Data Dictionary

Sex-Specific and Regional Analysis of Heart Disease Prediction Using Machine Learning Algorithms: Insights from the UCI Irvine Public Heart Disease Datasets (Cleveland and Long Beach)

Jonathan Asanjarani
City University of New York Graduate Center
DATA 79000: Capstone Project and Thesis
Advisor: Johanna Devaney

Significant Variables

1. **Age**
    - Type: Integer
    - Description: Patient's age in years.
2. **Sex**
    - Type: Binary (0 for Female, 1 for Male)
    - Description: Biological sex of the patient.
3. **Cp (Chest Pain Type)**
    - Type: Categorical (0–4)
    - Description: Chest pain severity levels, where higher values indicate more severe pain.
4. **Trestbps (Resting Blood Pressure)**
    - Type: Continuous (mmHg)
    - Description: Resting blood pressure in millimeters of mercury. Transformed using logarithmic scaling to reduce skewness.
5. **Chol (Serum Cholesterol)**
    - Type: Continuous (mg/dL)
    - Description: Serum cholesterol level in milligrams per deciliter. Transformed using logarithmic scaling to reduce skewness.
6. **Fbs (Fasting Blood Sugar)**
    - Type: Binary (0 for <120 mg/dL, 1 for ≥120 mg/dL)
    - Description: Indicator of whether fasting blood sugar exceeds 120 mg/dL.
7. **Restecg (Resting ECG Results)**
    - Type: Categorical (0–2)

- Description: Results of resting electrocardiographic tests (e.g., normal, ST-T wave abnormality, left ventricular hypertrophy).

8. **Thalach (Maximum Heart Rate Achieved)**
    - Type: Continuous (bpm)
    - Description: Maximum heart rate achieved during exercise. Transformed using a squared transformation to emphasize non-linear relationships.
9. **Exang (Exercise-Induced Angina)**
    - Type: Binary (0 for No, 1 for Yes)
    - Description: Presence of exercise-induced angina (chest pain).
10. **Oldpeak**
    - Type: Continuous
    - Description: ST depression induced by exercise relative to rest (ECG measure). Transformed using logarithmic scaling to reduce skewness.
11. **Slope (ST Segment Slope)**
    - Type: Categorical (1 for Upsloping, 2 for Flat, 3 for Downsloping)
    - Description: The slope of the peak exercise ST segment.
12. **Ca (Number of Major Vessels)**
    - Type: Integer (0–3)
    - Description: Number of major vessels (0–3) colored by fluoroscopy. Transformed using one-hot encoding.
13. **Thal (Thallium Stress Test Results)**
    - Type: Categorical (3 for Normal, 6 for Fixed Defect, 7 for Reversible Defect)
    - Description: Results of thallium stress tests. Transformed using one-hot encoding.
14. **Oldpeak_Slope_Combined**
    - Type: Continuous
    - Description: A derived feature combining Oldpeak (ST depression) and Slope (ECG segment pattern during peak exercise).
15. **Gender**-Based Interaction Terms
    - Type: Continuous
    - Description: Interaction features created by multiplying the "Sex" feature with key variables like Chol and Trestbps to account for demographic-specific variations.

Critical Functions

1. **Log Transformer**
    - Purpose: Reduces skewness in variables like Chol, Trestbps, and Oldpeak.
    - Inputs: Skewed numerical features.
    - Outputs: Log-transformed features.
2. **Squared Transformation**
    - Purpose: Captures non-linear relationships in features like Thalach.
    - Inputs: Thalach feature.
    - Outputs: Squared-transformed feature.
3. **Combine Oldpeak and Slope**
    - Purpose: Creates a new feature to enhance model accuracy.
    - Inputs: Oldpeak and Slope features.
    - Outputs: Combined feature reflecting ST segment depression and slope interaction.
4. **Gender-Based Interaction Creation**
    - Purpose: Generates gender-specific interaction terms to capture the influence of demographic variations on key features.
    - Inputs: Sex feature and numerical features such as Chol and Trestbps.
    - Outputs: Interaction features highlighting gender-based relevance.

Classifiers Used

1. **Random Forest Classifier**
    - Purpose: Constructs an ensemble of decision trees for binary classification.
    - Features: Robust against overfitting, useful for datasets with imbalanced classes.
    - Implementation: Optimized using GridSearchCV to select parameters like the number of estimators, maximum depth, and feature importance.
2. **XGBoost Classifier**
    - Purpose: Gradient boosting algorithm designed for efficiency and performance in binary classification tasks.
    - Features: Focuses on minimizing loss functions with parallelized tree construction.
3. **Ensemble Method**

- Purpose: Combines predictions from Random Forest, XGBoost, and Logistic Regression to improve robustness.
- Features: Weighted averaging of classifiers to leverage strengths of individual models.

4. **Logistic Regression**
    - Purpose: Serves as a baseline model to compare linear relationships between features and outcomes.
    - Features: Interpretable and effective for datasets with linear separability.