Machine Learning Assignment 2

Classification Models

Student Name:	Aravindan B
Student ID:	2025aa05026
Program:	M.Tech (AIML), BITS Pilani WILP
Course:	Machine Learning
Submission Date:	February 2026
Email:	2025aa05026@wilp.bits-pilani.ac.in

1. GitHub Repository Link

Repository URL:

```
https://github.com/algoyog/mlassignment2
```

Repository Contents:

- **app.py** -- Streamlit web application
- **requirements.txt** -- Python dependencies
- **README.md** -- Complete documentation
- **.gitignore** -- Git ignore rules
- **model/model_training.ipynb** -- Jupyter notebook for model training
- **input/wine_quality_red.csv** -- Wine Quality Red dataset
- **output/** -- Generated results and charts
- **utils/ml_utils.py** -- Shared ML utilities

2. Live Streamlit App Link

Application URL:

```
https://mlassignment2-kmepmbozytsyeyf3yerkyb.streamlit.app/
```

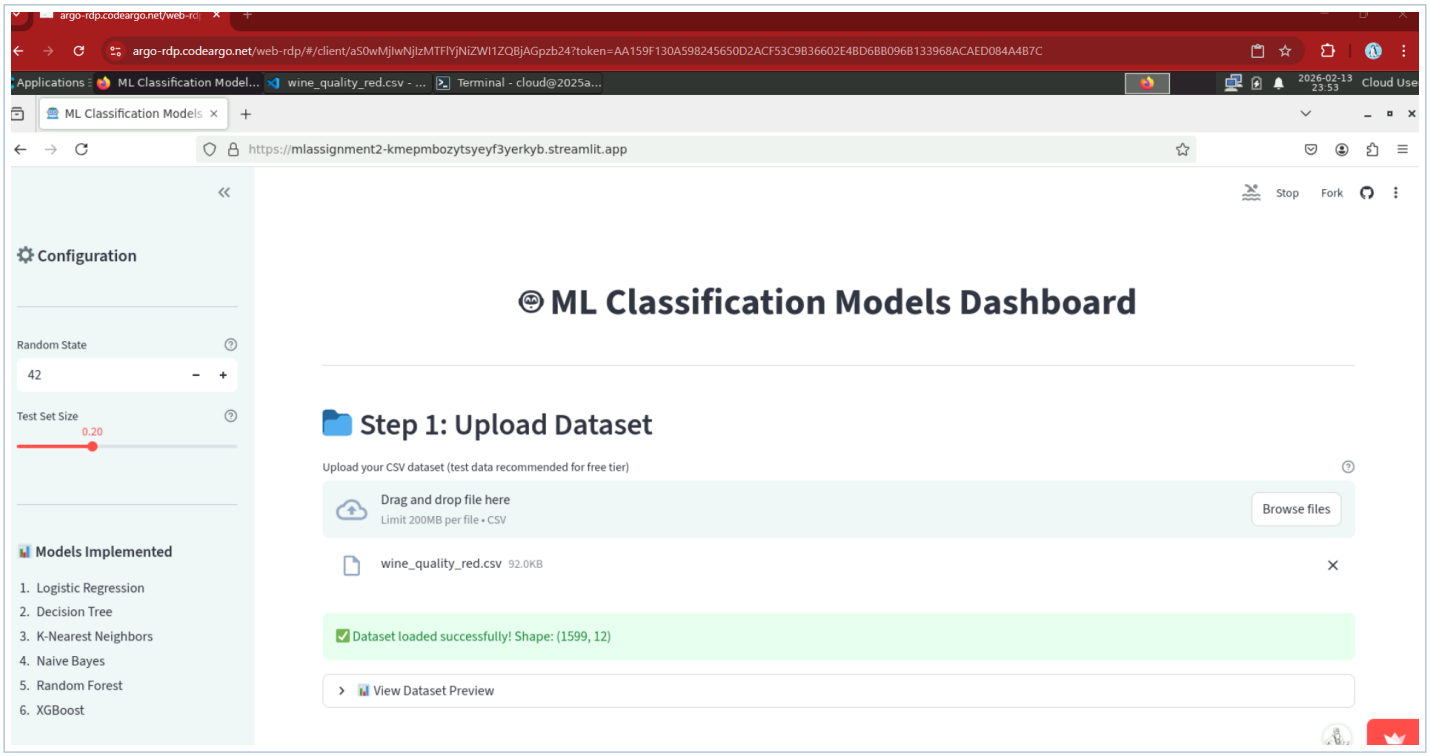
Implemented Features:

- CSV file upload with dataset preview
- Target column selection with class distribution
- Model selection dropdown (6 classification models)
- All 6 evaluation metrics displayed (Accuracy, AUC, Precision, Recall, F1, MCC)
- Confusion matrix visualization
- Classification report
- Model comparison table with best model highlighted
- Results download as CSV

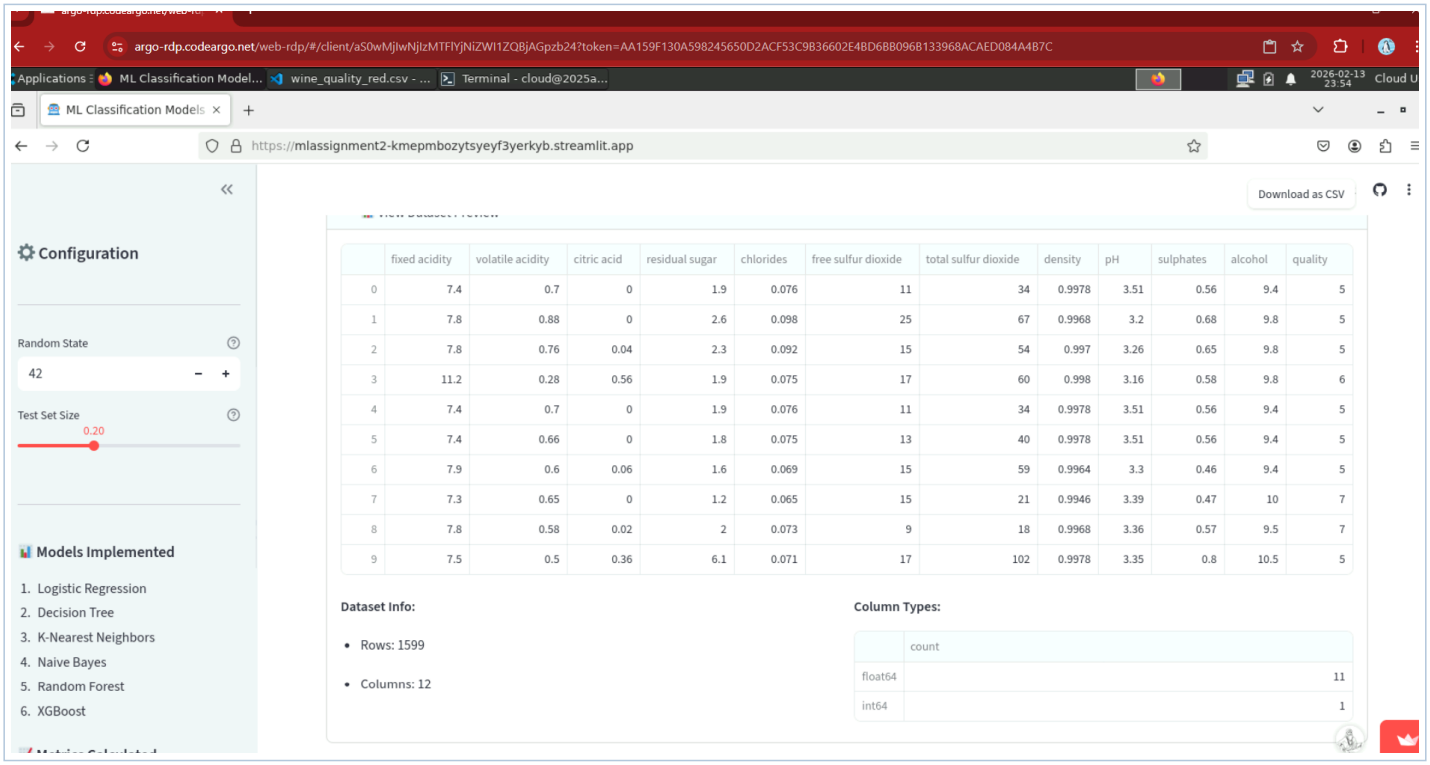
3. BITS Virtual Lab Screenshots

The following screenshots demonstrate the Streamlit application running on the BITS Virtual Lab (argo-rdp.codeargo.net) using the Wine Quality Red dataset (1599 instances, 12 features) as a test dataset.

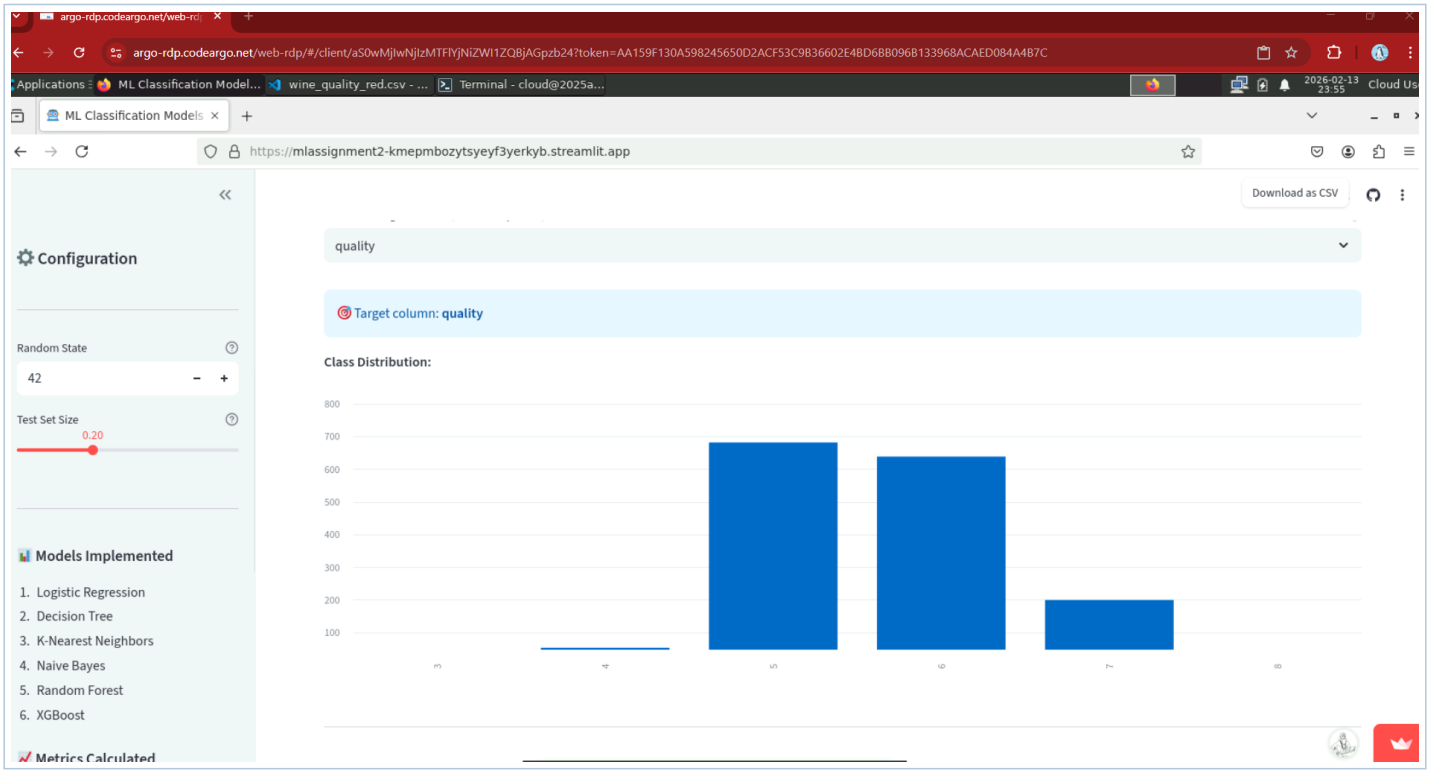
Screenshot 1: CSV Upload - wine_quality_red.csv loaded successfully (1599 rows, 12 features)



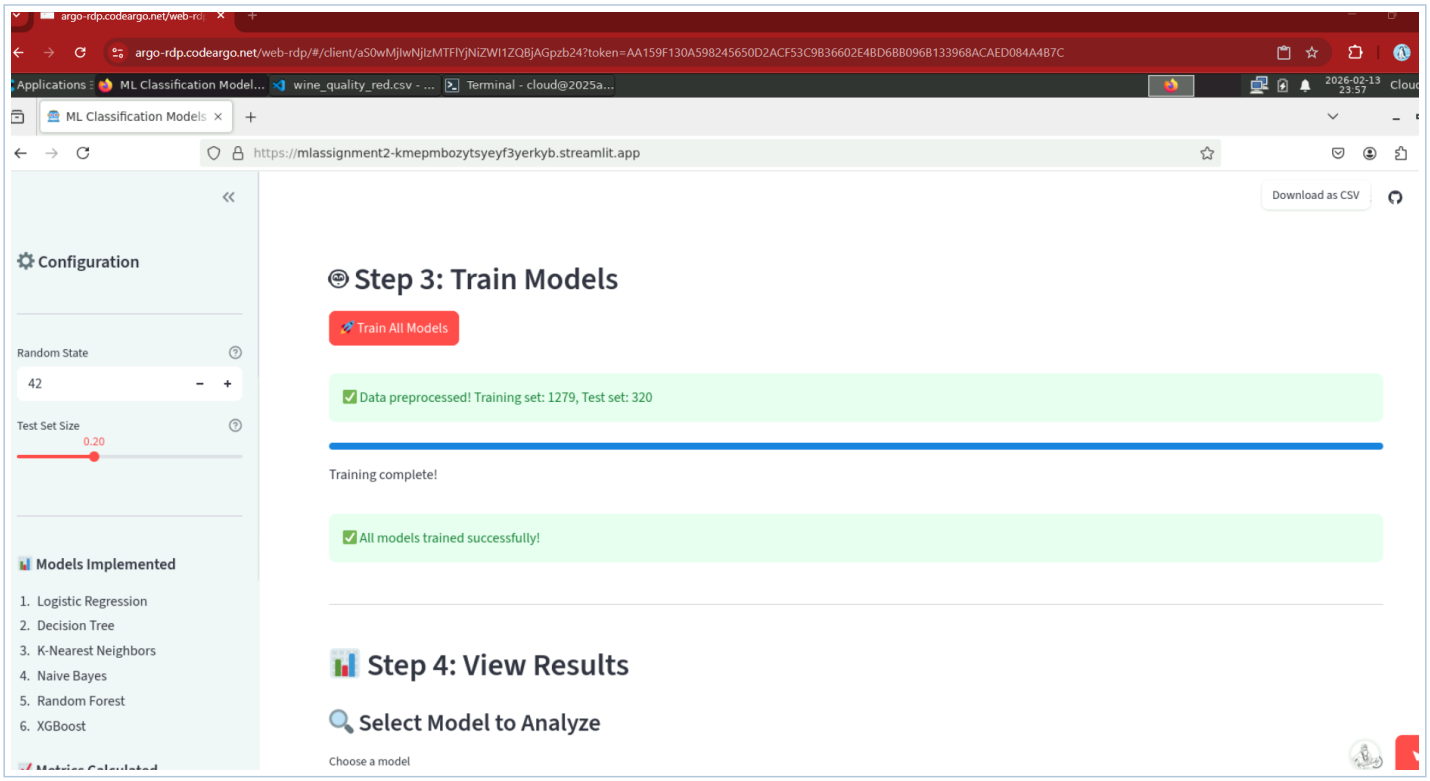
Screenshot 2: Dataset Preview - Feature table and dataset information



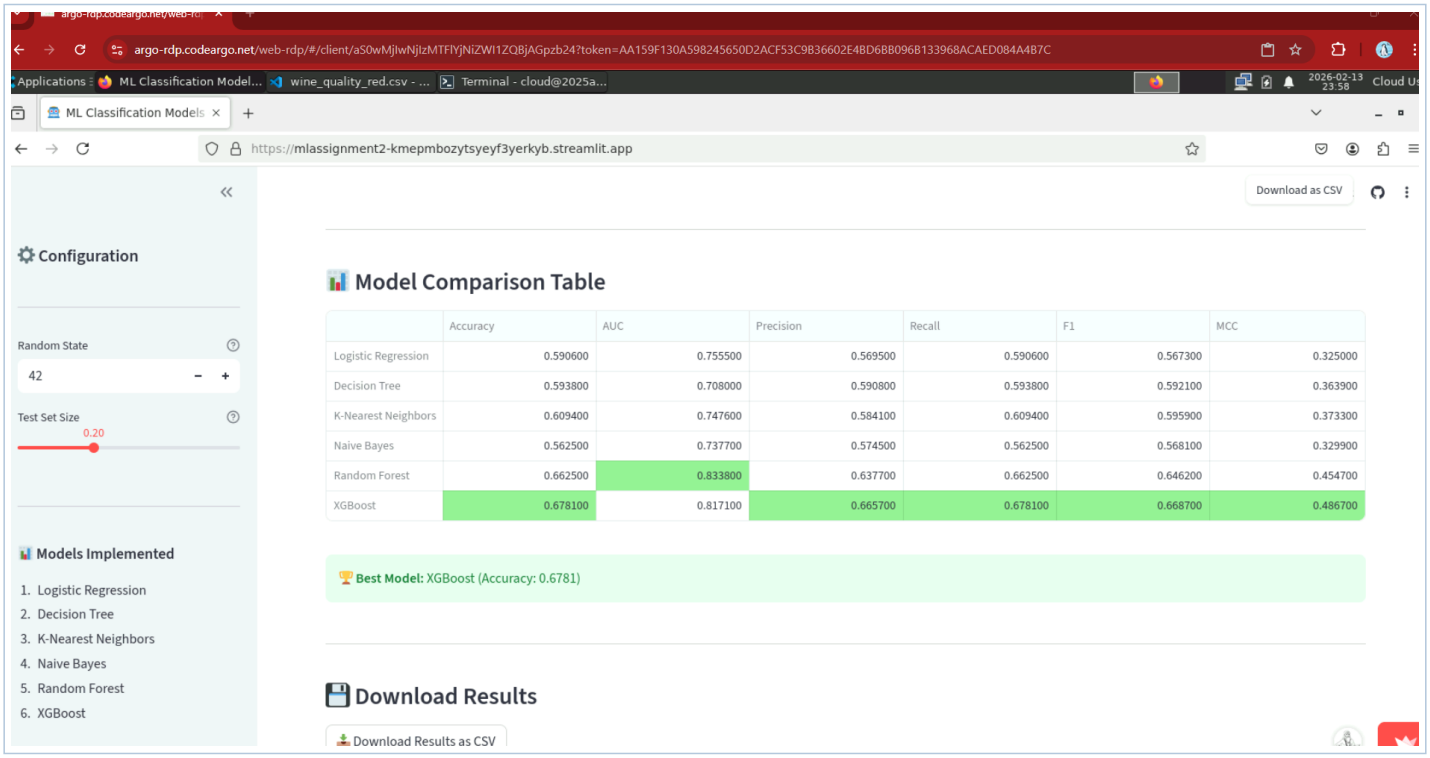
Screenshot 3: Target Column Selection - Class distribution visualization for "quality"



Screenshot 4: Model Training - All 6 models trained successfully



Screenshot 5: Model Comparison Table - All 6 models with all 6 evaluation metrics



	Accuracy	AUC	Precision	Recall	F1	MCC
Logistic Regression	0.590600	0.755500	0.569500	0.590600	0.567300	0.325000
Decision Tree	0.593800	0.708000	0.590800	0.593800	0.592100	0.363900
K-Nearest Neighbors	0.609400	0.747600	0.584100	0.609400	0.595900	0.373300
Naive Bayes	0.562500	0.737700	0.574500	0.562500	0.568100	0.329900
Random Forest	0.662500	0.833800	0.637700	0.662500	0.646200	0.454700
XGBoost	0.678100	0.817100	0.665700	0.678100	0.668700	0.486700

4. README Content

Problem Statement

This project implements a comprehensive machine learning classification pipeline featuring six different classification algorithms. The goal is to compare the performance of traditional ML models (Logistic Regression, Decision Tree, K-Nearest Neighbors, Naive Bayes) against ensemble methods (Random Forest and XGBoost) using six evaluation metrics. The pipeline is demonstrated using the Wine Quality Red dataset and includes an interactive Streamlit web application that supports any CSV classification dataset.

Dataset Description

Dataset Name: Wine Quality Red Dataset

Source: UCI Machine Learning Repository

Type: Multi-class Classification

Instances: 1,599 (Requirement >= 500: MET)

Features: 12 (Requirement >= 12: MET)

Target: quality (wine quality score)

Features (all numerical): fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, quality

Data Preprocessing Steps

- StandardScaler normalization on all features
- 80-20 stratified train-test split

Model Performance Comparison Table

ML Model Name	Accuracy	AUC	Precision	Recall	F1	MCC
Logistic Regression	0.5906	0.7555	0.5695	0.5906	0.5673	0.3250
Decision Tree	0.5938	0.7080	0.5908	0.5938	0.5921	0.3639
K-Nearest Neighbors	0.6094	0.7476	0.5841	0.6094	0.5959	0.3733
Naive Bayes	0.5625	0.7377	0.5745	0.5625	0.5681	0.3299
Random Forest	0.6625	0.8338	0.6377	0.6625	0.6462	0.4547
XGBoost	**0.6781**	**0.8171**	**0.6657**	**0.6781**	**0.6687**	**0.4867**

Best Model: XGBoost (Accuracy = 0.6781, AUC = 0.8171)

Model Performance Observations

ML Model Name	Observation about Model Performance
Logistic Regression	Achieves 59.06% accuracy with AUC 0.7555, serving as a solid linear baseline. It performs reasonably because several wine features like alcohol and volatile acidity have approximately linear relationships with quality. The model converges reliably with lbfgs solver and is the most interpretable among the six, though it falls short on non-linear patterns compared to tree-based models.
Decision Tree	Achieves 59.38% accuracy with AUC 0.7080 and MCC 0.3639. With max_depth=10, it captures non-linear interactions between features like alcohol, volatile acidity, and sulphates. It slightly outperforms Logistic Regression while remaining interpretable through decision rules. However, it is surpassed by ensemble methods which reduce its variance.
K-Nearest Neighbors	KNN with k=5 achieves 60.94% accuracy with AUC 0.7476, outperforming both linear models. Feature scaling via StandardScaler is essential for KNN distance calculations and was correctly applied. On 1,599 instances, it is computationally efficient at prediction time and benefits from the natural clustering present in the Wine Quality dataset.
Naive Bayes	Gaussian Naive Bayes achieves the lowest accuracy at 56.25% with MCC 0.3299. The feature independence assumption does not hold well here as features like fixed acidity, citric acid, and pH are correlated, limiting its performance. Despite this, it achieves AUC of 0.7377, indicating reasonable probability calibration, and is the fastest model to train.
Random Forest	Achieves 66.25% accuracy with AUC 0.8338 and MCC 0.4547, ranking second overall. The ensemble of 100 trees reduces overfitting through bagging and random feature subsets. Its AUC of 0.8338 reflects the best discriminative ability among all models. It consistently outperforms all traditional models across every metric and provides feature importance insights for interpretability.
XGBoost	Best performing model with 67.81% accuracy, highest F1 (0.6687), and highest MCC (0.4867). Its gradient boosting framework iteratively corrects errors from previous trees, capturing complex non-linear patterns. With learning_rate=0.1, max_depth=6, and 100 estimators, it demonstrates the clear advantage of advanced ensemble techniques over traditional classifiers on this multi-class dataset.

Overall Insights

- Best performing model: XGBoost with 67.81% accuracy, F1 0.6687, MCC 0.4867
- Ensemble methods (Random Forest and XGBoost) significantly outperform all traditional algorithms across every metric
- Naive Bayes is weakest (56.25%) due to its feature independence assumption not holding for correlated wine features
- All models achieved AUC > 0.70, indicating reasonable discriminative ability for this challenging multi-class problem

Aravindan B | M.Tech (AIML), BITS Pilani WILP | 2025aa05026@wilp.bits-pilani.ac.in

END OF SUBMISSION