**Heart Disease Prediction using Logistic Regression and K-Nearest Neighbors (KNN)**

**Overview**

This project aims to predict the presence of heart disease using Logistic Regression (LR) and K-nearest neighbours (KNN) classifiers. The models are evaluated using a variety of performance metrics, including accuracy, precision, recall, F1-score, and the ROC-AUC curve. The project involves feature selection to enhance model performance and interpretability.

**Project Objectives**

* Develop predictive models using Logistic Regression and KNN to classify patients as having heart disease or not.
* Evaluate the models using relevant performance metrics and visualize results.
* Identify the most important features influencing the predictions using Recursive Feature Elimination (RFE).

**Dataset**

The dataset used for this project is the **Heart Disease Dataset**, which includes various cardiovascular health indicators to determine the presence of heart disease (Target column).

**Key Features**

1. **Age**: Age of the patient.
2. **Sex**: Gender of the patient (1 = male, 0 = female).
3. **ChestPainType**: Type of chest pain experienced (4 types).
4. **RestingBP**: Resting blood pressure (in mm Hg).
5. **Cholesterol**: Serum cholesterol in mg/dl.
6. **FastingBS**: Fasting blood sugar > 120 mg/dl (1 = true, 0 = false).
7. **RestingECG**: Resting electrocardiographic results.
8. **MaxHR**: Maximum heart rate achieved.
9. **ExerciseAngina**: Exercise-induced angina (1 = yes, 0 = no).
10. **Oldpeak**: ST depression induced by exercise relative to rest.
11. **ST\_Slope**: Slope of the peak exercise ST segment.
12. **Target**: Presence of heart disease (1 = disease, 0 = no disease).

**Data Preprocessing**

* **Handling Missing Values**: No missing values were reported, assuming clean data.
* **Feature Scaling**: StandardScaler was used to normalize the feature values for improved model performance.
* **Train-Test Split**: The data was split into training (70%) and testing (30%) sets using stratified sampling to maintain class balance.

**Methodology**

**1. Logistic Regression (LR)**

* **Model Training**: A Logistic Regression model was trained using GridSearchCV to find the best hyperparameters (C and solver). GridSearchCV was used with 5-fold cross-validation.
* **Feature Selection**: Recursive Feature Elimination (RFE) was used to identify the most important features for the Logistic Regression model. The selected features were:
  + **Sex**
  + **ChestPainType**
  + **ExerciseInducedAngina**
  + **STDepression**
  + **NumberOfMajorVessels**
  + **Thalassemia**

**2. K-Nearest Neighbors (KNN)**

* **Model Training**: A KNN model was trained with n\_neighbors=5. The model was evaluated on both the full feature set and the selected features from RFE.

**Model Performance**

**Logistic Regression Model:**

Features selected: ['Sex', 'ChestPainType', 'ExerciseInducedAngina', 'STDepression', 'NumberOfMajorVessels', 'Thalassemia']

**Performance Metrics:**

**Accuracy:** 84%

**Precision:**

* Class 0: 0.88
* Class 1: 0.80

**Recall:**

* Class 0: 0.78
* Class 1: 0.90

**F1-Score:**

* Class 0: 0.83
* Class 1: 0.85

**Overall Weighted Scores:** Precision, Recall, and F1-Score: 0.84

**K-Nearest Neighbors (KNN) Model:**

**Performance Metrics:**

**Accuracy:** 88.31%

**Precision:**

* Class 0: 0.88
* Class 1: 0.89

**Recall:**

* Class 0: 0.89
* Class 1: 0.88

**F1-Score:**

* Class 0: 0.88
* Class 1: 0.88

**Overall Weighted Scores:** Precision, Recall, and F1-Score: 0.88

**Key Insights**

1. **Selected Features**: The features selected by RFE were highly indicative of heart disease, emphasizing specific clinical factors such as chest pain type, exercise-induced angina, and major vessels affected.
2. **Model Comparison**: KNN slightly outperformed Logistic Regression in terms of accuracy and F1-score, indicating a more nuanced handling of the data distribution. However, Logistic Regression provided interpretability and robustness, particularly in identifying critical features.
3. **Class Imbalance Handling**: Both models maintained consistent performance across classes, with slightly higher recall for class 1 in Logistic Regression, demonstrating a stronger sensitivity to detecting positive cases (heart disease).
4. **ROC Curve Analysis**: Logistic Regression exhibited a reliable trade-off between sensitivity and specificity, as visualized in the ROC curve, providing confidence in its predictive capacity even when probabilities are adjusted.

**Visualizations**

* **Correlation Heatmap**: Displayed correlations between all features, highlighting the strongest relationships.
* **Confusion Matrices**: For LR and KNN, showing the distribution of correct and incorrect predictions.
* **ROC Curve**: Illustrated the performance of Logistic Regression with an AUC score.

**Conclusion**

This project successfully developed and evaluated models to predict heart disease using Logistic Regression and KNN. Both models demonstrated strong performance, with KNN achieving slightly higher accuracy. Feature selection played a crucial role in enhancing model performance and interpretability, providing insights into the most influential predictors of heart disease.

**Future Work**

* Explore additional classification models like Support Vector Machines (SVM) or Random Forest for comparative analysis.
* Consider advanced feature engineering techniques or dimensionality reduction (e.g., PCA) to further refine model accuracy.
* Deploy the model as a user-friendly application for real-time heart disease risk assessment.