Assignment #04_Classification_Analysis

Dataset Summary:

The **heart disease dataset** has 918 entries and 12 columns, consisting of patient data related to cardiovascular health.This dataset is used to predict heart disease risk based on patient health and symptom indicators. Each feature, from cholesterol levels to chest pain types, gives an insight of dataset, allowing for the development of predictive models aimed at early heart disease detection and prevention.

Input/ Independent Variables:

1. **Age**: Age of the patient (integer).
2. **Sex**: Gender of the patient (M for Male, F for Female).
3. **ChestPainType**: Type of chest pain experienced, with categories such as ATA (Atypical Angina), NAP (Non-Anginal Pain), ASY (Asymptomatic), and TA (Typical Angina).
4. **RestingBP**: Resting blood pressure (in mm Hg).
5. **Cholesterol**: Serum cholesterol level (in mg/dL).
6. **FastingBS**: Fasting blood sugar (1 if fasting blood sugar > 120 mg/dL, 0 otherwise).
7. **RestingECG**: Resting electrocardiographic results with types like Normal, ST (ST-T wave abnormality), and LVH (left ventricular hypertrophy).
8. **MaxHR**: Maximum heart rate achieved.
9. **ExerciseAngina**: Exercise-induced angina (Y for Yes, N for No).
10. **Oldpeak**: ST depression induced by exercise relative to rest.
11. **ST_Slope**: Slope of the peak exercise ST segment, categorized as Up, Flat, or Down.

Output/ Dependent Variables:

- **HeartDisease**: Target variable indicating the presence (1) or absence (0) of heart disease.

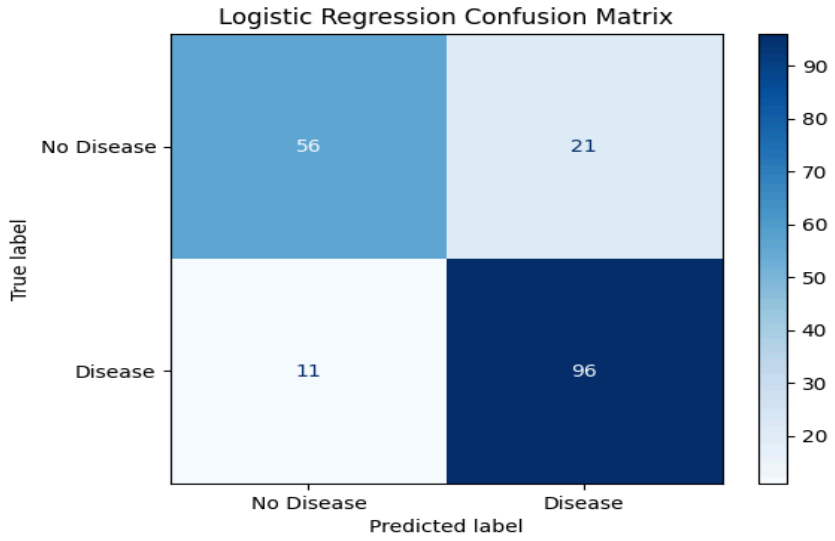
Confusion matrix and accuracy score:

Logistic Regression Accuracy (%): 82.60

Confusion Matrix for Logistic Regression:

[[56 21]

[11 96]]

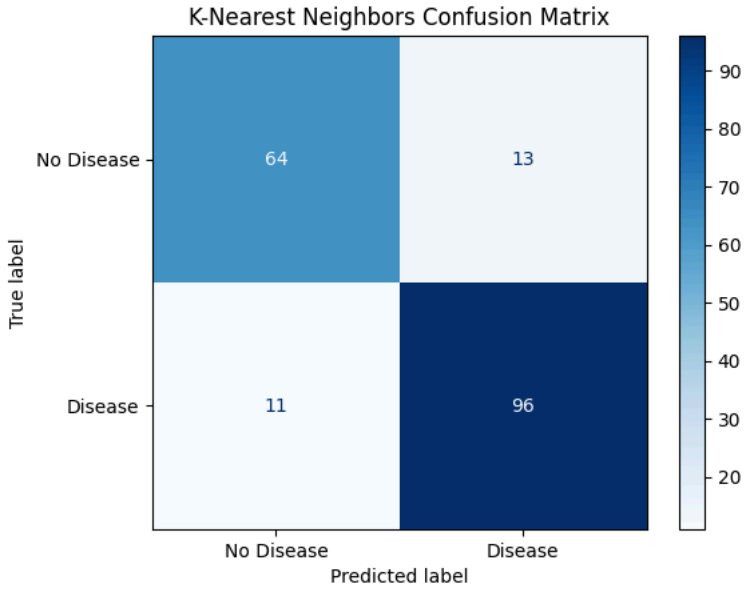


K-Nearest Neighbors Accuracy (%): 86.95

Confusion Matrix for K-Nearest Neighbors:

[[64 13]

[11 96]]

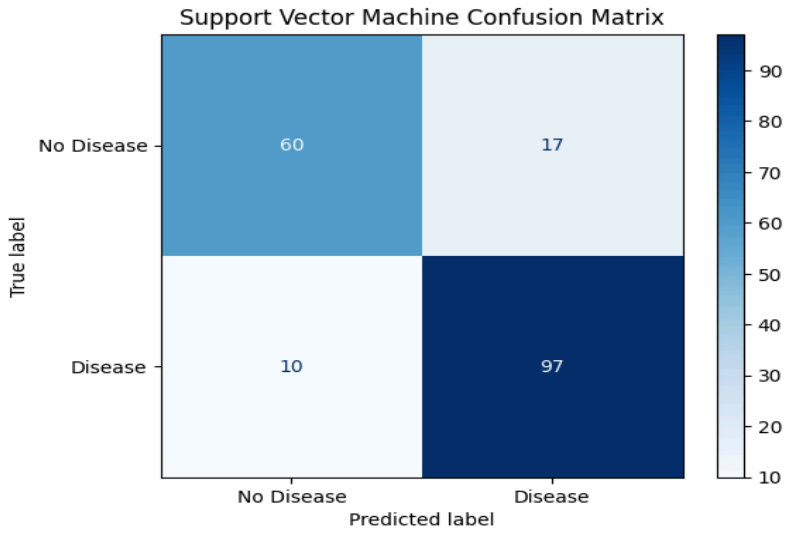


Support Vector Machine Accuracy (%): 85.32

Confusion Matrix for Support Vector Machine:

[[60 17]

[10 97]]



Conclusion:

Among the three models, **K-Nearest Neighbors (K-NN)** demonstrated the best performance with the highest accuracy of **86.96%** and fewer prediction errors. This suggests that K-NN is the most effective model for predicting heart disease within this dataset, as it maintains a good balance of true positive and true negative predictions, critical for accurate patient risk classification.

Both **Support Vector Machine (SVM)** and **Logistic Regression** also performed reasonably well, with accuracies of **85.33%** and **82.61%**, respectively. While they offer competitive accuracy, K-NN outperformed them slightly by better distinguishing between patients with and without heart disease, with lower false negatives. Therefore, K-NN is recommended as the preferred model for this heart disease dataset, as it minimizes missed heart disease cases while maintaining high accuracy.