Executive Summary: AI-Powered Heart Disease Prediction

Objective

This project aims to develop and evaluate machine learning models to accurately predict the likelihood of heart disease using clinical patient data from the UCI Heart Disease dataset. The ultimate goal is to support preventive healthcare strategies by identifying at-risk individuals early.

Dataset

- Source: UCI Heart Disease Dataset

- Size: 303 patient records

- Features: 13 original attributes + 4 engineered features (e.g., cholesterol-to-age ratio, heart rate reserve)

Methodology

1. Exploratory Data Analysis (EDA): Identified key correlations and visualized distributions.

2. Feature Engineering: Created derived metrics such as cholesterol per age, heart rate reserve, log-transformed ST depression, and interaction terms.

3. Preprocessing: Applied one-hot encoding and feature scaling.

4. Modeling: Trained Logistic Regression, Random Forest, and XGBoost; used grid search and built ensemble models (Voting, Weighted Voting, and Stacking).

Results

Model Performance:

- Logistic Regression: ~86% Accuracy, ~91% ROC AUC

- Random Forest: ~87% Accuracy, ~91% ROC AUC

- XGBoost: ~88% Accuracy, ~92% ROC AUC

- Voting Ensemble: ~89% Accuracy, ~93% ROC AUC

- Stacking Ensemble: ~90% Accuracy, ~94% ROC AUC

Interpretation with SHAP

- SHAP values provided model interpretability.

- Key Influential Features: Age, Max Heart Rate, ST Depression, Cholesterol × Resting BP.

- Waterfall plots illustrated patient-specific predictions.

Conclusion

The project demonstrates that ensemble learning and explainable AI can significantly enhance the prediction of heart disease risk, aiding data-driven clinical decision-making.