UE23CS352A Machine Learning Mini-Project Write-up
Heart Disease Prediction: A Comparative Classification Study
Problem statement - 16

Members: Kshitij - PES2UG23CS290 Likith N - PES2UG23CS306

1. Problem Statement

The objective of this project was to develop a machine learning solution for the **binary classification of heart disease presence** (Class 1) or absence (Class 0) in patients, leveraging the UCI Heart Disease Dataset. The primary challenge involved handling real-world data issues, specifically significant missing values and mixed data types, to build an accurate predictive model.

2. Approach and Methodology

The solution followed a standard Machine Learning pipeline, focusing on robust data preparation before comparative model training.

Data Preprocessing and Feature Engineering

The initial dataset contained 920 samples and 15 features, including several columns with substantial missingness.

1. **Missing Value Treatment:** Columns with highly sparse data (thal, ca, slope, dataset) were dropped. Missing numeric values (trestbps, chol, thalch, oldpeak) were imputed using **KNNImputer ()**, a proximity-based method. Missing categorical values (fbs, restecg, exang) were imputed using **mode imputation** often grouped by the sex feature for context.
2. **Transformation:** The original multi-class target (0-4) was converted to a **binary target (0/1)** to classify the presence of *any* heart disease.
3. **Encoding & Scaling:** Categorical features were converted via **One-Hot Encoding**. All features were then normalized using **StandardScaler** to prevent dominance by features with larger magnitude.
4. **Data Split:** The clean, processed data was split into an **80% training set** and a **20% testing set** (184 samples).

Model Implementation

Two distinct classification algorithms were chosen for training and comparison:

- **Logistic Regression:** Selected as a robust, interpretable linear baseline.
- **Support Vector Machine (SVM):** Implemented with an RBF kernel to capture potential non-linear relationships in the data.

3. Results and Conclusion

Both models demonstrated high performance in classifying the disease presence on the unseen test set, validating the integrity of the feature engineering pipeline. The Support Vector Machine (SVM) model was the top performer, achieving a test Accuracy of 83.70% (compared to 83.14% for Logistic Regression). Both models showed high precision for the positive class (0.88), but the SVM demonstrated a slightly superior Recall (0.84) and F1-Score (0.86) for predicting heart disease. The ability of the SVM to handle non-linear decision boundaries likely contributed to this marginal edge over the linear Logistic Regression.