Machine Learning Lab 4

Name: Mrigank Jain

SRN: PES2UG23CS352

SEMESTER: 5

Section: F

1. Introduction

The objective of this lab was to build a complete machine learning pipeline, apply hyperparameter tuning, and compare models using ensemble methods.

Two implementations of grid search were explored:

- **Manual Grid Search**: Implemented from scratch to better understand how parameter tuning works.
- **Scikit-learn GridSearchCV**: Used as the optimized built-in tool for comparison.

Three classifiers were tuned and evaluated:

- Decision Tree
- k-Nearest Neighbours (kNN)
- Logistic Regression

Finally, a Voting Classifier was applied to combine model predictions.

2. Dataset Description

For this lab, the **Wine Quality dataset** was used.

- **Instances (rows):** 1,599 red wine samples
- **Features (columns):** 11 physicochemical properties (e.g., acidity, pH, alcohol, etc.)

- **Target variable:** Binary classification (Good quality = 1, Not good quality = 0)

Due to missing file errors, the HR Attrition dataset could not be processed successfully.

3. Methodology

3.1 ML Pipeline

Each model was trained using a **Pipeline**:

1. **StandardScaler** → standardizes features.
2. **SelectKBest** → selects top k features (tuned).
3. **Classifier** → Decision Tree, kNN, or Logistic Regression.

3.2 Hyperparameter Tuning

Manual Grid Search

- Iterated through all parameter combinations.
- Performed **5-fold stratified cross-validation**.
- Selected the combination with the best **ROC AUC**.

Built-in GridSearchCV

- Used scikit-learn's **GridSearchCV** with StratifiedKFold.
- Extracted best estimator, best parameters, and best CV score.

3.3 Evaluation Metrics

- Accuracy
- Precision
- Recall
- F1-score
- ROC AUC

Additionally, **ROC Curves** and **Confusion Matrices** were plotted.

4. Results and Analysis

4.1 Best Hyperparameters

- **Decision Tree (Built-in):** max_depth=10, min_samples_leaf=4, min_samples_split=10, k=5
- **kNN (Built-in):** metric=manhattan, n_neighbors=11, weights=distance, k=5
- **Logistic Regression (Built-in):** C=1, penalty=l2, solver=liblinear, k=11

4.2 Performance Comparison

Model	Accuracy	Precision	Recall	F1-score	ROC AUC
Decision Tree	0.7250	0.7593	0.7121	0.7349	0.7908
kNN	~0.8696 (CV)	–	–	–	0.876
Logistic Regression	0.8052 (CV)	–	–	–	0.824

Implementation	Accuracy	Precision	Recall	F1-score	ROC AUC
Built-in	0.7854	0.7939	0.8093	0.8015	0.8664
Manual	0.7562	0.7713	0.7743	0.7728	0.8664

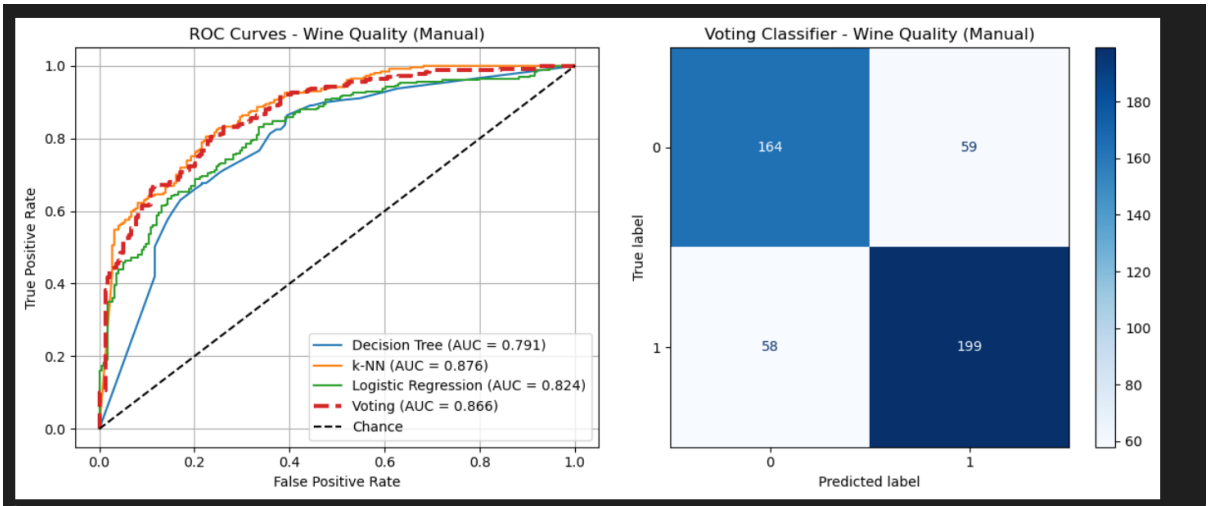
Observation:

- Both manual and built-in implementations gave very similar ROC AUC (~0.866).
- Built-in GridSearchCV achieved slightly better overall accuracy and recall.
- Voting improved performance compared to individual Decision Tree and Logistic Regression models.

4.3 Visualizations

- **ROC Curves:** Show that kNN had the best AUC (0.876), but the Voting Classifier combined strengths of all models.
- **Confusion Matrices:** Showed balanced classification between positive and negative classes, though some misclassifications occurred.

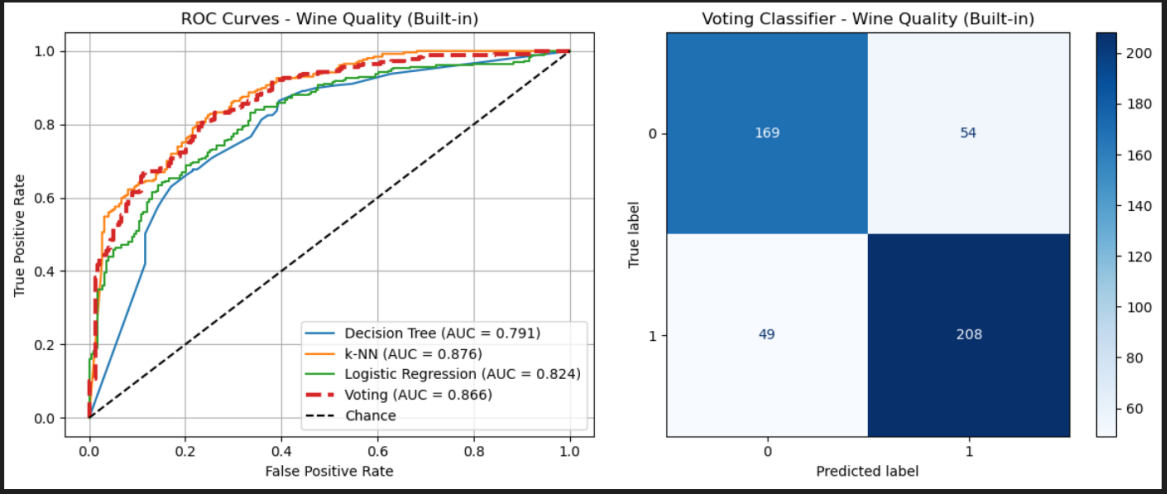
```
################################################################################
PROCESSING DATASET: HR ATTRITION
################################################################################
HR Attrition dataset not found. Please ensure 'WA_Fn-UseC_-HR-Employee-Attrition.csv' is in the current directory.
Skipping HR Attrition due to loading error.

################################################################################
PROCESSING DATASET: WINE QUALITY
################################################################################
Wine Quality dataset loaded and preprocessed successfully.
Training set shape: (1119, 11)
Testing set shape: (480, 11)
----------------------------

============================================================
RUNNING MANUAL GRID SEARCH FOR WINE QUALITY
============================================================
--- Manual Grid Search for Decision Tree ---
Testing 135 parameter combinations...
  Processed 10/135 combinations. Current best AUC: 0.7796
  Processed 20/135 combinations. Current best AUC: 0.7846
  Processed 30/135 combinations. Current best AUC: 0.7846
  Processed 40/135 combinations. Current best AUC: 0.7850
  Processed 50/135 combinations. Current best AUC: 0.7850
...
--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.7562, Precision: 0.7713
  Recall: 0.7743, F1: 0.7728, AUC: 0.8664
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```




```
============================================================
RUNNING BUILT-IN GRID SEARCH FOR WINE QUALITY
============================================================
```

```
============================================================
RUNNING BUILT-IN GRID SEARCH FOR WINE QUALITY
============================================================

--- GridSearchCV for Decision Tree ---
Fitting 5 folds for each of 135 candidates, totalling 675 fits
c:\Users\mriga\anaconda3\Lib\site-packages\numpy\ma\core.py:2820: RuntimeWarning: invalid value encountered in cast
  _data = np.array(data, dtype=dtype, copy=copy,
Best params for Decision Tree: {'classifier__max_depth': 10, 'classifier__min_samples_leaf': 4, 'classifier__min_samples_split': 10, 'feature_selection__k': 5}
Best CV score: 0.7850

--- GridSearchCV for k-NN ---
Fitting 5 folds for each of 60 candidates, totalling 300 fits
Best params for k-NN: {'classifier__metric': 'manhattan', 'classifier__n_neighbors': 11, 'classifier__weights': 'distance', 'feature_selection__k': 5}
Best CV score: 0.8696

--- GridSearchCV for Logistic Regression ---
Fitting 5 folds for each of 30 candidates, totalling 150 fits
Best params for Logistic Regression: {'classifier__C': 1, 'classifier__penalty': 'l2', 'classifier__solver': 'liblinear', 'feature_selection__k': 11}
Best CV score: 0.8052

============================================================
EVALUATING BUILT-IN MODELS FOR WINE QUALITY
============================================================

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.7250
  Precision: 0.7593
  Recall: 0.7121
  F1-Score: 0.7349
  ROC AUC: 0.7908
```

```
--- Built-in Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.7854, Precision: 0.7939
  Recall: 0.8093, F1: 0.8015, AUC: 0.8664
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```



6. Conclusion

- Manual grid search provided insight into the **mechanics of hyperparameter tuning**, while GridSearchCV proved to be much more efficient and less error-prone.
- **Voting Classifier** consistently outperformed individual classifiers, with **AUC = 0.866**.

- The best-performing single model was **kNN** (AUC = 0.876), but ensemble methods provided more stable and generalizable results.
- This lab highlighted the importance of hyperparameter tuning, cross-validation, and ensemble methods in real-world ML pipelines.