Heart Disease Prediction using Classification Algorithms

Objective
The objective of this project is to build and compare Logistic Regression, Decision Tree, and K-Nearest Neighbors (KNN) models to predict heart disease using the Heart Disease UCI dataset, with emphasis on minimizing false negatives.

Dataset Overview
The dataset contains clinical and demographic attributes such as age, sex, chest pain type, cholesterol level, resting blood pressure, ECG results, and maximum heart rate achieved. The target variable represents the presence or absence of heart disease.

Data Cleaning
Missing numerical values were imputed using median values, categorical features using mode, and invalid medical values such as cholesterol equal to zero were corrected using median replacement.

One-Hot Encoding
Categorical features including sex, dataset, chest pain type, fasting blood sugar, and resting ECG were one-hot encoded with drop-first strategy to avoid multicollinearity.

Feature Scaling
Numerical features such as resting blood pressure, cholesterol, maximum heart rate, and ST depression were standardized using StandardScaler after data splitting to prevent leakage.

Feature Engineering
New features included cholesterol risk categories, high blood pressure indicator, and heart rate efficiency to enhance model performance and interpretability.

Data Splitting
The dataset was split into training (60%), validation (20%), and test (20%) sets using stratified sampling.

Exploratory Data Analysis
EDA included target distribution analysis, cholesterol vs target boxplots, and cholesterol vs resting blood pressure scatter plots to understand feature relationships.

Model Implementation & Evaluation
Logistic Regression, Decision Tree, and KNN classifiers were implemented and evaluated using confusion matrix, precision, recall, and F1-score with recall prioritized.

Conclusion
Logistic Regression performed best with the highest recall and lowest false negatives, making it most suitable for heart disease prediction. Decision Trees showed strong performance but risked overfitting, while KNN underperformed in recall.