**Heart Disease Prediction Report**

**Introduction**

Heart disease is one of the leading causes of death worldwide, and early detection is essential for effective treatment. Machine learning methods can support medical professionals by predicting the likelihood of heart disease based on patient data. In this study, a heart disease dataset was analyzed using two classification models: **Random Forest Classifier** and **Logistic Regression**.

The dataset contains information such as **age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG results, maximum heart rate, exercise-induced angina, ST depression (Oldpeak), and ST slope**. The target variable is **HeartDisease (0 = No, 1 = Yes)**.

**Data Preprocessing**

* The dataset originally included a serial number column (Unnamed: 0) which was removed.
* Categorical variables (e.g., Sex, ChestPainType, RestingECG, ExerciseAngina, ST\_Slope) were encoded using **OneHotEncoder**.
* The data was then scaled using **StandardScaler** to normalize numeric features.
* Data was split into **training (80%) and testing (20%) sets** for evaluation.

**Models Used**

**Logistic Regression**

Logistic Regression is a linear model commonly used for binary classification. It estimates probabilities using the logistic function and applies a threshold (default = 0.5) to classify outcomes.

**Random Forest Classifier**

Random Forest is an ensemble learning method that builds multiple decision trees and averages their predictions. It typically provides higher accuracy and robustness by reducing overfitting compared to a single decision tree.

**Results**

**Confusion Matrices**

**Random Forest (Accuracy = 84.76%)**

[[36 13]

[ 7 49]]

**Logistic Regression (Accuracy = 80.95%)**

[[36 13]

[ 7 49]]

Both models produced the same confusion matrix, meaning they made identical predictions on the test set. The difference in accuracy is due to slight differences in how probabilities were handled.

**ROC Curve & AUC**

* Random Forest achieved a higher **AUC score**, indicating better discrimination between positive and negative cases.
* Logistic Regression showed competitive performance but slightly lower AUC.

**Discussion**

* Both models correctly identified most positive and negative cases of heart disease.
* **Random Forest** performed slightly better with an accuracy of ~85%, compared to Logistic Regression’s ~81%.
* Logistic Regression, being simpler and easier to interpret, is still valuable in medical settings.
* Random Forest provides better overall predictive power and can capture complex, non-linear relationships in the data.

**Conclusion**

This study demonstrates that machine learning models can successfully predict heart disease using clinical and demographic features. While Logistic Regression provides interpretability and simplicity, Random Forest achieved higher accuracy and AUC, making it more reliable for predictive performance.

For practical applications, a combination of models may be used: Logistic Regression for medical interpretability and Random Forest for improved prediction accuracy.