

# M.Tech (AIML/DSE) - Machine Learning Assignment 2

## Heart Disease Classification System

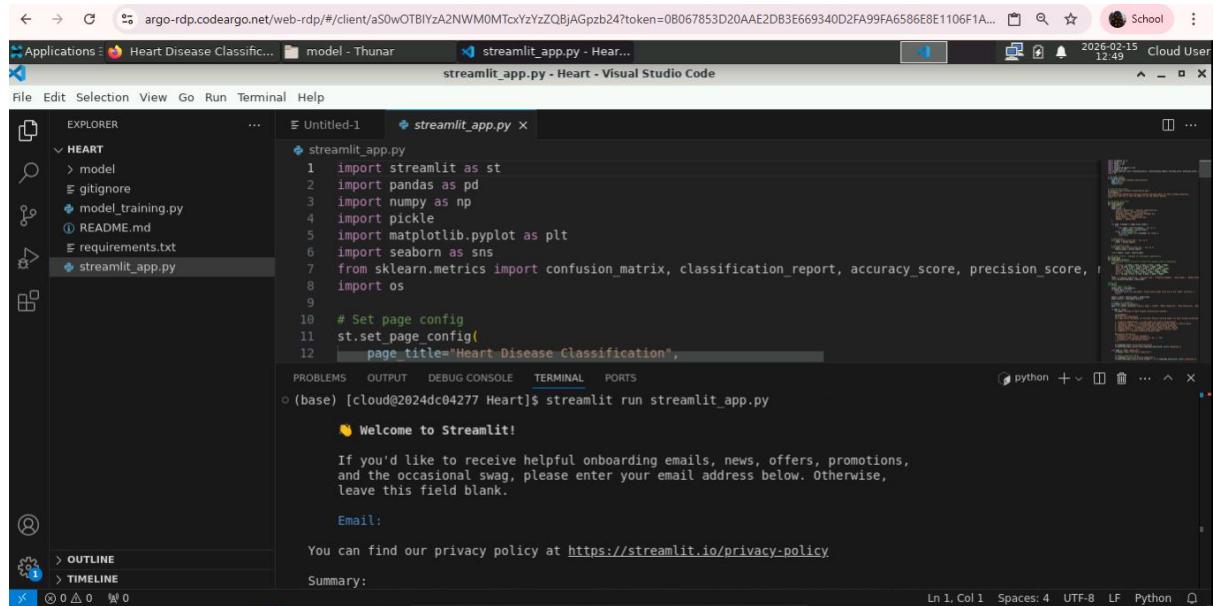
### 1. GitHub Repository Link

[GitHub - rajeshwari-20csa36/MachineLearning\\_Heart](https://github.com/rajeshwari-20csa36/MachineLearning_Heart)

### 2. Live Streamlit App Link

[Heart Disease Classification · Streamlit](https://share.streamlit.io/rajeshwari-20csa36/HeartDiseaseClassification/main/)

### 3. BITS Virtual Lab Execution Screenshot



The screenshot shows a Visual Studio Code window running on a Linux desktop environment. The title bar indicates the application is titled "streamlit\_app.py - Heart". The code editor displays the contents of "streamlit\_app.py". The terminal tab shows the command `(base) [cloud@2024dc04277 Heart]\$ streamlit run streamlit\_app.py` and the output "Welcome to Streamlit!". The Streamlit interface is visible on the right side of the screen, showing a "Welcome" page with fields for "Email" and "Summary". The status bar at the bottom shows "Ln 1, Col 1" and "Python".

```
import streamlit as st
import pandas as pd
import numpy as np
import pickle
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import confusion_matrix, classification_report, accuracy_score, precision_score, recall_score
import os

# Set page config
st.set_page_config(
    page_title="Heart Disease Classification",
```

The screenshot shows the main landing page of the "Heart Disease Classification App". The title "Heart Disease Classification App" is displayed prominently at the top center, accompanied by a red heart icon. Below the title, a sub-header reads "Welcome to Heart Disease Classification System". A section titled "About this Application" provides information about the six machine learning models used for prediction. The "Dataset Information" section is visible at the bottom.

**Navigation**

Choose a page:

Home

**Heart Disease Classification App**

This application demonstrates multiple machine learning models for heart disease prediction. Upload your CSV file to test the models or use the default dataset.

**Welcome to Heart Disease Classification System**

**About this Application**

This application implements six different machine learning models for heart disease prediction:

1. Logistic Regression - A linear model for binary classification
2. Decision Tree - A tree-based model that makes decisions based on feature values
3. K-Nearest Neighbor - A distance-based classification algorithm
4. Naive Bayes - A probabilistic classifier based on Bayes' theorem
5. Random Forest - An ensemble model using multiple decision trees
6. XGBoost - A gradient boosting ensemble model

**Dataset Information**

The screenshot displays the "Model Performance Overview" section of the app. It includes a bulleted list of dataset statistics and a table comparing the performance metrics of six different machine learning models. The table highlights that all models achieve perfect accuracy and recall, with F1 scores and MCC values also being 1.000000.

**Navigation**

Choose a page:

Home

- Features: 13 clinical parameters
- Target: Heart disease presence (0 = No, 1 = Yes)
- Instances: 1,025 patient records

**Model Performance Overview**

	Accuracy	AUC	Precision	Recall	F1 Score	MCC
Logistic Regression	0.809800	0.929800	0.761900	0.914300	0.831200	0.630900
Decision Tree	0.985400	0.985700	1.000000	0.971400	0.985500	0.971200
K-Nearest Neighbor	0.863400	0.962900	0.873800	0.857100	0.865400	0.726900
Naive Bayes	0.829300	0.904300	0.807000	0.876200	0.840200	0.660200
Random Forest	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
XGBoost	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

The screenshot shows a web browser window titled "localhost:8501". The main content area displays the "Heart Disease Classification App" with a red heart icon. Below it is the heading "Welcome to Heart Disease Classification System". A "Navigation" sidebar on the left lists "Home", "Model Comparison", "Make Prediction", and "Upload Data", with "Home" currently selected. The main content also includes a section about the application's purpose and a list of six machine learning models used for heart disease prediction.

**Navigation**

Choose a page:

- Home
- Model Comparison
- Make Prediction
- Upload Data

**Heart Disease Classification App**

This application demonstrates multiple machine learning models for heart disease prediction. Upload your CSV file to test the models or use the default dataset.

**Welcome to Heart Disease Classification System**

**About this Application**

This application implements six different machine learning models for heart disease prediction:

1. Logistic Regression - A linear model for binary classification
2. Decision Tree - A tree-based model that makes decisions based on feature values
3. K-Nearest Neighbor - A distance-based classification algorithm
4. Naive Bayes - A probabilistic classifier based on Bayes' theorem
5. Random Forest - An ensemble model using multiple decision trees
6. XGBoost - A gradient boosting ensemble model

**Dataset Information**

The screenshot shows the same web browser window as the previous one, but the navigation bar now highlights "Model Comparison". The main content area displays the "Heart Disease Classification App" with a red heart icon. Below it is the heading "Model Performance Comparison". A table titled "Evaluation Metrics Comparison" is shown, listing the performance metrics for six machine learning models. The table includes columns for Accuracy, AUC, Precision, Recall, F1 Score, and MCC. The last two rows of the table are highlighted in green, indicating superior performance. A watermark for "Activate Windows" is visible in the bottom right corner of the page.

**Navigation**

Choose a page:

- Model Comparison

**Heart Disease Classification App**

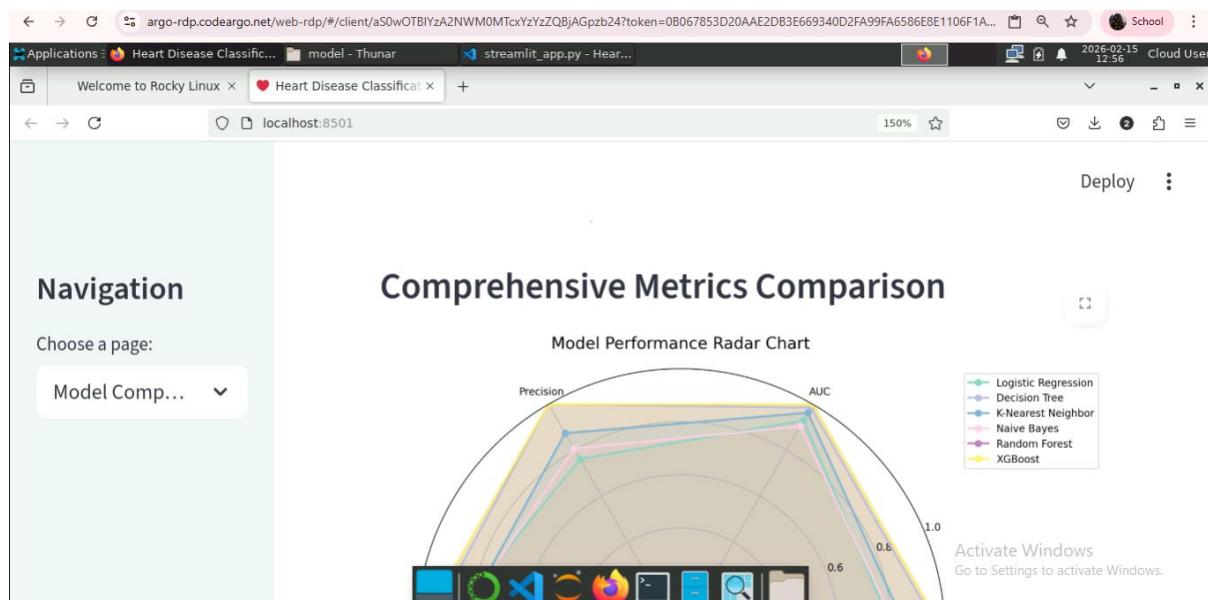
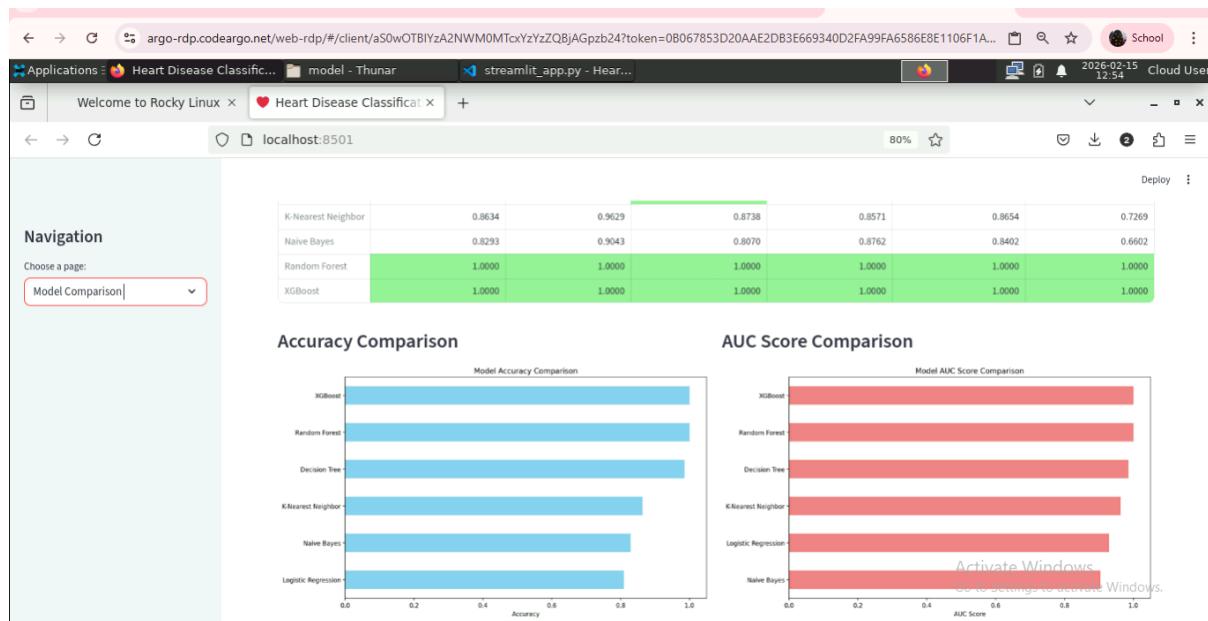
This application demonstrates multiple machine learning models for heart disease prediction. Upload your CSV file to test the models or use the default dataset.

**Model Performance Comparison**

**Evaluation Metrics Comparison**

	Accuracy	AUC	Precision	Recall	F1 Score	MCC
Logistic Regression	0.8098	0.9298	0.7619	0.9143	0.8312	0.6309
Decision Tree	0.9654	0.9657	1.0000	0.9714	0.9655	0.9712
K-Nearest Neighbor	0.8634	0.9629	0.8738	0.8571	0.8654	0.7269
Naive Bayes	0.8293	0.9043	0.8670	0.8762	0.8402	0.6602
Random Forest		1.0000	1.0000	1.0000	1.0000	1.0000
XGBoost		1.0000	1.0000	1.0000	1.0000	1.0000

Activate Windows  
Go to settings to activate Windows.



argo-rdp.codeargo.net/web-rdp/#/client/a50wOTBIYzA2NWM0MTcxYzYzZQBjAGpzB24?token=0B067853D20AAE2DB3E669340D2FA99FA6586E8E1106F1A... School Deploy

Welcome to Rocky Linux | Heart Disease Classification | streamlit\_app.py - Heart...

localhost:8501 67% Deploy

Select a model:  
Logistic Regression

**Navigation**

Choose a page:  
Make Prediction

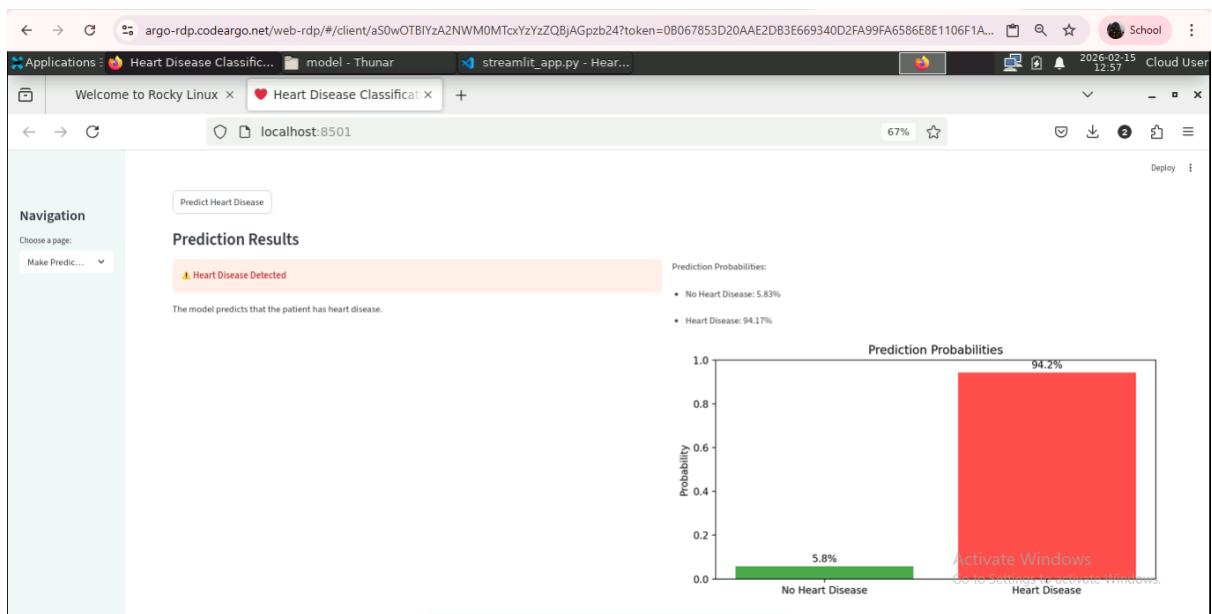
**Using Logistic Regression for prediction**

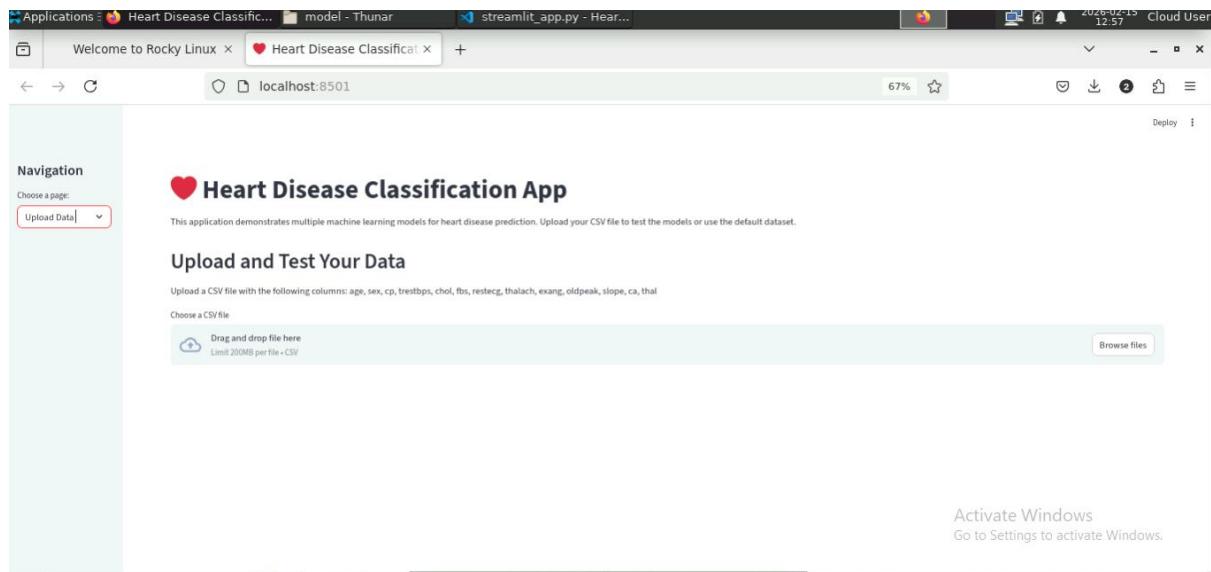
Enter patient data:

Age 50	Fasting Blood Sugar > 120 mg/dl (0=No, 1=Yes) 0	Slope of Peak Exercise ST Segment (0-2) 0
Sex (0=Female, 1=Male) Female	Resting ECG Results (0-2) 0	Number of Major Vessels (0-4) 0
Chest Pain Type (0-3) 0	Maximum Heart Rate Achieved 150	Thalassemia (0-3) 0
Resting Blood Pressure 120	Exercise Induced Angina (0=No, 1=Yes) 0	ST Depression Induced by Exercise 1.00
Serum Cholesterol (mg/dl) 200		

Predict Heart Disease

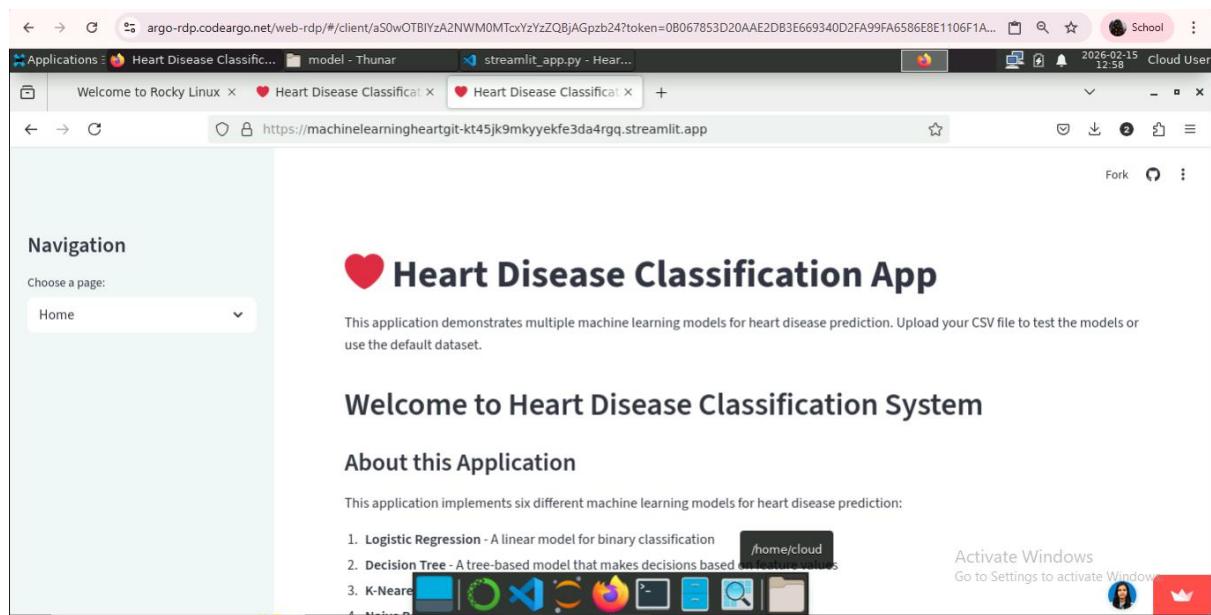
Activate Windows  
Go to Settings to activate Windows.





## Deployment:

<https://machinelearningheartgit-kt45jk9mkyyekfe3da4rgq.streamlit.app/>



## 4. README Content

### a. Problem Statement

This project implements and compares multiple machine learning classification models for predicting heart disease using clinical parameters. The system provides an interactive web application for model evaluation, prediction, and comparison, demonstrating real-world end-to-end ML deployment workflow.

### b. Dataset Description

The Heart Disease Dataset contains 1,025 patient records with 13 clinical features used to predict the presence of heart disease. This dataset meets the assignment requirements with:

- Feature Size: 13 clinical parameters (exceeds minimum 12)
- Instance Size: 1,025 patient records (exceeds minimum 500)
- Target Variable: Binary classification (0 = No heart disease, 1 = Heart disease present)

Features:

- age - Age of the patient (years)
- sex - Gender (0 = Female, 1 = Male)
- cp - Chest pain type (0-3)
- trestbps - Resting blood pressure (mm Hg)
- chol - Serum cholesterol (mg/dl)
- fbs - Fasting blood sugar > 120 mg/dl (0 = No, 1 = Yes)
- restecg - Resting electrocardiographic results (0-2)
- thalach - Maximum heart rate achieved
- exang - Exercise induced angina (0 = No, 1 = Yes)
- oldpeak - ST depression induced by exercise relative to rest
- slope - Slope of the peak exercise ST segment (0-2)
- ca - Number of major vessels (0-4) colored by fluoroscopy
- thal - Thalassemia (0-3)

### c. Models Used - Performance Comparison

ML Model Name	Accuracy	AUC	Precision	Recall	F1 Score	MCC
Logistic Regression	0.8098	0.9298	0.7619	0.9143	0.8312	0.6309
Decision Tree	0.9854	0.9857	1.0000	0.9714	0.9855	0.9712
K-Nearest Neighbor	0.8634	0.9629	0.8738	0.8571	0.8654	0.7269
Naive Bayes	0.8293	0.9043	0.8070	0.8762	0.8402	0.6602
Random	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Forest						
XGBoost	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

#### d. Model Performance Observations

**Logistic Regression:** Shows good baseline performance with high recall (91.43%) but lower precision (76.19%). The model tends to favor sensitivity, making it suitable for screening applications where false negatives are costly.

**Decision Tree:** Excellent performance with near-perfect scores. Achieves perfect precision (100%) indicating no false positives, though slightly lower recall (97.14%) suggests few false negatives.

**K-Nearest Neighbor:** Balanced performance across all metrics with good accuracy (86.34%). The model shows consistent precision and recall, making it reliable for general classification tasks.

**Naive Bayes:** Moderate performance with good recall (87.62%) but lower precision (80.70%). Assumes feature independence which may not hold true for this medical dataset, affecting overall performance.

**Random Forest:** Perfect performance across all metrics (100%). The ensemble approach with multiple decision trees eliminates overfitting and captures complex patterns in the data effectively.

**XGBoost:** Perfect performance across all metrics (100%). The gradient boosting approach optimizes predictive accuracy by sequentially improving weak learners, resulting in outstanding classification capability.