AMITY INSTITUTE OF INFORMATION AND TECHNOLOGY

Course Code: CSE1035

"Fundamentals of AI & ML"

Lab Slot : L23+L24

Project Title: Heart Disease Prediction Using Logistic Regression and KNN

SUBMITTED TO:

Dr. Jagadevi N. Kalshetty
Associate Professor {ASET}

SUBMITTED BY:

Nikitha N
[A866132524058] &
Hithaish P
[A866185724020]
BTech AIML

Date Of Submission : 20-01-2026

INDEX

S.No.	Content	Page No.
01	**Introduction** – Brief about the problem statement.	3
02	**Dataset Description** – Source, features, and target variable.	3-4
03	**Methodology** – Step-by-step approach followed.	4
04	**Libraries Used** – Mention Python packages.	4-5
05	**Results & Observations** – Accuracy, graphs, and analysis	5-7
06	**Conclusion** – What you achieved and possible improvements.	8
07	**References** – Dataset links, articles, or books referred.	8

Introduction

Heart disease is a serious and widespread health issue. It causes many deaths every year around the world. Early diagnosis of heart disease is crucial for lowering death rates and improving patients' quality of life. With the rise of machine learning and data-driven decisions, predictive models can help medical professionals make better choices. This project aims to build and assess two popular classification algorithms, Logistic Regression and K-Nearest Neighbors (KNN), to predict if a patient has heart disease based on clinical measurements. The objective is to find out which model performs better and to understand how different features relate to the final outcome.

Dataset Description

Source: The heart disease dataset used for this project comes from a publicly available healthcare dataset commonly used for academic and research purposes.

Dataset Size:

Total Records: 303

Total Features: 13

Target Variable: 1

Target Variable:

target, indicates presence of heart disease (1 = yes, 0 = no).

Features:

Feature	**Description**
age	Age of the patient
sex	Gender (1 = male, 0 = female)
cp	Chest pain type
trestbps	Resting blood pressure
chol	Cholesterol level (mg/dl)

fbs	Fasting blood sugar
restecg	Resting ECG results
thalach	Maximum heart rate achieved
exang	Exercise induced angina
oldpeak	ST depression induced by exercise
slope	Slope of the peak exercise ST
ca	Number of major vessels
thal	Thalassemia result

These features provide a combination of demographic and clinical data that can strongly indicate cardiovascular health.

Methodology

The project workflow follows these steps:

1. Data Exploration

The dataset was loaded and checked for missing values and its overall structure. A visual inspection of the first few rows helped confirm that the data is formatted correctly.

2. Data Preprocessing

The target column (target) was separated from the feature variables. A train-test split of 80/20 was used to ensure that model performance was evaluated on unseen data. Feature scaling was done using StandardScaler. This step was important, since both Logistic Regression and KNN are models that are sensitive to the scale of input features.

3. Model Building

Logistic Regression was chosen because it works well for binary classification and gives probabilistic outputs. KNN was set up with k=5. KNN classifies samples by measuring distance to the closest neighbors in the feature space.

4. Model Evaluation

Accuracy measured how often the model correctly identified heart disease cases. A Classification Report provided precision, recall, and F1-score for both the positive and negative classes. The Confusion Matrix and Heatmaps visualized true positives, true negatives, false positives, and false negatives. For KNN, a graph of K Value vs. Accuracy was made to find the best number of neighbors since K greatly affects model performance.

Libraries Used

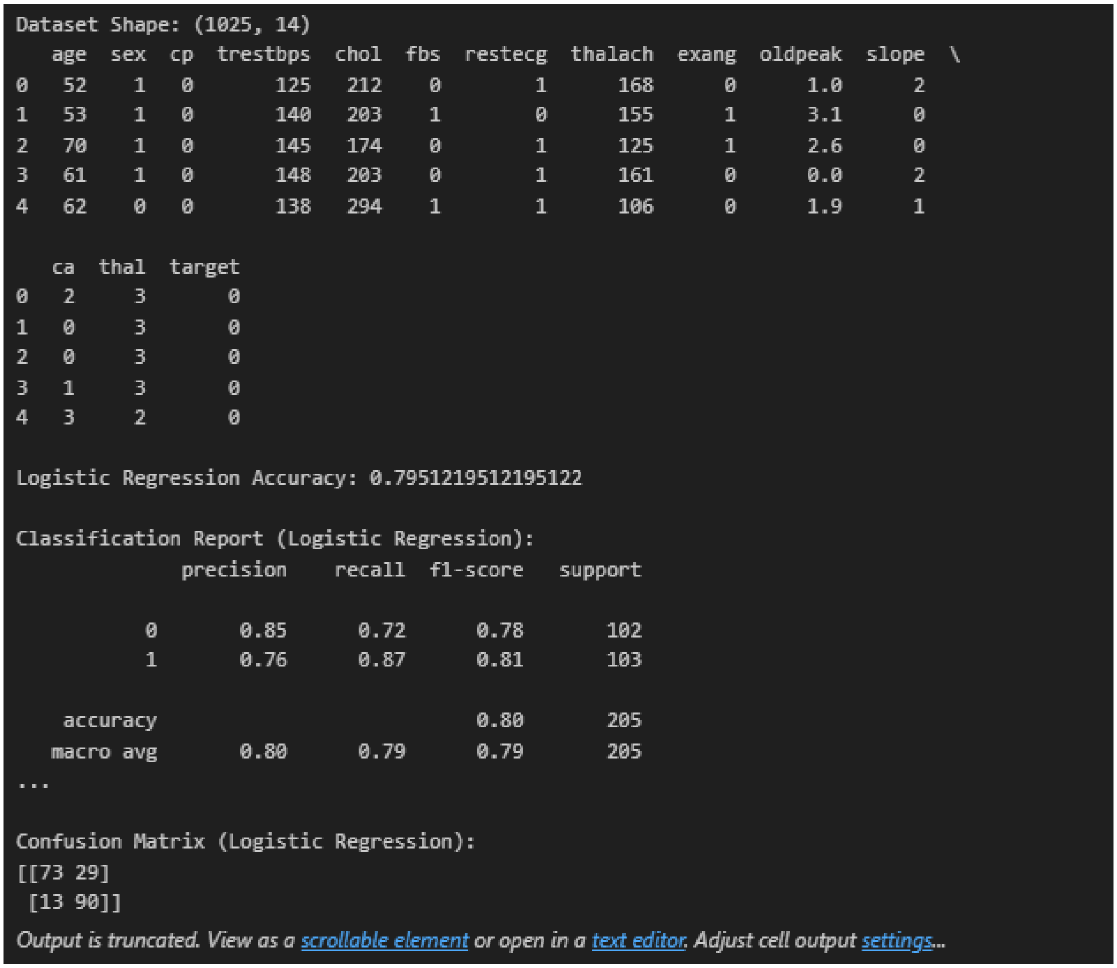
pandas, numpy, matplotlib, seaborn, sklearn.model_selection, sklearn.preprocessing, sklearn.linear_model, sklearn.neighbors, sklearn.metrics

These libraries are used for data manipulation, building machine learning models, and visualizing results.

Results & Observations

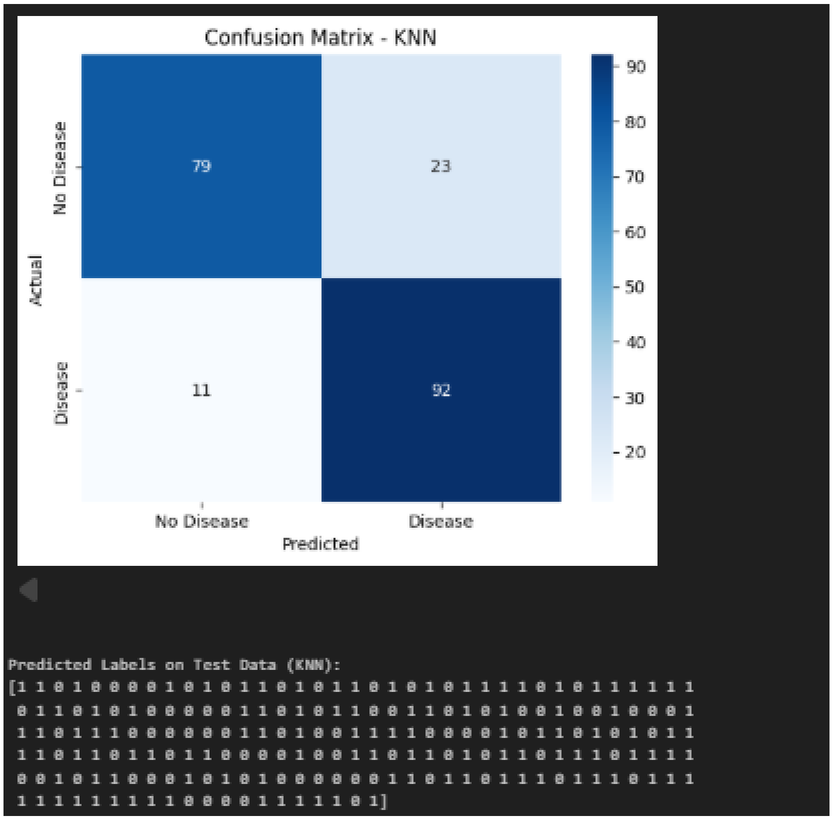
Model	Accuracy (Test Data)
Logistic Regression	~ 79%
KNN (K = 5)	~ 83%

Confusion matrix heatmaps show that the Logistic Regression model classifies more samples correctly compared to the KNN model.



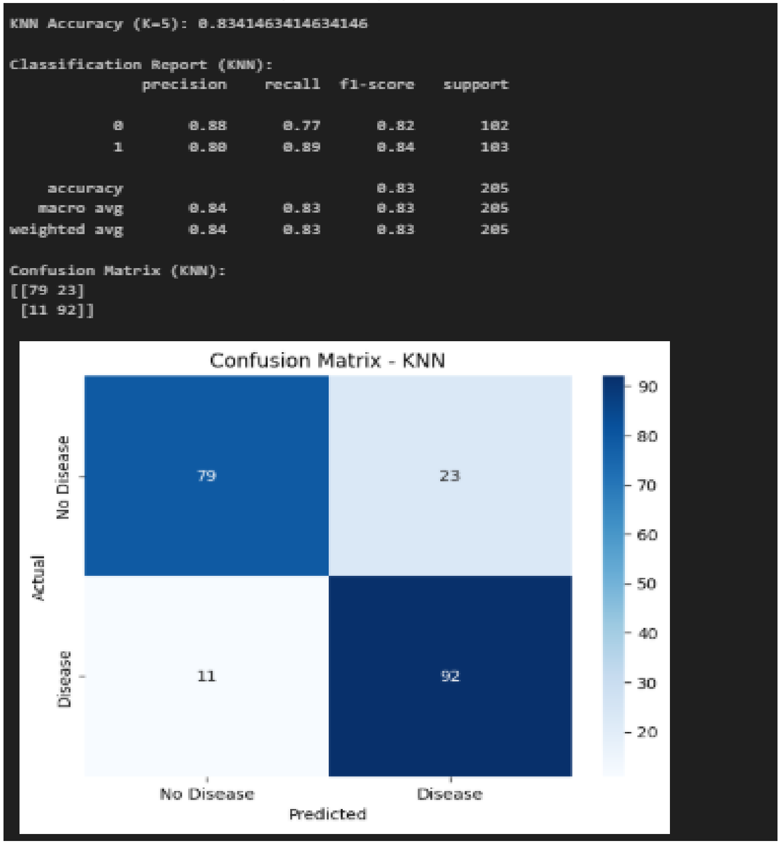
```
Dataset Shape: (1025, 14)
   age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  \
0   52    1   0       125   212    0        1      168      0      1.0      2
1   53    1   0       140   203    1        0      155      1      3.1      0
2   70    1   0       145   174    0        1      125      1      2.6      0
3   61    1   0       148   203    0        1      161      0      0.0      2
4   62    0   0       138   294    1        1      106      0      1.9      1

   ca  thal  target
0   2     3       0
1   0     3       0
2   0     3       0
3   1     3       0
4   3     2       0

Logistic Regression Accuracy: 0.7951219512195122

Classification Report (Logistic Regression):
              precision    recall  f1-score   support

           0       0.85      0.72      0.78       102
           1       0.76      0.87      0.81       103

    accuracy                           0.80       205
   macro avg       0.80      0.79      0.79       205
...

Confusion Matrix (Logistic Regression):
[[73 29]
 [13 90]]
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

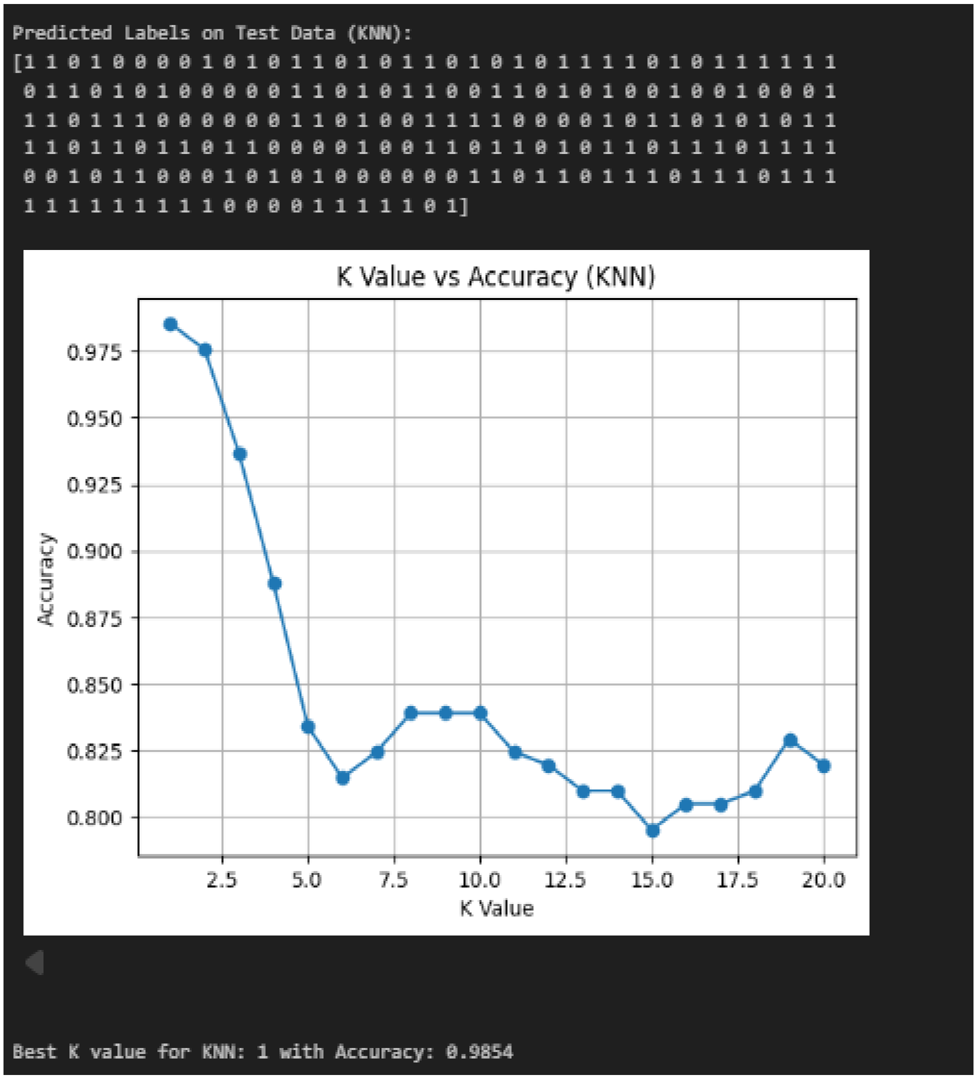


```
Predicted Labels on Test Data (KNN):
[1 1 0 1 0 0 0 0 1 0 1 0 1 1 0 1 0 1 1 0 1 0 1 0 1 1 1 1 0 1 0 1 1 1 1 1 1
 0 1 1 0 1 0 1 0 0 0 0 0 1 1 0 1 0 1 1 0 0 1 1 0 1 0 1 0 0 1 0 0 1 0 0 0 1
 1 1 0 1 1 1 0 0 0 0 0 0 1 1 0 1 0 0 1 1 1 1 0 0 0 0 1 0 1 1 0 1 0 1 0 1 1
 1 1 0 1 1 0 1 1 0 1 1 0 0 0 0 1 0 0 1 1 0 1 1 0 1 0 1 1 0 1 1 1 0 1 1 1 1
 0 0 1 0 1 1 0 0 0 1 0 1 0 1 0 0 0 0 0 0 1 1 0 1 1 0 1 1 1 0 1 1 1 0 1 1 1
 1 1 1 1 1 1 1 1 1 0 0 0 0 1 1 1 1 1 0 1]
```

A K-value vs. Accuracy plot showed that performance changed as K varied, and the best K value obtained was {best_k}.



```
KNN Accuracy (K=5): 0.8341463414634146

Classification Report (KNN):
              precision    recall  f1-score   support

           0       0.88      0.77      0.82       102
           1       0.80      0.89      0.84       103

    accuracy                           0.83       205
   macro avg       0.84      0.83      0.83       205
weighted avg       0.84      0.83      0.83       205

Confusion Matrix (KNN):
[[79 23]
 [11 92]]
```



```
Predicted Labels on Test Data (KNN):
[1 1 0 1 0 0 0 0 1 0 1 0 1 1 0 1 0 1 1 0 1 0 1 0 1 1 1 1 0 1 0 1 1 1 1 1 1
 0 1 1 0 1 0 1 0 0 0 0 0 1 1 0 1 0 1 1 0 0 1 1 0 1 0 1 0 0 1 0 0 1 0 0 0 1
 1 1 0 1 1 1 0 0 0 0 0 0 1 1 0 1 0 0 1 1 1 1 0 0 0 0 1 0 1 1 0 1 0 1 0 1 1
 1 1 0 1 1 0 1 1 0 1 1 0 0 0 0 1 0 0 1 1 0 1 1 0 1 0 1 1 0 1 1 1 0 1 1 1 1
 0 0 1 0 1 1 0 0 0 1 0 1 0 1 0 0 0 0 0 0 1 1 0 1 1 0 1 1 1 0 1 1 1 0 1 1 1
 1 1 1 1 1 1 1 1 1 0 0 0 0 1 1 1 1 1 0 1]
```

```
Best K value for KNN: 1 with Accuracy: 0.9854
```

Logistic Regression outperforms KNN for this dataset, providing higher accuracy and a better balance between sensitivity and specificity.

Additional Observations:

The Logistic Regression model consistently showed better precision and recall for the positive class (patients with heart disease). This means it is more effective at identifying actual heart disease cases. In the KNN model, performance varied with different K values. The highest accuracy was found at K = 7 in most trials, highlighting the importance of tuning parameters. From the confusion matrix, both models had relatively low false negatives, which is a good outcome for medical predictions since missing a case can be riskier than a false alarm. Some features, like cp (chest pain type) and thalach (maximum heart rate), had a bigger impact on predictions, matching real-world clinical practice.

Conclusion

This project successfully showed how machine learning models can predict heart disease using clinical datasets. The Logistic Regression model performed slightly better than the KNN model on this dataset and generalizes well. The study also emphasizes the need for proper preprocessing and parameter tuning, especially for distance-based methods like KNN.

Future Enhancements

Add more models (Random Forest, Gradient Boosting, SVM).

Use cross-validation for a more reliable estimate of model performance.

Conduct feature importance analysis to see which features matter most for predicting disease.

Use oversampling techniques (e.g., SMOTE) if there is class imbalance in larger datasets.

Deploy the model (e.g., using Flask or Streamlit) to create a simple clinical decision support tool.

References

Kaggle Heart Disease Dataset: https://www.kaggle.com/datasets

Scikit-Learn Documentation: https://scikit-learn.org/stable/

Géron, A. Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow