Model Selection and Comparative Analysis

1. Introduction

This lab explored model selection and hyperparameter tuning using both a manual grid search (to understand the process) and Scikit-learn's GridSearchCV (for efficiency). Three models were tested—Decision Trees, kNN, and Logistic Regression—and later combined in a Voting Classifier.

The work involved building pipelines with preprocessing and feature selection, tuning parameters through grid search, and evaluating models with k-fold cross-validation. Performance was compared using metrics like Accuracy, Precision, Recall, F1-score, ROC-AUC, along with visual analysis through Confusion Matrices and ROC curves.

In short, the lab highlighted how different approaches to tuning and combining models affect overall performance**.**

2. Dataset Description

1. Wine Quality Dataset

● Number of Instances: 1599 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 samples

● Number of Features: 41 molecular descriptors.

● Target Variable: Biodegradable (1) vs. non-biodegradable (0).

● Training = 738, Testing = 317.

3. Methodology

The core idea of this lab was to find the best set of hyperparameters for different models and evaluate them using a consistent process.

To ensure fair comparison, we built a machine learning pipeline with:

- StandardScaler for feature normalization (mean = 0, std = 1),

- SelectKBest (with ANOVA F-test) for feature selection, and
- A classifier (Decision Tree, kNN, or Logistic Regression).

For tuning and evaluation, two approaches were followed:

Part 1: Manual Grid Search

- Defined parameter grids separately for each model.
- Iterated through all possible parameter combinations.
- Applied 5-fold Stratified Cross-Validation for evaluation.
- Selected the best parameters based on ROC-AUC scores.

Part 2: GridSearchCV (Scikit-learn)

- Replicated the same pipeline setup using the built-in GridSearchCV.
- Used scoring = 'roc_auc' with StratifiedKFold (k=5) for consistent evaluation.
- Extracted both the best hyperparameters and the corresponding cross-validation performance.

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	00.7628	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	00.7667	0.6449	0.7005	0.8868

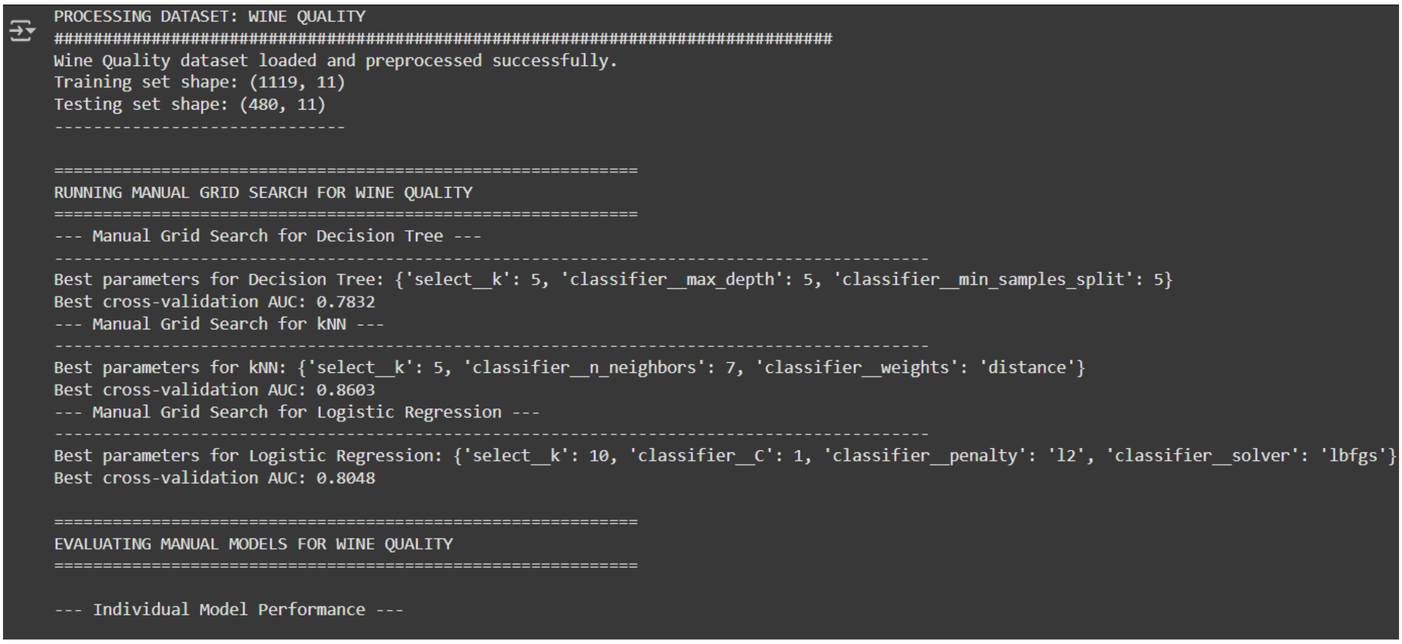
Voting classifier	0.8076(manual)
	0.8139(built-in)
		~0.75	~0.65-0.67	~0.70	0.8898

Therefore the Best Model: Logistic Regression (highest ROC AUC = 0.8868).

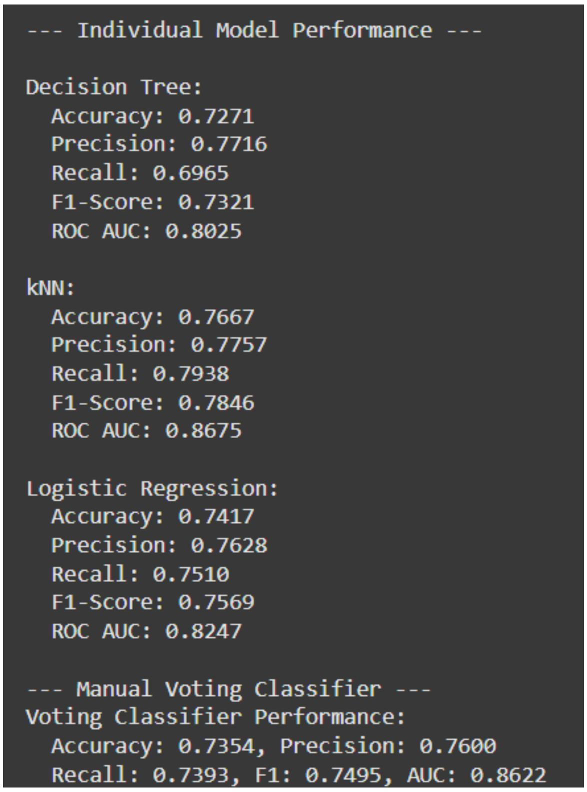
Because Built-in Voting achieved slightly higher accuracy and recall compared to manual Voting.

5. Screenshots

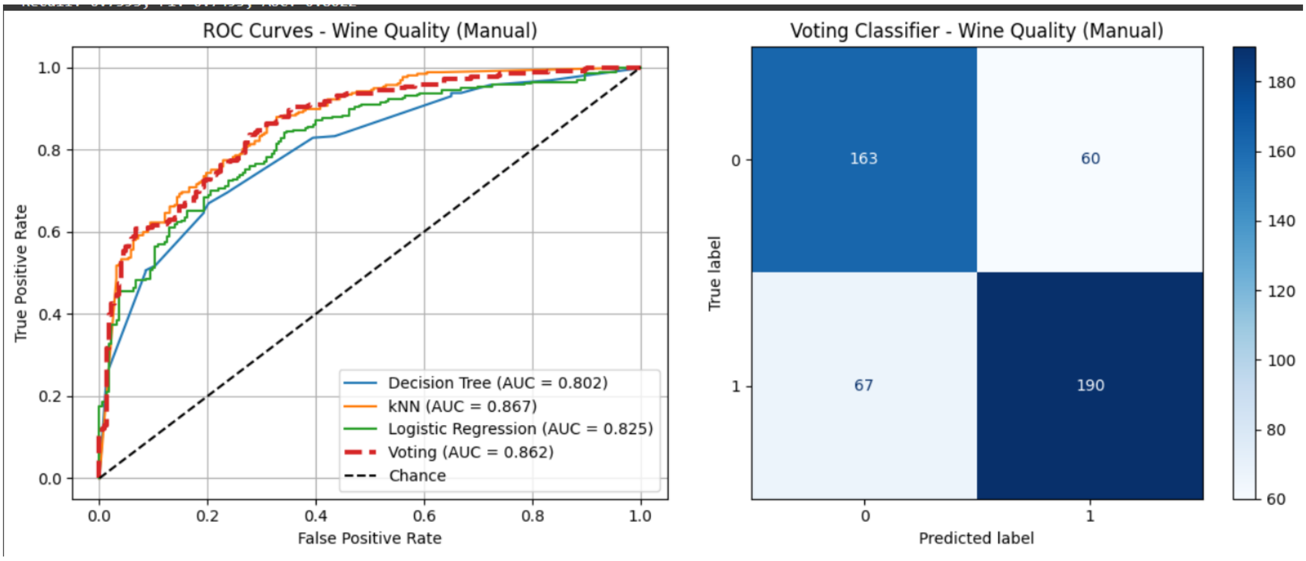
Wine dataset:

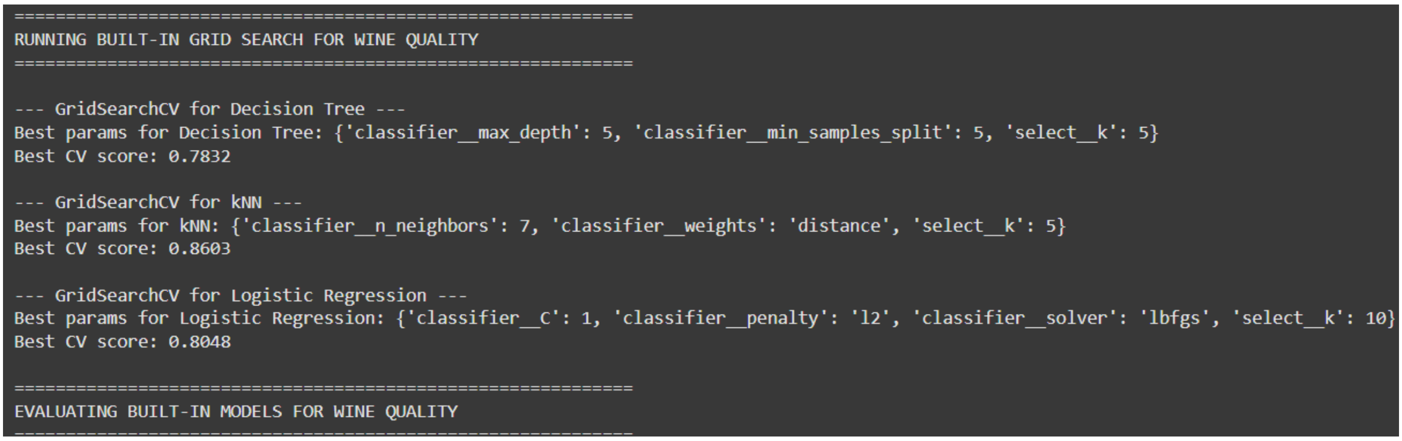


```
PROCESSING DATASET: WINE QUALITY
################################################################################
Wine Quality dataset loaded and preprocessed successfully.
Training set shape: (1119, 11)
Testing set shape: (480, 11)
------------------------------

============================================================
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 ---
```

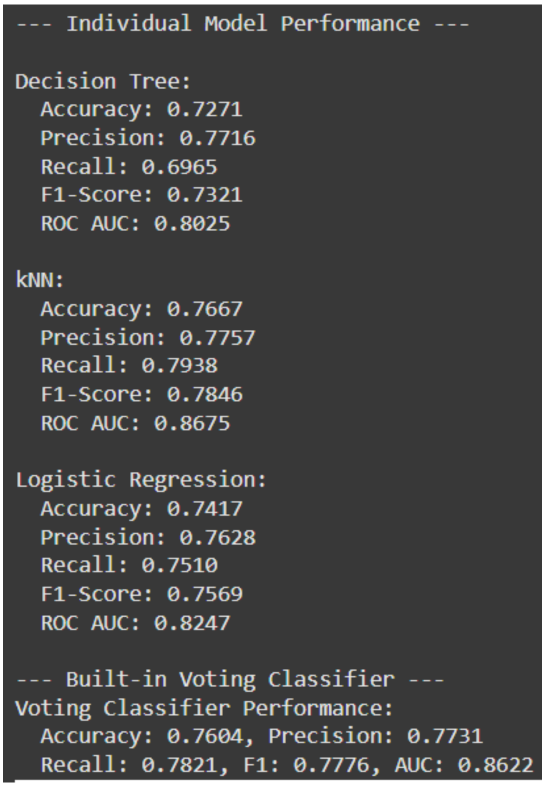


```
--- Individual Model Performance ---

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
```

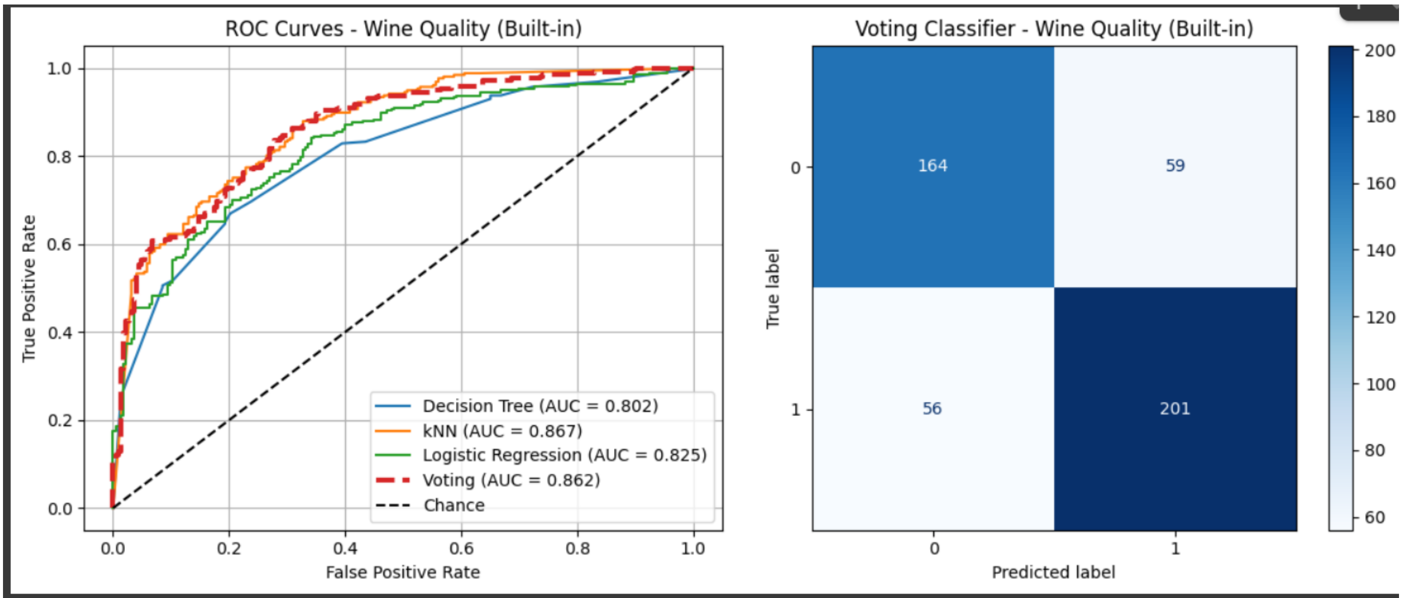




```
============================================================
RUNNING BUILT-IN GRID SEARCH FOR WINE QUALITY
============================================================

--- 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
============================================================
```

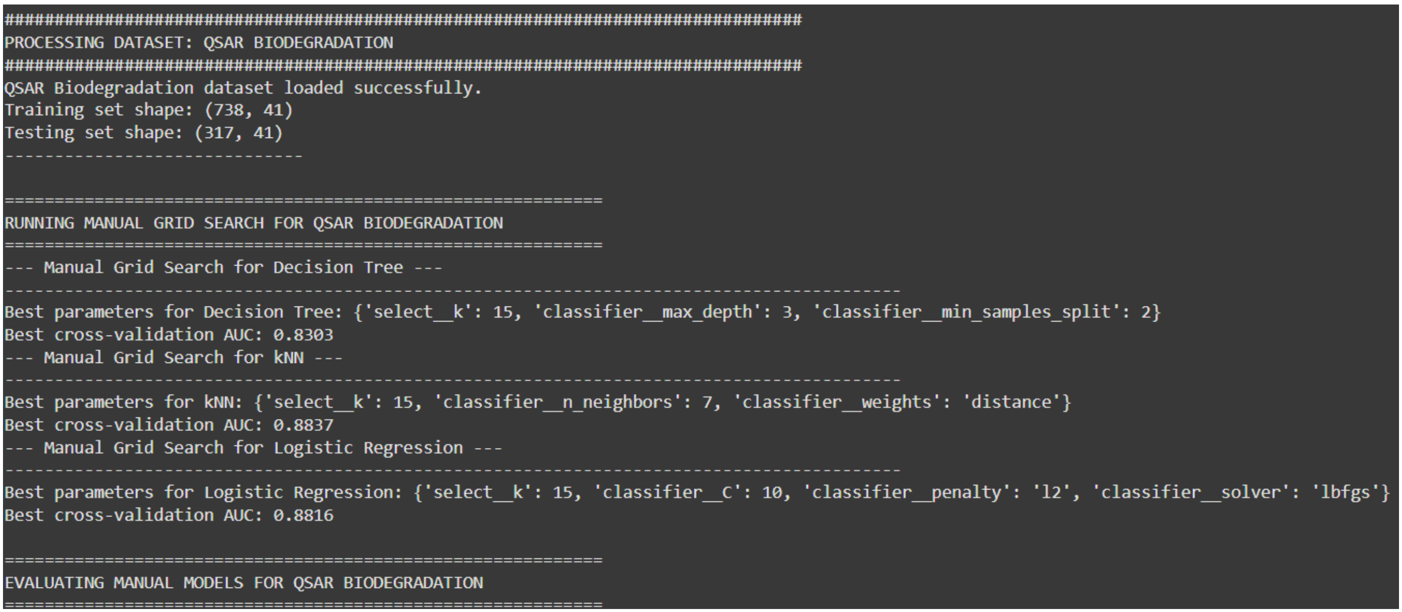


```
--- Individual Model Performance ---

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
```

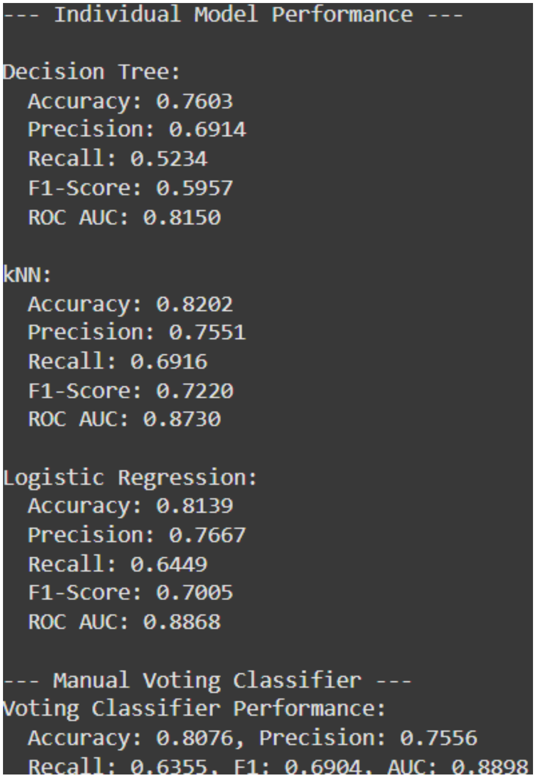


```
Completed processing for Wine Quality
================================================================================

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

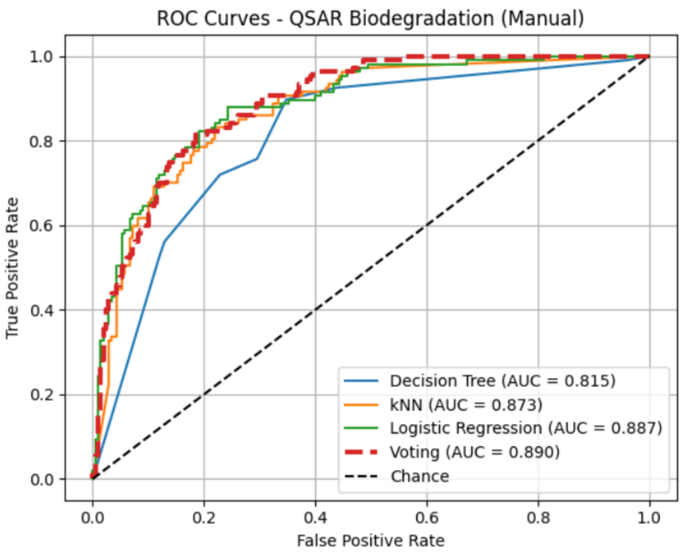
QSAR BIODEGRADATION dataset:

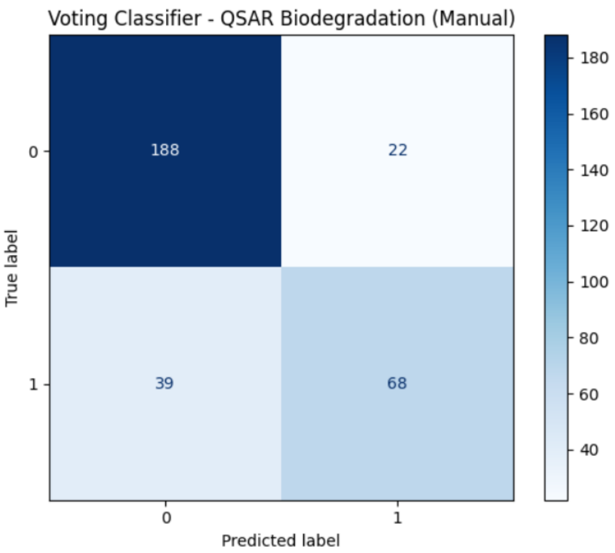


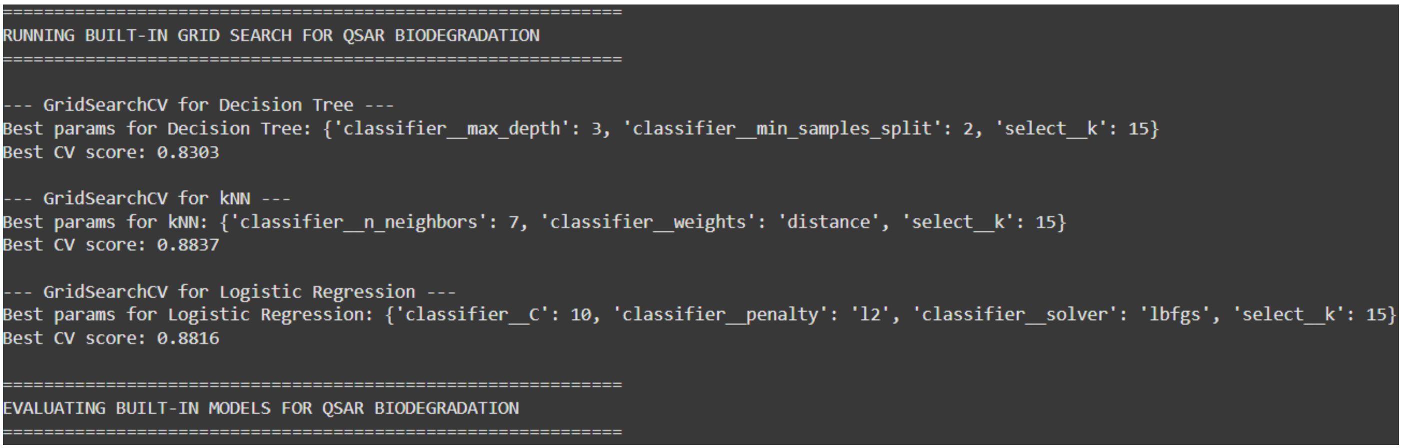
```
################################################################################
PROCESSING DATASET: QSAR BIODEGRADATION
################################################################################
QSAR Biodegradation dataset loaded successfully.
Training set shape: (738, 41)
Testing set shape: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



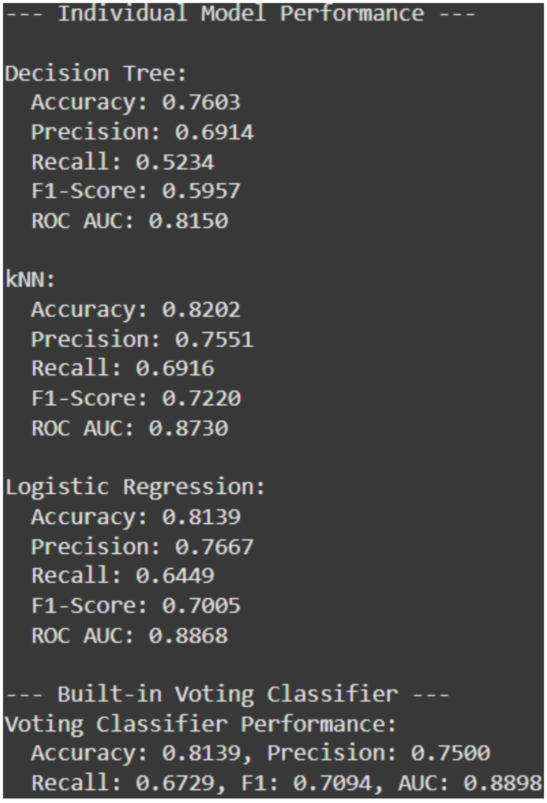
```
--- Individual Model Performance ---

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
```



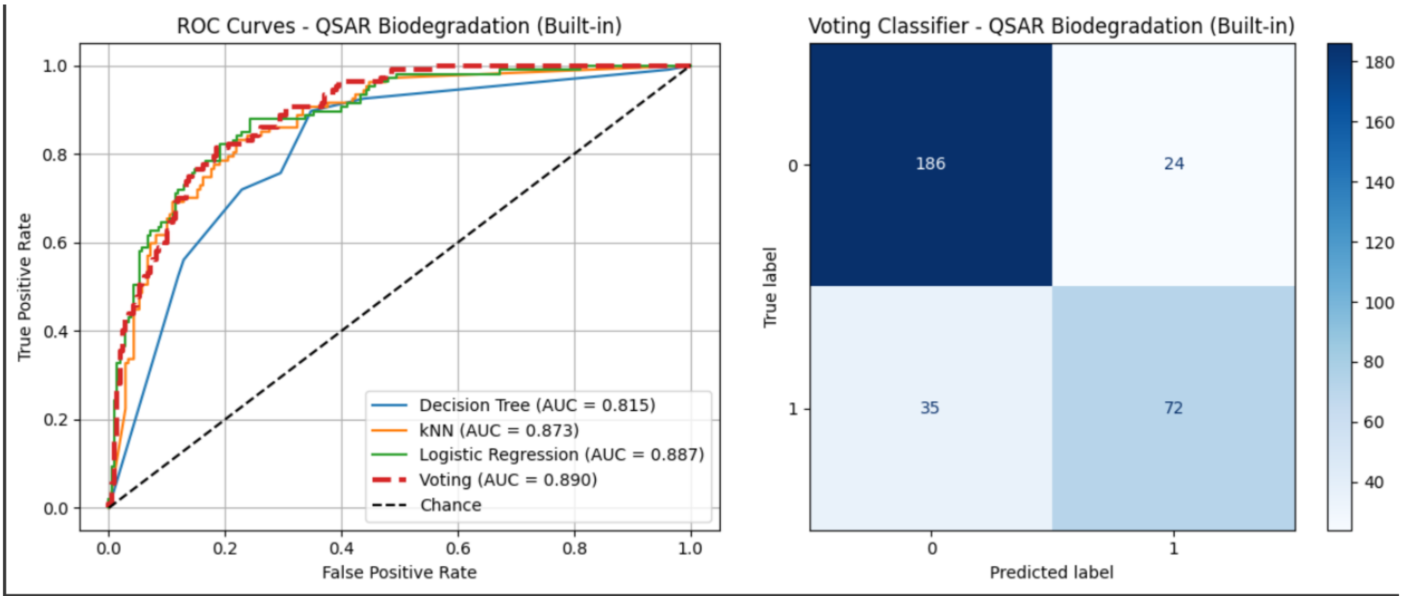


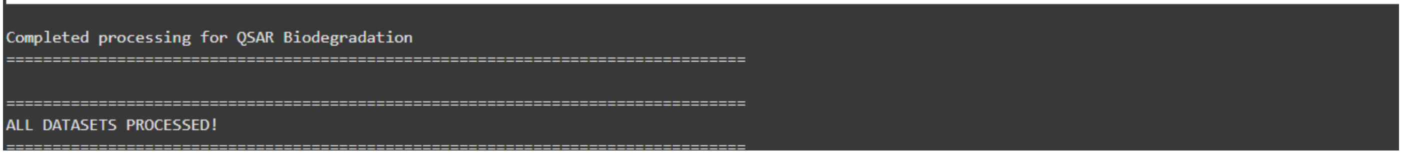


```
============================================================
RUNNING BUILT-IN GRID SEARCH FOR QSAR BIODEGRADATION
============================================================

--- 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 the manual grid search and Scikit-learn's GridSearchCV arrived at the same best hyperparameters and performance results. However, GridSearchCV proved to be far more efficient, reducing the coding effort and minimizing chances of error.

For the Wine Quality dataset, kNN outperformed Decision Tree and Logistic Regression. This is because the dataset involves non-linear class boundaries in continuous chemical features where:

- Decision Tree underfit due to shallow depth,
- Logistic Regression struggled with its linear assumption,
- kNN worked best by adapting to the local and complex feature relationships.

For the QSAR Biodegradation dataset, Logistic Regression achieved the highest ROC-AUC. Its strength in handling high-dimensional feature spaces gave it an edge, while:

- kNN struggled with the curse of dimensionality,
- Decision Tree leaned toward overfitting.

The Voting Classifier overall provided balanced performance, benefiting from the bias-variance trade-off, with the built-in version slightly ahead of individual models.

Key Takeaways:

- Manual grid search is valuable for understanding the process, but in practice, GridSearchCV and pipelines are the go-to tools.
- Ensemble methods like Voting can improve results, but the "best" model often depends on the specific characteristics of the dataset.