ML Assignment – 2

Evaluation of Classification Models

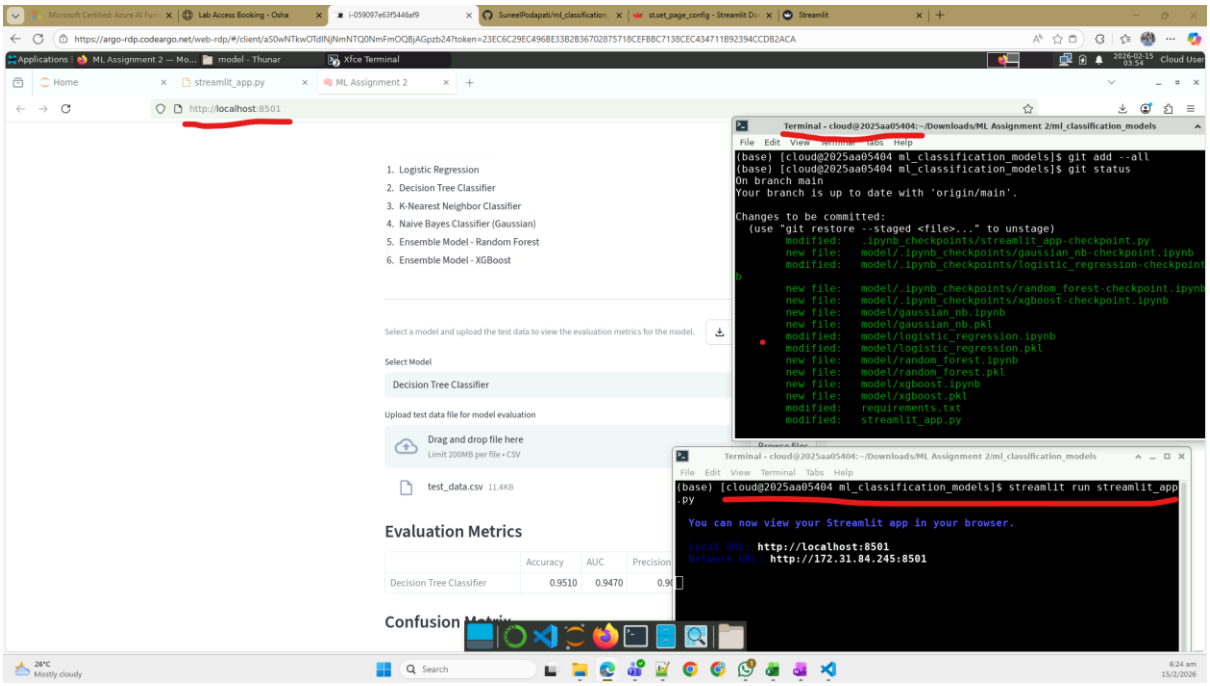
Name: **Podapati Suneel**

BITS Id: **2025AA05404**

1. GitHub Repo Link: https://github.com/SuneelPodapati/ml_classification_models

2. Live Streamlit App Link: https://sp-ml-assignment2.streamlit.app/

3. Screenshot of BITS Virtual Lab



4. GitHub README content

ML Assignment 2 – Heart Failure Prediction 🫀

Evaluate 6 classification models

a. Problem statement

The objective of this assignment is to train and evaluate multiple classification models. I have selected a data set to train and predict whether a patient with heart failure will survive or not based on their clinical records. The goal is to identify the high-risk patients, so required medical procedures can be suggested.

The target/outcome variable:

- **y = 0 => Not Dead → Survived**
- **y = 1 => Dead → Not Survived**

b. Dataset description

Heart Failure Prediction data from Kaggle.

The Heart Failure Clinical Records data contains clinical and demographic information of patients diagnosed with heart failure. It contains details about existing medical conditions and clinical measurements such as ejection fraction, serum creatinine, and serum sodium. These attributes are highly useful for predicting a patient's risk of mortality during the follow-up period.

Features: ***12*** - age creatinine_phosphokinase ejection_fraction platelets serum_creatinin e serum_sodium time anaemia diabetes high_blood_pressure sex smoking

Data set size: ***5000*** (before cleaning)

Instances/sample size: ***1319*** (after removing the contradictory and duplicate samples)

Train/Test split: 80:20 → *1055:264*

Attribute details

- age: age of the patient (years)
- anaemia: decrease of red blood cells or haemoglobin (boolean)
- creatinine phosphokinase (CPK): level of the CPK enzyme in the blood (mcg/L)
- diabetes: if the patient has diabetes (boolean)

- ejection fraction: percentage of blood leaving the heart at each contraction (percentage)
- high blood pressure: if the patient has hypertension (boolean)
- platelets: platelets in the blood (kiloplatelets/mL)
- sex: woman or man (binary)
- serum creatinine: level of serum creatinine in the blood (mg/dL)
- serum sodium: level of serum sodium in the blood (mEq/L)
- smoking: if the patient smokes or not (boolean)
- time: follow-up period (days)
- DEATH_EVENT: if the patient died during the follow-up period (boolean)

Feature Engineering

- Instances with same feature values but different (contradicting) target values has been cleaned to have single instance for each combination of features.
- All categorical features are One-hot encoded
- All numerical features are z-score normalized (standardization)

c. Models used

The following machine learning classification models were implemented and evaluated using the same dataset. The evaluation metrics used for comparison are Accuracy, AUC Score, Precision, Recall, F1 Score, and Matthews Correlation Coefficient (MCC).

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

Comparison Table

ML Model Name	Accuracy	AUC	Precision	Recall	F1	MCC
Logistic Regression	0.856	0.886	0.797	0.696	0.743	0.647
Decision Tree Classifier	0.951	0.947	0.902	0.937	0.919	0.884
K-Nearest Neighbor Classifier	0.852	0.936	0.917	0.557	0.693	0.636
Naive Bayes Classifier (Gaussian)	0.799	0.850	0.703	0.570	0.629	0.499
Random Forest (Ensemble)	0.962	0.985	0.960	0.911	0.935	0.909
XGBoost (Ensemble)	0.936	0.984	0.943	0.835	0.886	0.844

Observations on Model Performance

ML Model Name	Observation about model performance
Logistic Regression	A solid baseline model with balanced performance. Accuracy (0.856) and AUC (0.886) indicate good overall discrimination. Precision (0.797) is stronger than recall (0.696), meaning it predicts positive cases conservatively and may miss some true death events. Could not improve the performance even after increasing the iterations by 10x
Decision Tree Classifier	Very strong performance with high accuracy (0.951) and excellent recall (0.937). It captures most positive cases and maintains high precision (0.902). However, single trees can overfit, so this performance may not generalize as consistently as ensemble methods.
K-Nearest Neighbor Classifier	Shows mixed behaviour: high precision (0.917) but low recall (0.557). This means it is very good at avoiding false positives but misses many true positives. Accuracy (0.852) and AUC (0.936) are strong, but the imbalance between precision and recall suggests sensitivity issues. Increasing the k value after 6 was also increasing the error
Naive Bayes (Gaussian)	The weakest performer in your set. Accuracy (0.799), recall (0.570), and F1 (0.629) are noticeably lower. The independence assumption likely limits its ability to model the clinical relationships in this dataset.
Random Forest	The best overall performer. Outstanding accuracy (0.962), AUC (0.985), and F1 (0.935). Precision (0.960) and recall (0.911) are both high, showing excellent balance. MCC (0.909) confirms strong reliability across both classes.
XGBoost	Another top-tier model with excellent accuracy (0.936) and AUC (0.984). Precision (0.943) and recall (0.835) are well-balanced, though slightly lower than Random Forest. Still a highly dependable model with strong generalization.

Conclusion

Ensemble Model - Random Forest achieved the highest AUC and strongest overall metrics on this dataset. This model has a good trade-off between accuracy, F1, and MCC, making it the preferred choice for production use.

Streamlit 🚀

The trained models are saved into *pkl* files and are deployed to Streamlit with an interactive evaluation application at https://sp-ml-assignment2.streamlit.app/

The application can be used to

- download sample test data
- select a model for evaluation
- upload test data
- view the model evaluation metrics: Accuracy, AUC, Precision, Recall, F1, and MCC.
- view confusion matrix and classification report.