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/adult_income.csv** -- UCI Adult Income 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/
```

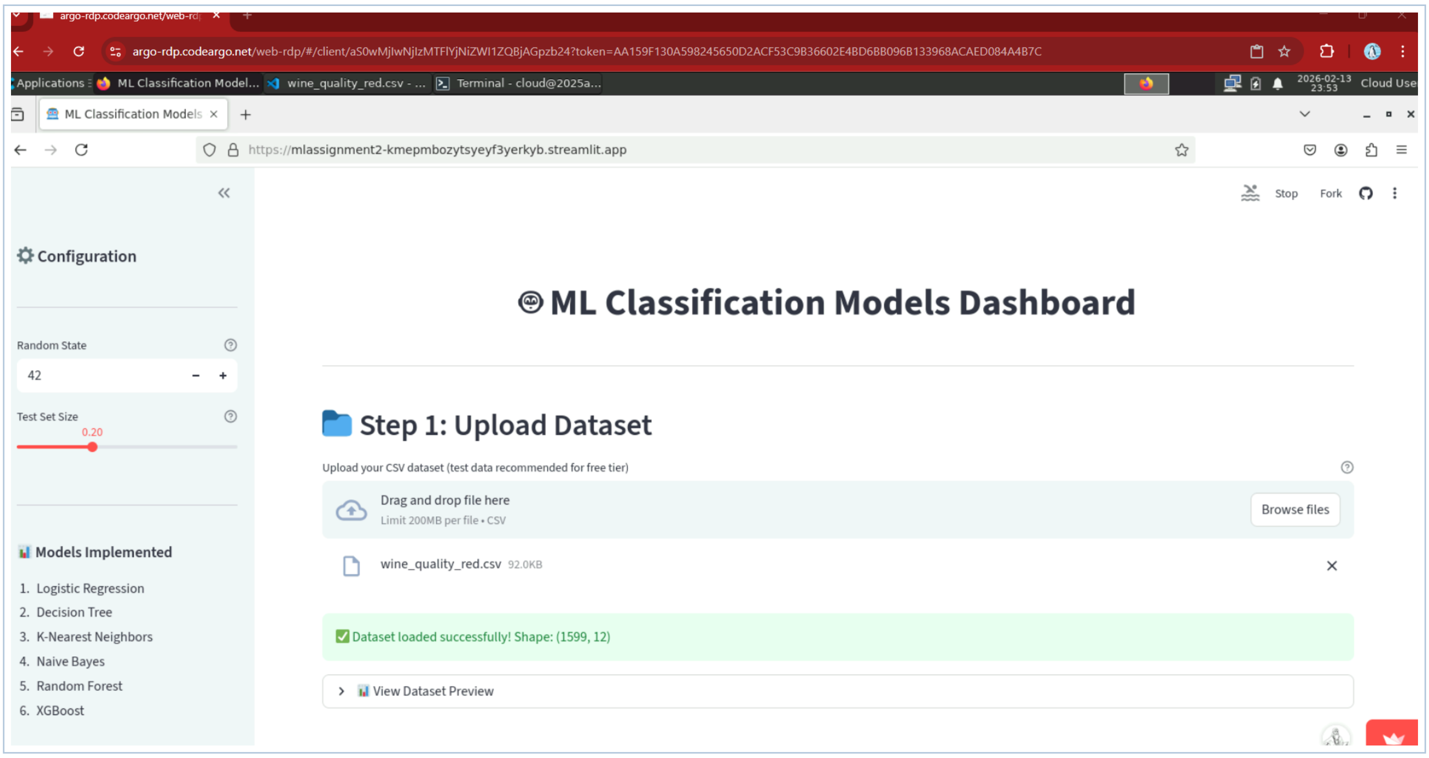
Implemented Features:

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



argo-rdp.codeargo.net/web-rdp/#/client/aS0wMjIwNjIzMTFlYjNiZWI1ZQBjAGpzb24?token=AA159F130A598245650D2ACF53C9B36602E4BD6BB096B133968ACAED084A4B7C

https://mlassignment2-kmepmbozytsyeyf3yerkyb.streamlit.app

Applications
ML Classification Model...
wine_quality_red.csv - ...
Terminal - cloud@2025a...
2026-02-13 23:53
Cloud Use
ML Classification Models
Stop
Fork

⚙ Configuration

Random State: 42

Test Set Size: 0.20

Models Implemented

1. Logistic Regression
2. Decision Tree
3. K-Nearest Neighbors
4. Naive Bayes
5. Random Forest
6. XGBoost

ML Classification Models Dashboard

Step 1: Upload Dataset

Upload your CSV dataset (test data recommended for free tier)

Drag and drop file here
Limit 200MB per file • CSV

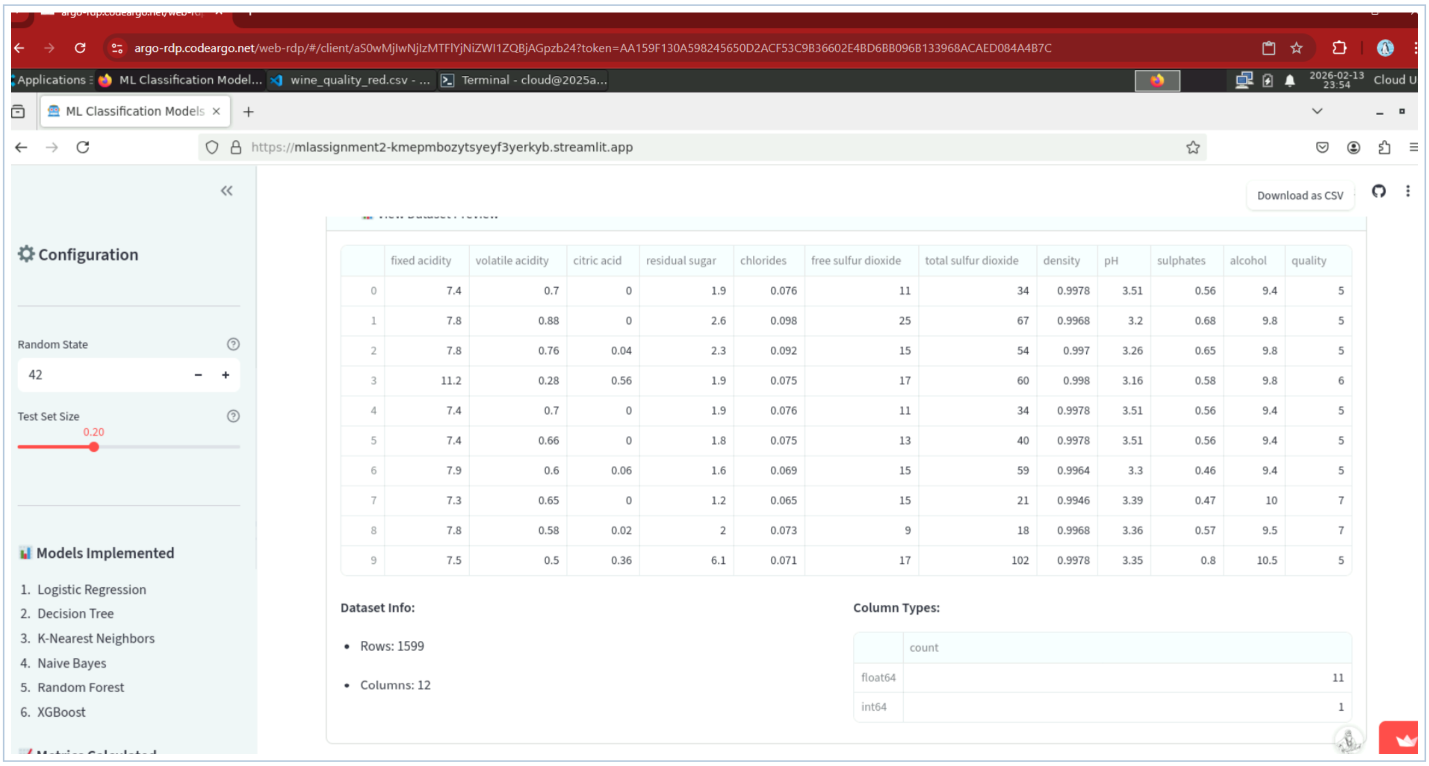
Browse files

wine_quality_red.csv 92.0KB

✅ Dataset loaded successfully! Shape: (1599, 12)

> View Dataset Preview

Screenshot 2: Step 2: Dataset Preview - Feature table and dataset information displayed



argo-rdp.codeargo.net/web-rdp/#/client/aS0wMjIwNjIzMTFlYjNiZWI1ZQBjAGpzb24?token=AA159F130A598245650D2ACF53C9B36602E4BD6BB096B133968ACAED084A4B7C

https://mlassignment2-kmepmbozytsyeyf3yerkyb.streamlit.app

Applications
ML Classification Model...
wine_quality_red.csv - ...
Terminal - cloud@2025a...
2026-02-13 23:54
Cloud U
ML Classification Models
Download as CSV

⚙ Configuration

Random State: 42

Test Set Size: 0.20

Models Implemented

1. Logistic Regression
2. Decision Tree
3. K-Nearest Neighbors
4. Naive Bayes
5. Random Forest
6. XGBoost

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
4	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
5	7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5
6	7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5
7	7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7
8	7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7
9	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5

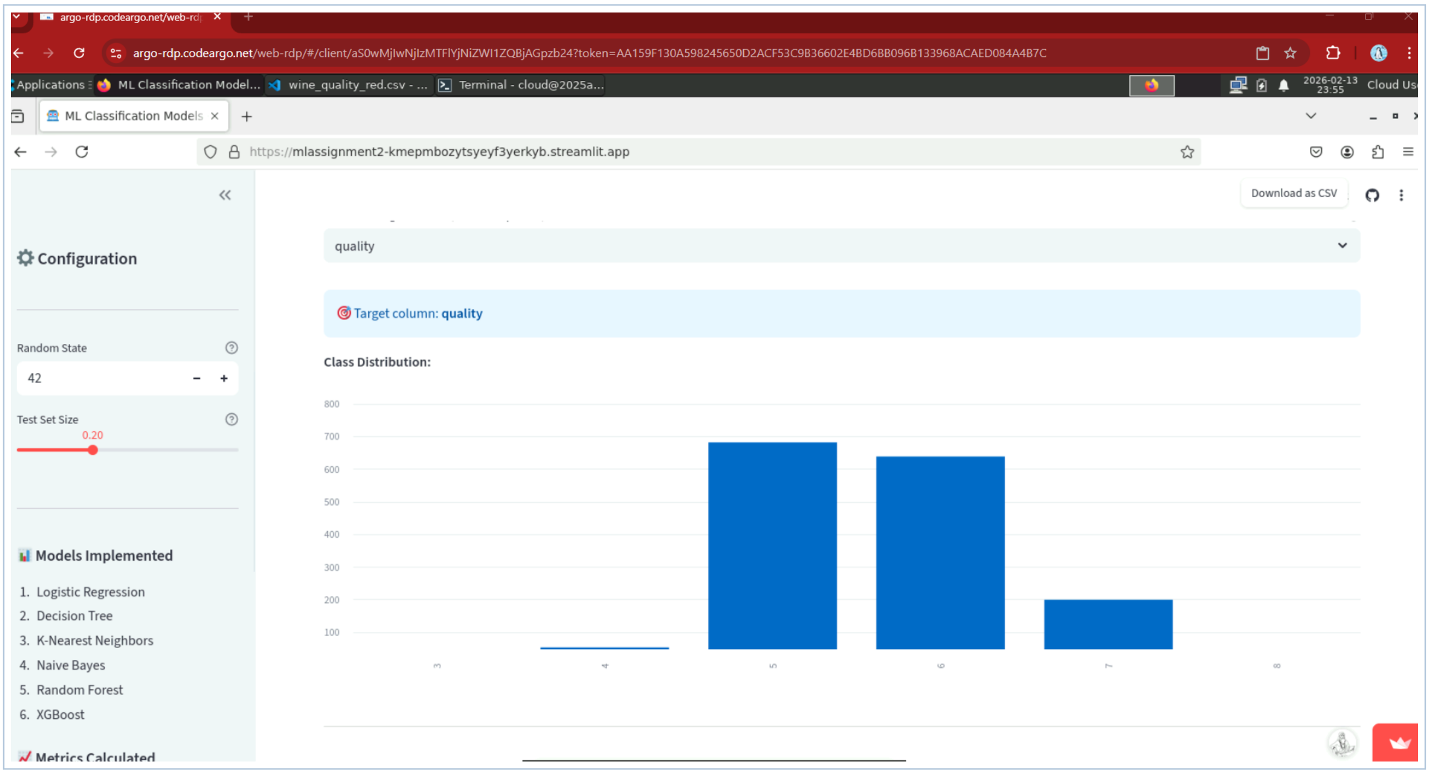
Dataset Info:

- Rows: 1599
- Columns: 12

Column Types:

	count
float64	11
int64	1

Screenshot 3: Step 2: Target Column Selection - Class distribution visualization for "quality"


argo-rdp.codeargo.net/web-rdp
argo-rdp.codeargo.net/web-rdp/#/client/aS0wMjlwNjlzMTFlYjNiZWI1ZQBjAGpzb24?token=AA159F130A598245650D2ACF53C9B36602E4BD6BB096B133968ACAED084A4B7C
Applications
ML Classification Model...
wine_quality_red.csv - ...
Terminal - cloud@2025a...
2026-02-13 23:55
Cloud Us
ML Classification Models
https://mlassignment2-kmepmbozytsyeyf3yerkyb.streamlit.app
Download as CSV

⚙ Configuration

Random State: 42

Test Set Size: 0.20

quality

🎯 Target column: **quality**

Class Distribution:

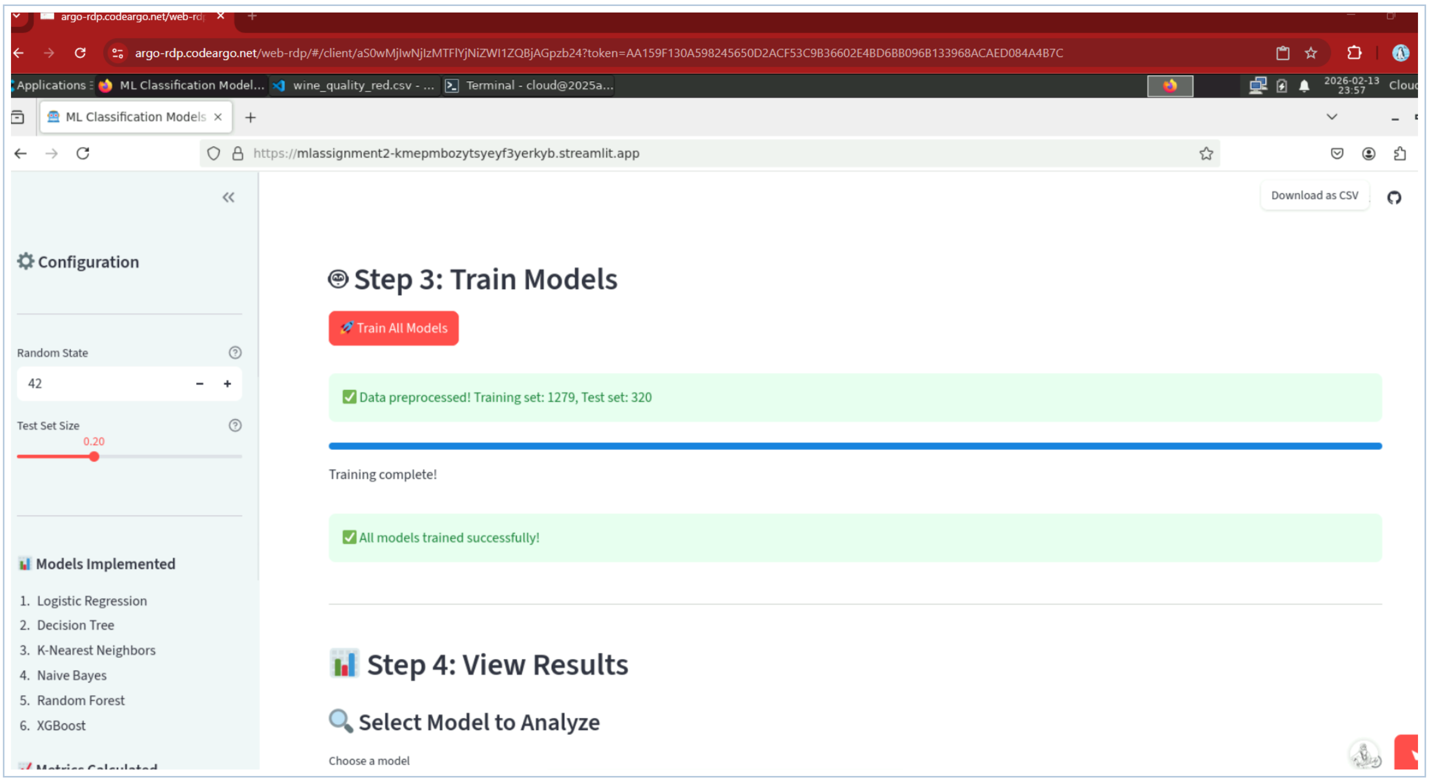
800
700
600
500
400
300
200
100
3
4
5
6
7
8

📊 Models Implemented

1. Logistic Regression
2. Decision Tree
3. K-Nearest Neighbors
4. Naive Bayes
5. Random Forest
6. XGBoost

📈 Metrics Calculated

Screenshot 4: Step 3: Model Training - All 6 models trained successfully on BITS Virtual Lab


argo-rdp.codeargo.net/web-rdp
argo-rdp.codeargo.net/web-rdp/#/client/aS0wMjlwNjlzMTFlYjNiZWI1ZQBjAGpzb24?token=AA159F130A598245650D2ACF53C9B36602E4BD6BB096B133968ACAED084A4B7C
Applications
ML Classification Model...
wine_quality_red.csv - ...
Terminal - cloud@2025a...
2026-02-13 23:57
Cloud
ML Classification Models
https://mlassignment2-kmepmbozytsyeyf3yerkyb.streamlit.app
Download as CSV

⚙ Configuration

Random State: 42

Test Set Size: 0.20

📊 Models Implemented

1. Logistic Regression
2. Decision Tree
3. K-Nearest Neighbors
4. Naive Bayes
5. Random Forest
6. XGBoost

📈 Metrics Calculated

Step 3: Train Models

🚀 Train All Models

✅ Data preprocessed! Training set: 1279, Test set: 320

Training complete!

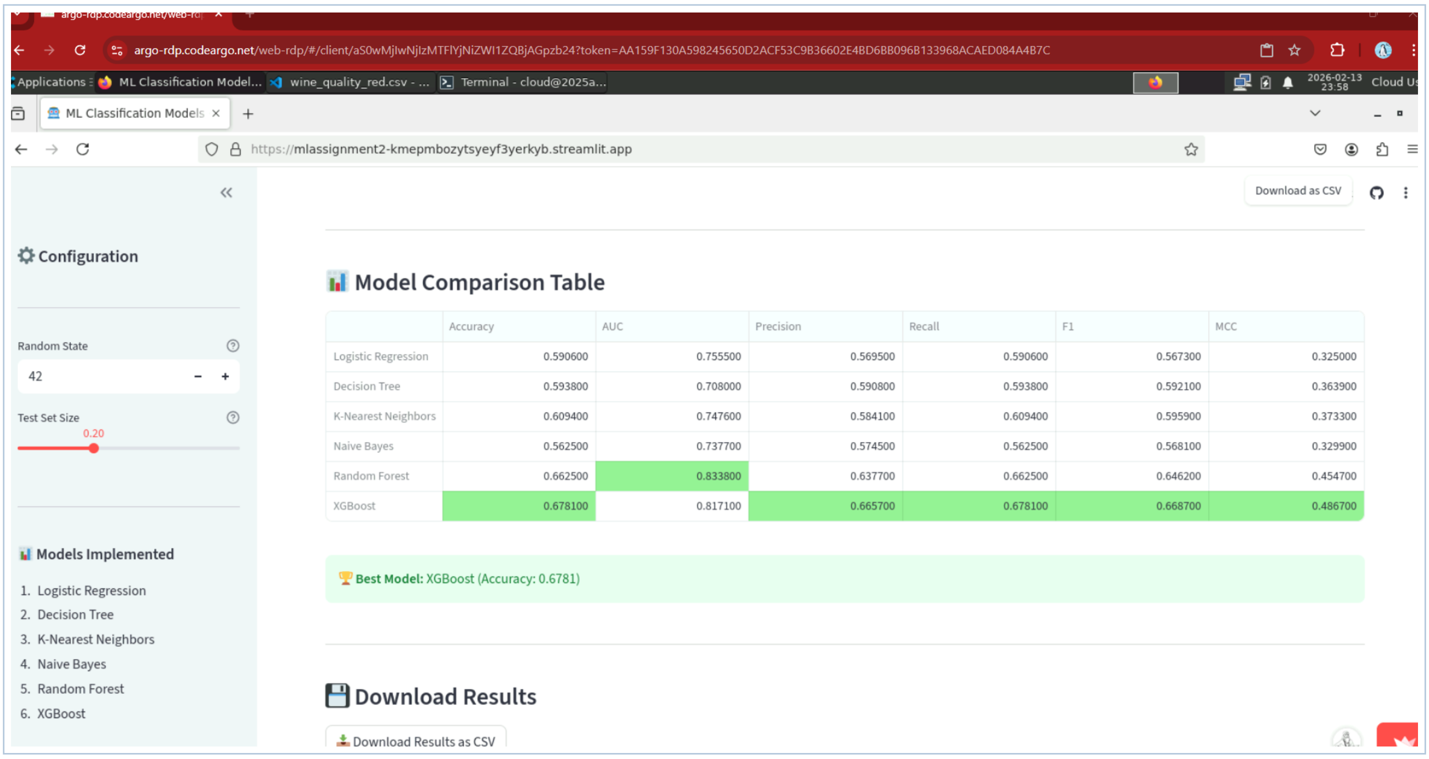
✅ All models trained successfully!

📊 Step 4: View Results

🔍 Select Model to Analyze

Choose a model

Screenshot 5: Step 4: Model Comparison Table - All 6 models with all 6 evaluation metrics


argo-rdp.codeargo.net/#/client/aS0wMjlwNjIzMTFlYjNiZWI1ZQBjAGpzb24?token=AA159F130A598245650D2ACF53C9B36602E4BD6BB096B133968ACAED084A4B7C
Applications
ML Classification Model...
wine_quality_red.csv - ...
Terminal - cloud@2025a...
2026-02-13 23:58
ML Classification Models
https://mlassignment2-kmepmbozytsyeyf3yerkyb.streamlit.app
Download as CSV

Configuration

Random State: 42

Test Set Size: 0.20

Models Implemented

1. Logistic Regression
2. Decision Tree
3. K-Nearest Neighbors
4. Naive Bayes
5. Random Forest
6. XGBoost

Model Comparison Table

	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

🏆 **Best Model:** XGBoost (Accuracy: 0.6781)

Download Results

Download Results as CSV

4. README Content

Problem Statement

This project implements a comprehensive machine learning classification pipeline featuring six different classification algorithms applied to the UCI Adult Income dataset. The goal is to predict whether an individual's annual income exceeds $50K based on census attributes, and 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.

Dataset Description

Dataset Name: UCI Adult Income Dataset (Census Income)

Source: UCI Machine Learning Repository

Type: Binary Classification

Instances: 30,162 (Requirement >= 500: MET)

Features: 14 (Requirement >= 12: MET)

Target: income (<=50K or >50K)

Class Distribution: Class 0 (<=50K): 22,654 samples (75.1%) | Class 1 (>50K): 7,508 samples (24.9%)

Numerical Features (6): age, fnlwgt, education_num, capital_gain, capital_loss, hours_per_week

Categorical Features (8): workclass, education, marital_status, occupation, relationship, race, sex, native_country

Data Preprocessing Steps

- Label Encoding on all 8 categorical features
- Target encoded: <=50K to 0, >50K to 1
- 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	0.8175	0.8501	0.8060	0.8175	0.8018	0.4613
Decision Tree	0.8508	0.8855	0.8446	0.8508	0.8451	0.5789
K-Nearest Neighbors	0.8190	0.8498	0.8133	0.8190	0.8154	0.4993
Naive Bayes	0.7978	0.8498	0.7830	0.7978	0.7697	0.3798
Random Forest	0.8589	0.9136	0.8534	0.8589	0.8526	0.6003
XGBoost	**0.8671**	**0.9243**	**0.8624**	**0.8671**	**0.8624**	**0.6269**

Best Model: XGBoost (Accuracy = 0.8671, AUC = 0.9243)

Model Performance Observations

ML Model Name	Observation about Model Performance
Logistic Regression	Achieves 81.75% accuracy with AUC 0.8501, serving as a solid linear baseline. It performs well because features like education_num, age, and hours_per_week have approximately linear relationships with income. 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 85.08% accuracy with AUC 0.8855 and MCC 0.5789. With max_depth=10, it successfully captures non-linear interactions between occupation, marital_status, and education without extreme overfitting. It outperforms linear models significantly 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 81.90% accuracy with AUC 0.8498, slightly edging out Logistic Regression. Feature scaling via StandardScaler is essential for KNN distance calculations and was correctly applied. On 30,162 instances, it is computationally heavier at prediction time but benefits from the demographic clustering present in the Adult Income dataset.
Naive Bayes	Gaussian Naive Bayes achieves the lowest accuracy at 79.78% with MCC 0.3798. The feature independence assumption does not hold well here as features like education, occupation, and marital_status are correlated, limiting its performance. Despite this, it achieves a competitive AUC of 0.8498, indicating reasonable probability calibration, and is the fastest model to train.
Random Forest	Achieves 85.89% accuracy with AUC 0.9136 and MCC 0.6003, ranking second overall. The ensemble of 100 trees reduces overfitting through bagging and random feature subsets. Its AUC of 0.9136 reflects excellent discriminative ability. It consistently outperforms all traditional models across every metric and provides feature importance insights for interpretability.
XGBoost	Best performing model with 86.71% accuracy, highest AUC (0.9243), F1 (0.8624), and MCC (0.6269). 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 tabular dataset.

Overall Insights

- Best performing model: XGBoost with 86.71% accuracy, AUC 0.9243, F1 0.8624, MCC 0.6269
- Ensemble methods (Random Forest and XGBoost) outperform all traditional algorithms across every metric
- Naive Bayes is weakest (79.78%) due to its feature independence assumption not holding for this correlated dataset
- All models achieved AUC > 0.84, indicating good discriminative ability despite class imbalance (75%/25%)

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

END OF SUBMISSION