Heart Disease Prediction - Analysis Documentation

Project Overview

This Jupyter Notebook presents a comprehensive machine learning project for predicting heart disease presence in patients based on various medical parameters. The project follows a structured data science workflow from data exploration to model implementation.

Dataset Description

The dataset contains 14 medical attributes with 303 total observations. Key features include:

1. Demographic & Clinical Attributes:
    - age: Patient's age in years
    - sex: Gender (1 = male; 0 = female)
    - cp: Chest pain type (4 categories from typical angina to asymptomatic)
    - trestbps: Resting blood pressure (mm Hg)
    - chol: Serum cholesterol (mg/dl)
    - fbs: Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
2. Cardiac Test Results:
    - restecg: Resting electrocardiographic results
    - thalach: Maximum heart rate achieved
    - exang: Exercise induced angina
    - oldpeak: ST depression induced by exercise
    - slope: Slope of peak exercise ST segment
    - ca: Number of major vessels colored by fluoroscopy
    - thal: Thallium stress test result
3. Target Variable:
    - target: Presence of heart disease (1 = yes, 0 = no)

Technical Implementation

1. Libraries Used:
    - Data Handling: pandas, numpy
    - Visualization: matplotlib, seaborn
    - Machine Learning: scikit-learn, XGBoost, CatBoost
    - Model Evaluation: classification_report, confusion_matrix, accuracy_score
2. Key Analysis Steps:
    - Data Loading & Initial Exploration: Examination of dataset structure and basic statistics
    - Data Preprocessing: Standardization using StandardScaler
    - Model Training: Implementation of multiple classifiers including:
        - XGBoost
        - CatBoost
        - Random Forest
        - K-Nearest Neighbors
        - Support Vector Machines

- Hyperparameter Tuning: Using RandomizedSearchCV for optimization
- Model Evaluation: Comprehensive performance assessment using multiple metrics