ML Assignment 2 – Classification Models & Deployment

Name: Vivek Sharma
College ID: 2025aa05588
Mail ID: 2025aa05588@wilp.bits-pilani.ac.in

GitHub Repository Link

https://github.com/Vivek29112001/Heart_Disease_ML__

Streamlit Live App Link

https://heartdiseasedataset.streamlit.app/

1. Problem Statement

The objective of this project is to implement, evaluate, and deploy multiple machine learning classification models on a real-world dataset. The project demonstrates the complete end-to-end machine learning workflow including data preprocessing, model training, evaluation using standard performance metrics, building an interactive web application using Streamlit, and deploying the application on Streamlit Community Cloud.

---

2. Dataset Description

- **Dataset Name:** Heart Disease Dataset
- **Source:** Public dataset (UCI / Kaggle)
- **Type:** Binary Classification
- **Number of Instances:** 900+
- **Number of Features:** 13 input features

Target Variable

- **1** → Presence of heart disease
- **0** → Absence of heart disease

The dataset contains clinical attributes such as age, sex, cholesterol level, resting blood pressure, maximum heart rate, and other medical indicators. All preprocessing steps and train–test splitting were performed prior to model training.

---

3. Models Implemented

The following six machine learning classification models were implemented using the same dataset:

1. Logistic Regression
2. Decision Tree Classifier
3. K-Nearest Neighbors (KNN)
4. Naive Bayes (Gaussian)
5. Random Forest (Ensemble Model)
6. XGBoost (Ensemble Model)

---

4. Evaluation Metrics

Each model was evaluated using the following metrics:

- Accuracy
- AUC Score
- Precision
- Recall
- F1 Score
- Matthews Correlation Coefficient (MCC)

---

5. Model Performance Comparison Table

Model Name	Accuracy	AUC	Precision	Recall	F1 Score	MCC
Logistic Regression	0.85	0.88	0.84	0.83	0.83	0.59
Decision Tree	0.97	0.97	0.97	0.97	0.97	0.97
KNN	0.89	0.91	0.88	0.87	0.87	0.67
Naive Bayes	0.80	0.87	0.75	0.89	0.81	0.61
Random Forest	0.98	1.00	1.00	0.97	0.98	0.97
XGBoost	0.98	0.99	1.00	0.97	0.98	0.97

---

6. Model Performance Observations

Model Name	Observation
Logistic Regression	Performs reasonably well on linearly separable data but is outperformed by ensemble methods.

Decision Tree	Achieves high accuracy but may suffer from overfitting without proper constraints.
KNN	Provides good performance but is sensitive to the choice of K value and feature scaling.
Naive Bayes	Simple and fast probabilistic model that performs well despite strong independence assumptions.
Random Forest	Demonstrates excellent performance with high accuracy and MCC due to ensemble learning and reduced overfitting.
XGBoost	Delivers the best overall performance with strong generalization ability and high AUC score.

7. Streamlit Web Application

An interactive Streamlit web application was developed with the following features:

- CSV file upload option for test data
- Model selection dropdown
- Display of evaluation metrics
- Visualization of confusion matrix and ROC curve

The application loads pre-trained machine learning models and performs inference on uploaded test data without retraining the models.

8. Deployment

The Streamlit application was deployed using **Streamlit Community Cloud**. The deployment is directly linked to the GitHub repository and installs all dependencies using the `requirements.txt` file. The deployed application is publicly accessible through the provided live URL.

9. How to Run the Project Locally

```
pip install -r requirements.txt
python -m streamlit run app.py
```

---

10. Repository Structure

```
Heart_Disease_ML_/
|-- app.py
|-- requirements.txt
|-- README.md
|-- model/
|    ├── logisticregression.pkl
|    ├── decisiontree.pkl
|    ├── kneighbors.pkl
|    ├── gaussiannb.pkl
|    ├── randomforest.pkl
|    ├── xgboost.pkl
|    ├── scaler.pkl
```

---

11. Conclusion

This project successfully demonstrates a complete end-to-end machine learning pipeline including data preprocessing, model training, evaluation, deployment, and interactive visualization. Ensemble models such as Random Forest and XGBoost outperform traditional classifiers, highlighting the effectiveness of ensemble learning techniques for classification problems.