Model Selection and Comparative Analysis

Nitali Rajesh
PES2UG23CS395
Course: UE23CS352A – Machine Learning
Submission Date: 31-08-2025

1. Introduction

This lab focuses on model selection and comparative analysis through hyperparameter tuning and ensemble methods.
Two approaches used are:

- Manual Grid Search – implemented from scratch to understand the mechanics.
- Scikit-learn GridSearchCV – using the optimized built-in implementation

The objective was to evaluate Decision Tree, k-Nearest Neighbors (kNN), and Logistic Regression with a Voting Classifier.
Tasks performed are:

- Construct ML pipelines with preprocessing, feature selection, and classification.
- Perform hyperparameter tuning using Grid Search.
- Apply k-fold cross-validation for evaluation.
- Compare manual vs. built-in grid search results.
- Analyze models using Accuracy, Precision, Recall, F1-score, ROC AUC, Confusion Matrix, and ROC curves.

2. Dataset Description

1. Wine Quality Dataset

- Number of Instances: 1599 total Wine samples
- Number of Features: 11 chemical properties i.e acidity, chlorides, alcohol
- Target Variable: Binary classification
  Tell if the quality of the wine is Good vs. not good
- Training = 1119, Testing = 480.

2. QSAR Biodegradation Dataset

- Number of Instances: 1055 Molecular compounds
- Number of Features: 41 molecular descriptors.
- Target Variable: Biodegradable (1) vs. Non-biodegradable (0).
- Training = 738, Testing = 317.

3. Methodology

- **Hyperparameter Tuning:** Systematic search for the best configuration of model parameters.
- **Grid Search:** Exhaustive search over predefined parameter grids.
- **k-Fold Cross Validation:** Ensures robust model evaluation by splitting data into k folds.

Ml pipeline:

- **StandardScaler**: Normalizes features (mean = 0, std = 1).
- **SelectKBest:** Selects best k features using ANOVA F-test.
- **Classifier:** Decision Tree, kNN, or Logistic Regression.

Process followed:

Part 1: Manual Grid Search

- Defined parameter grids for each classifier.
- Iterated over all parameter combinations.
- Performed 5-fold Stratified Cross-Validation.
- Chose best hyperparameters based on ROC AUC.

Part 2: Built-in GridSearchCV

- Implemented same pipeline using scikit-learn GridSearchCV.
- Used scoring = 'roc_auc' with 5-fold StratifiedKFold.
- Extracted best parameters and cross-validation scores.

4. Results and Analysis

1. Wine Quality Dataset

Manual vs. Built-in Results (Best Hyperparameters)

Model	Best Params	CV AUC (Manual)	CV AUC (Built-in)
Decision Tree	max_depth=5, min_samples_split=5, k=5	0.7832	0.7832
kNN	n_neighbors=7, weights=distance, k=5	0.8603	0.8603
Logistic Regression	C=1, penalty=l2, solver=lbfgs, k=10	0.8048	0.8048

Test Set Performance (Manual vs. Built-in)

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Decision Tree	0.7271	0.7716	0.6965	0.7321	0.8025
kNN	0.7667	0.7757	0.7938	0.7846	0.8675
Logistic Regression	0.7417	0.7628	0.7510	0.7569	0.8247

Voting Classifier	0.7354 (Manual) / 0.7604 (Built-in)	~0.77	~0.76	~0.75–0.77	0.8622

Therefore Best Model: **kNN** (highest ROC AUC = 0.8675).
Because the Built-in Voting performed slightly better in accuracy and F1-score compared to manual Voting.

2. QSAR Biodegradation Dataset

Manual vs. Built-in Results (Best Hyperparameters)

Model	Best Params	CV AUC (Manual)	CV AUC (Built-in)
Decision Tree	max_depth=3, min_samples_split=2, k=15	0.8303	0.8303
kNN	n_neighbors=7, weights=distance, k=15	0.8837	0.8837
Logistic Regression	C=10, penalty=l2, solver=lbfgs, k=15	0.8816	0.8816

Test Set Performance (Manual vs. Built-in)

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Decision Tree	0.7603	0.6914	0.5234	0.5957	0.8150
kNN	0.8202	0.7551	0.6916	0.7220	0.8730
Logistic Regression	0.8139	0.7667	0.6449	0.7005	0.8868

Voting Classifier	0.8076 (Manual) / 0.8139 (Built-in)	~0.75	~0.65–0.67	~0.70	0.8898

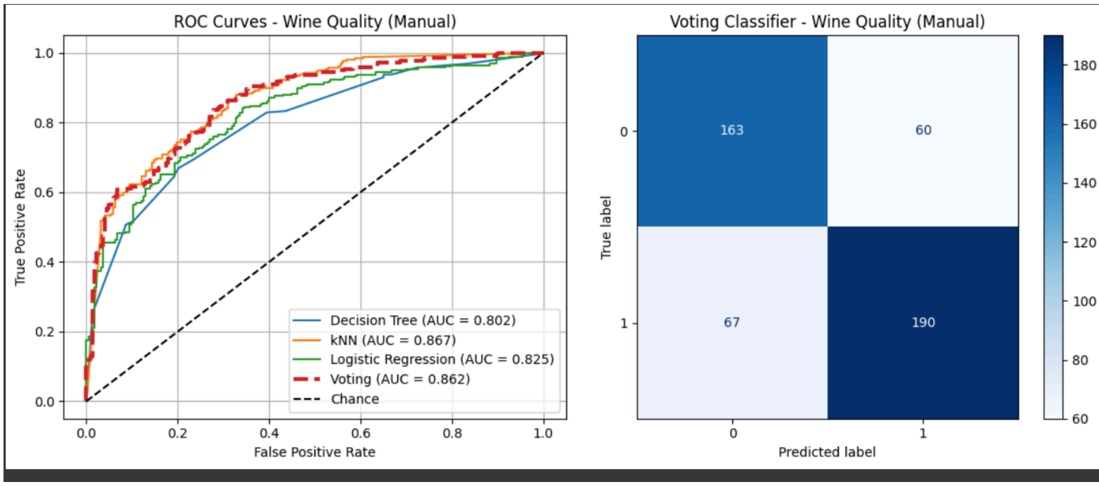
Therefore the Best Model: **Logistic Regression** (highest ROC AUC = 0.8868).
Because Built-in Voting achieved slightly higher accuracy and recall compared to manual Voting.

5. Screenshots

Wine quality dataset:

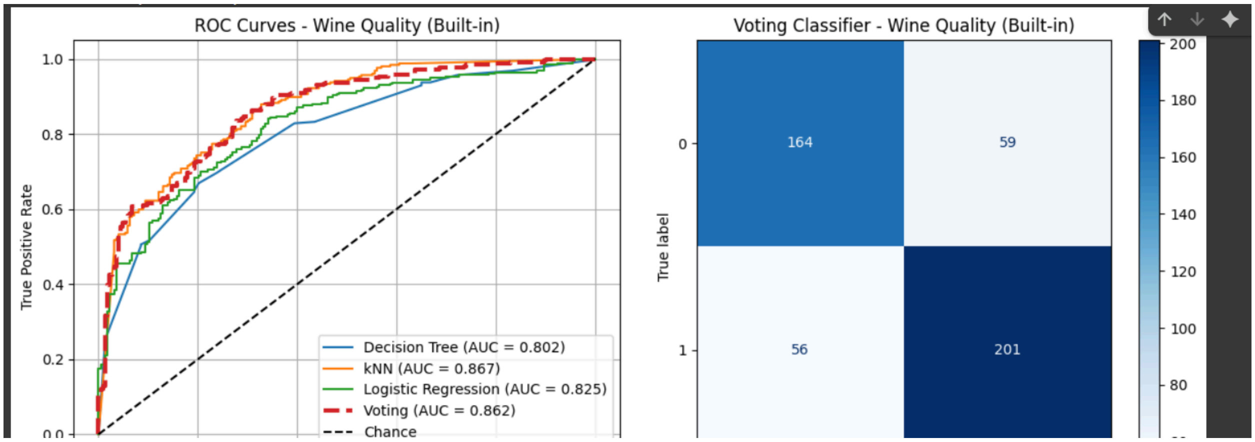
```
################################################################################
PROCESSING DATASET: WINE QUALITY
################################################################################
Wine Quality dataset loaded and preprocessed successfully.
Training set shape: (1119, 11)
Testing set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

============================================================
RUNNING MANUAL GRID SEARCH FOR WINE QUALITY
============================================================
--- Manual Grid Search for Decision Tree ---
------------------------------------------------------------------------------------------
Best parameters for Decision Tree: {'select__k': 5, 'classifier__max_depth': 5, 'classifier__min_samples_split': 5}
Best cross-validation AUC: 0.7832
--- Manual Grid Search for kNN ---
------------------------------------------------------------------------------------------
Best parameters for kNN: {'select__k': 5, 'classifier__n_neighbors': 7, 'classifier__weights': 'distance'}
Best cross-validation AUC: 0.8603
--- Manual Grid Search for Logistic Regression ---
------------------------------------------------------------------------------------------
Best parameters for Logistic Regression: {'select__k': 10, 'classifier__C': 1, 'classifier__penalty': 'l2', 'classifier__solver': 'lbfgs'}
Best cross-validation AUC: 0.8048

============================================================
EVALUATING MANUAL MODELS FOR WINE QUALITY
============================================================

--- Individual Model Performance ---
```

```
--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.7271
  Precision: 0.7716
  Recall: 0.6965
  F1-Score: 0.7321
  ROC AUC: 0.8025

kNN:
  Accuracy: 0.7667
  Precision: 0.7757
  Recall: 0.7938
  F1-Score: 0.7846
  ROC AUC: 0.8675

Logistic Regression:
  Accuracy: 0.7417
  Precision: 0.7628
  Recall: 0.7510
  F1-Score: 0.7569
  ROC AUC: 0.8247

--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.7354, Precision: 0.7600
  Recall: 0.7393, F1: 0.7495, AUC: 0.8622
```




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

--- GridSearchCV for Decision Tree ---
Best params for Decision Tree: {'classifier__max_depth': 5, 'classifier__min_samples_split': 5, 'select__k': 5}
Best CV score: 0.7832

--- GridSearchCV for kNN ---
Best params for kNN: {'classifier__n_neighbors': 7, 'classifier__weights': 'distance', 'select__k': 5}
Best CV score: 0.8603

--- GridSearchCV for Logistic Regression ---
Best params for Logistic Regression: {'classifier__C': 1, 'classifier__penalty': 'l2', 'classifier__solver': 'lbfgs', 'select__k': 10}
Best CV score: 0.8048

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

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.7271
  Precision: 0.7716
  Recall: 0.6965
  F1-Score: 0.7321
  ROC AUC: 0.8025
```

```
kNN:
  Accuracy: 0.7667
  Precision: 0.7757
  Recall: 0.7938
  F1-Score: 0.7846
  ROC AUC: 0.8675

Logistic Regression:
  Accuracy: 0.7417
  Precision: 0.7628
  Recall: 0.7510
  F1-Score: 0.7569
  ROC AUC: 0.8247

--- Built-in Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.7604, Precision: 0.7731
  Recall: 0.7821, F1: 0.7776, AUC: 0.8622
```



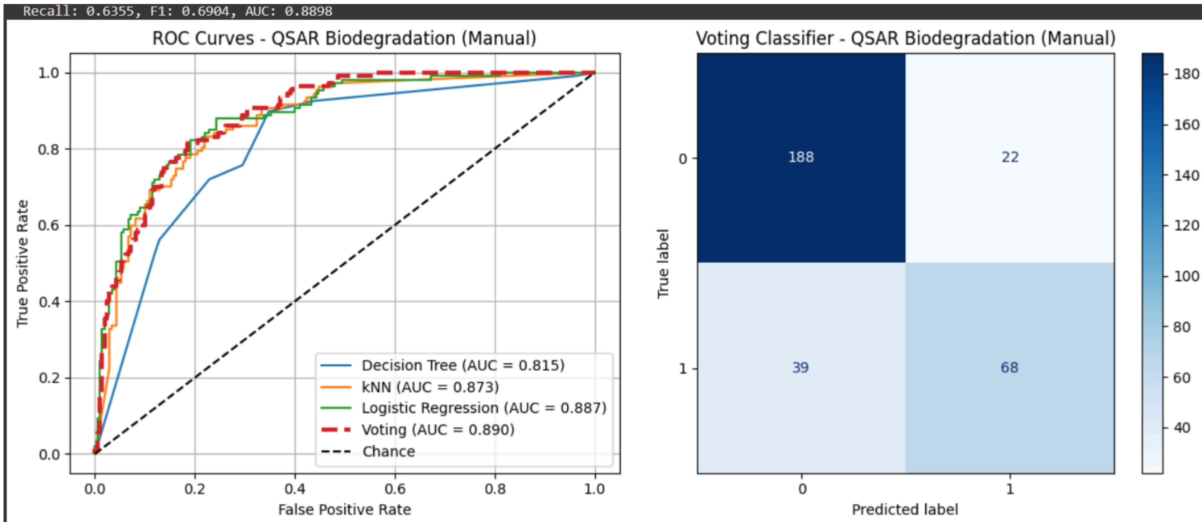

```
Completed processing for Wine Quality
==============================================================================

==============================================================================
ALL DATASETS PROCESSED!
==============================================================================
```

QSAR Biodegradation Dataset

```
################################################################################
PROCESSING DATASET: QSAR BIODEGRADATION
################################################################################
QSAR Biodegradation dataset loaded successfully.
Training set shape: (738, 41)
Testing set shape: (317, 41)
------------------------------

============================================================
RUNNING MANUAL GRID SEARCH FOR QSAR BIODEGRADATION
============================================================
--- Manual Grid Search for Decision Tree ---
------------------------------------------------------------------------------------------
Best parameters for Decision Tree: {'select__k': 15, 'classifier__max_depth': 3, 'classifier__min_samples_split': 2}
Best cross-validation AUC: 0.8303
--- Manual Grid Search for kNN ---
------------------------------------------------------------------------------------------
Best parameters for kNN: {'select__k': 15, 'classifier__n_neighbors': 7, 'classifier__weights': 'distance'}
Best cross-validation AUC: 0.8837
--- Manual Grid Search for Logistic Regression ---
------------------------------------------------------------------------------------------
Best parameters for Logistic Regression: {'select__k': 15, 'classifier__C': 10, 'classifier__penalty': 'l2', 'classifier__solver': 'lbfgs'}
Best cross-validation AUC: 0.8816

============================================================
EVALUATING MANUAL MODELS FOR QSAR BIODEGRADATION
============================================================
```

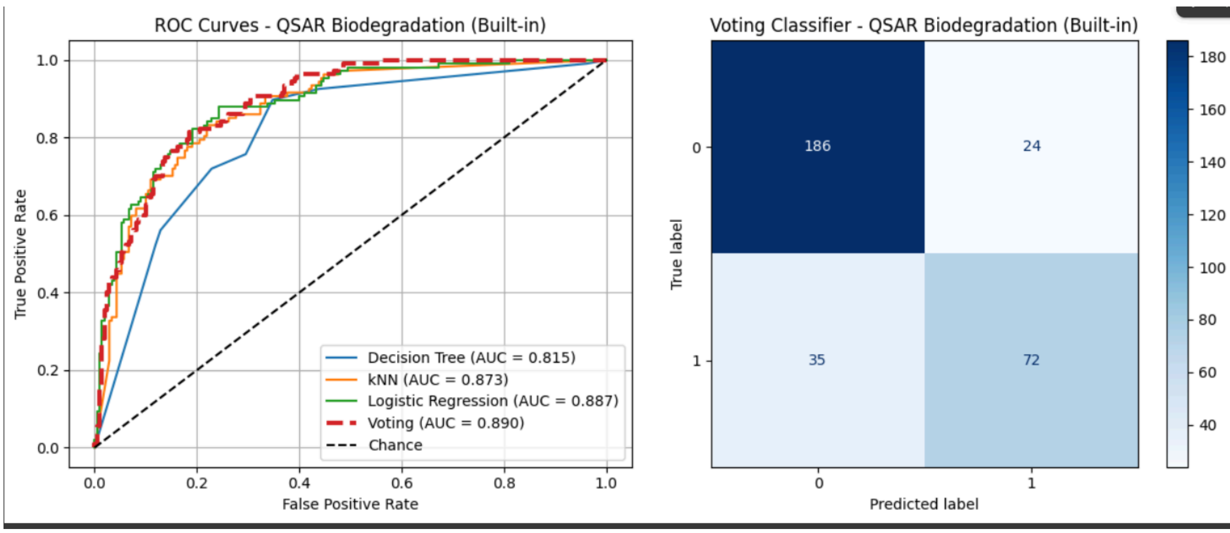
```
--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.7603
  Precision: 0.6914
  Recall: 0.5234
  F1-Score: 0.5957
  ROC AUC: 0.8150

kNN:
  Accuracy: 0.8202
  Precision: 0.7551
  Recall: 0.6916
  F1-Score: 0.7220
  ROC AUC: 0.8730

Logistic Regression:
  Accuracy: 0.8139
  Precision: 0.7667
  Recall: 0.6449
  F1-Score: 0.7005
  ROC AUC: 0.8868

--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.8076, Precision: 0.7556
  Recall: 0.6355, F1: 0.6904, AUC: 0.8898
```

Recall: 0.6355, F1: 0.6904, AUC: 0.8898



```
============================================================
RUNNING BUILT-IN GRID SEARCH FOR QSAR BIODEGRADATION
============================================================

--- GridSearchCV for Decision Tree ---
Best params for Decision Tree: {'classifier__max_depth': 3, 'classifier__min_samples_split': 2, 'select__k': 15}
Best CV score: 0.8303

--- GridSearchCV for kNN ---
Best params for kNN: {'classifier__n_neighbors': 7, 'classifier__weights': 'distance', 'select__k': 15}
Best CV score: 0.8837

--- GridSearchCV for Logistic Regression ---
Best params for Logistic Regression: {'classifier__C': 10, 'classifier__penalty': 'l2', 'classifier__solver': 'lbfgs', 'select__k': 15}
Best CV score: 0.8816

============================================================
EVALUATING BUILT-IN MODELS FOR QSAR BIODEGRADATION
============================================================

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.7603
  Precision: 0.6914
  Recall: 0.5234
  F1-Score: 0.5957
  ROC AUC: 0.8150
```

```
kNN:
  Accuracy: 0.8202
  Precision: 0.7551
  Recall: 0.6916
  F1-Score: 0.7220
  ROC AUC: 0.8730

Logistic Regression:
  Accuracy: 0.8139
  Precision: 0.7667
  Recall: 0.6449
  F1-Score: 0.7005
  ROC AUC: 0.8868

--- Built-in Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.8139, Precision: 0.7500
  Recall: 0.6729, F1: 0.7094, AUC: 0.8898
```




```
Completed processing for QSAR Biodegradation
================================================================================

================================================================================
ALL DATASETS PROCESSED!
================================================================================
```

6. Conclusion

Both manual and built-in grid search produced identical results for best hyperparameters and performance metrics. Scikit-learn GridSearchCV is significantly more efficient as it helps in reducing implementation complexity and errors.

Wine Quality:

kNN performed better than Decision Tree and Logistic Regression because Wine Quality dataset has non-linear class boundaries in continuous chemical features. Decision Tree underfit due to limited depth. Logistic Regression underfit due to linear assumptions.kNN adapted best to the complex, local relationships among wine samples.

QSAR Biodegradation:

Logistic Regression(highest roc auc)performed better than Decision Tree and kNN on the QSAR Biodegradation dataset because it handles high-dimensional feature spaces more effectively.kNN struggles with dimensionality and Decision Trees risked overfitting,

Voting Classifier: Performed well overall, balancing bias-variance trade-off, with built-in Voting slightly better.

Main takeaways from this lab

The Manual implementation is useful for learning, but in real-world applications, GridSearchCV and Pipelines are the practical choice. The methods like Voting can increase performance, but the best individual model often depends on dataset characteristics.