Experiment 3: Implementation of Supervised Learning (Classification) on Heart Disease Dataset

1. Database Source

2. Dataset Description :

The experiment utilizes the **Heart Disease Dataset** (heart.csv).

- **Size:** The dataset contains **1,025 records** and **14 columns**.
- **Features:**
  - age: Age in years.
  - sex: (1 = male; 0 = female).
  - cp: Chest pain type (4 values).
  - trestbps: Resting blood pressure.
  - chol: Serum cholesterol in mg/dl.
  - fbs: Fasting blood sugar > 120 mg/dl.
  - restecg: Resting electrocardiographic results (values 0,1,2).
  - thalach: Maximum heart rate achieved.
  - exang: Exercise induced angina.
  - oldpeak: ST depression induced by exercise relative to rest.
  - slope: The slope of the peak exercise ST segment.
  - ca: Number of major vessels (0-3) colored by fluoroscopy.
  - thal: 0 = normal; 1 = fixed defect; 2 = reversible defect.
- **Target Variable:** target (0 = no heart disease, 1 = presence of heart disease).
- **Characteristics:** The dataset provides a comprehensive set of physiological indicators. It is a multivariate dataset that requires careful scaling due to the varying ranges of numerical features like chol and thalach.

3. Mathematical Formulation of the Algorithm :

For this classification task, **Logistic Regression** or **Support Vector Machines (SVM)** are commonly used.

- **Logistic Regression Hypothesis:**

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

- **Decision Boundary:** The model predicts 1 if $h_\theta(x) \geq 0.5$ and 0 otherwise.
- **Cost Function (Log Loss):**

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} [y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))]$$

Optimization is achieved using **Gradient Descent** to find the weights (θ) that minimize $J(\theta)$.

4. Algorithm Limitations :

- **Assumption of Linearity:** Assumes a linear decision boundary between the classes; if the relationship is highly non-linear, the model will underperform.
- **Sensitivity to Outliers:** Extreme values in features like chol can skew the decision boundary.
- **Independence of Features:** Assumes that the input features are not highly correlated (multicollinearity), which can lead to unstable coefficient estimates.
- **Feature Scaling:** Requires features to be on a similar scale for the optimization algorithm to converge efficiently.

5. Methodology / Workflow :

1. **Data Loading:** Ingest heart.csv using the Pandas library.
2. **Exploratory Data Analysis (EDA):** Check for missing values and visualize the distribution of classes.
3. **Data Preprocessing:**
   - Split the data into features (X) and target (y).
   - Apply **Standard Scaling** to numerical features to normalize the ranges of age, trestbps, chol, etc.
4. **Data Splitting:** Partition the dataset into **Training (80%)** and **Testing (20%)** sets.
5. **Model Training:** Instantiate the classifier and fit it to the training data.
6. **Model Evaluation:** Predict the target for the test set and calculate performance metrics.

6. Performance Analysis :

- **Evaluation Metrics:**
    - **Accuracy:** ~85% - 88% (Standard for this dataset).
    - **Precision:** Measures the accuracy of positive predictions.
    - **Recall (Sensitivity):** Crucial in medical contexts, measuring the ability to find all heart disease cases.
- **Interpretation:** A high recall is preferred in this experiment because missing a heart disease diagnosis (False Negative) is more critical than a false alarm (False Positive). The model demonstrates a strong ability to differentiate between healthy individuals and those with heart conditions based on clinical data.

7. Hyperparameter Tuning :

To optimize the model, **Regularization Strength (**C**)** and **Optimization Solvers** were tuned:

- **Process:** Used GridSearchCV to test various values of C (from 0.001 to 100) and different solvers like liblinear and lbfgs.
- **Impact:**
    - Tuning the C parameter helped prevent overfitting by penalizing large weights.
    - A smaller C (e.g., 0.1) provided better generalization on the test set, reducing the gap between training and testing accuracy.
    - The choice of solver ensured the mathematical optimization reached the global minimum efficiently.

Conclusion: The implementation of a supervised classification model on the heart disease dataset successfully predicted patient outcomes with high accuracy. The experiment highlights the importance of feature scaling and recall-oriented evaluation in medical machine learning applications.