Heart Disease Prediction Using Machine Learning

Introduction

This project focuses on predicting heart disease using machine learning algorithms. The dataset used is heart.csv, which contains multiple features related to heart health.

Dependencies

Ensure you have the following Python libraries installed:

- pandas
- numpy
- matplotlib
- plotly
- scikit-learn

Data Preprocessing

1. **Load Dataset**: The dataset is loaded using pandas.read_csv().
2. **Data Exploration**:
    - .info(), .describe(), .isna().sum() are used to understand dataset structure and check for missing values.
    - df['target'].value_counts(), df['chol'].value_counts(), df['sex'].value_counts() provide class distributions.
3. **Feature Selection**:
    - Dropped irrelevant columns to optimize model performance.
4. **Data Visualization**:
    - Histogram of age distribution using plotly.express.

Model Training and Evaluation

1. Linear Regression

- Model: Linear Regression
- Performance Metrics:
    - Mean Squared Error (MSE)
    - R-squared Score (R2)

2. Logistic Regression

- Model: Logistic Regression
- Performance Metrics:
    - Accuracy
    - Confusion Matrix

    - Precision, Recall, and F1-score

3. Decision Tree Classifier

- Model: Decision Tree Classifier
- Evaluated using accuracy scores.

4. K-Nearest Neighbors (KNN)

- Model: KNeighbors Classifier
- Performance Metrics:
    - Accuracy
    - Precision
    - Recall
    - F1-score

5. K-Means Clustering

- Model: KMeans
- Performance Metrics:
    - Inertia
    - Silhouette Score
    - Davies-Bouldin Score
    - Calinski-Harabasz Score
    - Homogeneity, Completeness, and V-measure
- Visualization of clusters using matplotlib.

6. Linear Discriminate Analysis (LDA)

- Model: Linear Discriminat Analysis
- Visualization of LDA-transformed data.

Conclusion

This project applies multiple machine learning models to classify heart disease. Performance metrics help evaluate model effectiveness, and data visualization enhances interpretability.