Executive Summary: Heart Disease Prediction Using Machine Learning

This project focuses on predicting heart disease using machine learning techniques, with a strong emphasis on data preprocessing, model selection, performance evaluation, and explainability. The dataset was sourced from Kaggle and contained both numerical and categorical features relevant to heart health.

Data Exploration and Preprocessing

Initial exploration involved basic functions like head() and describe() to understand the structure and summary statistics of the data. A major challenge emerged in handling categorical variables, especially non-boolean types.

While boolean values were easily encoded as 0 and 1, encoding other categorical Variables required deeper understanding and transformation using techniques like one-hot encoding (dummy variables).

Data Cleaning and Feature Engineering

To ensure compatibility with machine learning models, data types were rectified and non-numeric features were converted. Feature engineering was approached iteratively — beginning with dropping highly correlated redundant features, and later refining with proper one-hot encoding and correlation analysis. A correlation heatmap helped identify relationships between features and the target variable, guiding the removal of multicollinear features and those with low predictive power.

Exploratory Data Analysis (EDA)

A series of visualizations were created to understand the relationship between features and presence of heart disease. Comparisons such as age vs. heart disease, gender vs. heart disease, and various other pairings were explored. Seaborn visualizations and layout optimizations helped derive insights, such as identifying risk factors and patterns in the dataset.

Model Building and Evaluation

The **Random Forest Classifier** was selected due to its ensemble nature, robustness to overfitting, and interpretability. The model was configured with 1000 estimators and a fixed random_state of 42 for reproducibility. Performance metrics were strong:

- **Accuracy**: 86.6%
- **Precision**: 87.7%
- **Recall**: 86.0%
- **F1-Score**: 86.9%

The confusion matrix showed that the model correctly predicted the majority of both heart disease and non-disease cases, confirming its reliability.

A comparison with the **Support Vector Machine (SVM)** model was conducted. Although the confusion matrix for SVM was promising, its overall metrics underperformed compared to the Random Forest, justifying the final choice.

Model Explainability with SHAP and LIME

To enhance trust and transparency, SHAP (SHapley Additive exPlanations) was applied. This analysis revealed that features like **maximum heart rate**, **thalassemia type**, and **number of major vessels** had significant influence on predictions. Specifically, lower maximum heart rate values were associated with a higher risk of heart disease. Less influential features (with SHAP values below 0.025) were dropped and tested; however, they had negligible impact on model performance.

Additionally, LIME (Local Interpretable Model-agnostic Explanations) was used to explain individual predictions, further validating the model’s behavior.

Conclusion

This project successfully developed a robust and interpretable machine learning model for heart disease prediction. The Random Forest classifier demonstrated strong performance, and SHAP/LIME ensured the model’s decision-making could be trusted. With further validation and real-world integration, this model could serve as a valuable tool in clinical settings, supporting early detection and diagnosis of heart disease.