non-linear relationships between variables.

By combining these, the ensemble provides a more "voted" and reliable prediction than any individual model.

4. Model Evaluation Metrics

Model performance was evaluated using Accuracy, Precision, Recall, F1-score, and ROC-AUC. Particular emphasis was placed on recall, as false negatives in medical diagnosis may delay treatment and increase patient risk.

5. Visualizations and Insights

5.1 Insights

- **Accuracy Improvement:** The ensemble reached **87% accuracy**, a 3% gain over baseline models.

  This improvement is clinically significant, as higher recall and accuracy directly reduce the risk of missed diagnoses in early-stage heart disease.

- **Feature Importance:** Feature engineering (PCA/RFE) proved critical; raw datasets without these steps showed lower reliability.

6. Recommendations

Based on this analysis, the following actions are recommended for clinical implementation:

- **Initial Screening:** Hospitals should use the EFEM-HDP model to flag high-risk patients for immediate specialist consultation.
- **Integration:** The model should be deployed as a web-based decision support system for easier access by general practitioners.
- **Data Growth:** Periodically update the model with new patient data to maintain accuracy across different demographics.