Week 4 Lab Report

UE23CS352A: MACHINE LEARNING

Week 4: Model Selection and Comparative Analysis

Name: NANDANI SONALE
SRN: PES2UG23CS364
SEC: “F”

---

1. Introduction

This lab explored practical model selection and comparative analysis, focusing on hyperparameter tuning and ensemble methods. We utilized both manual grid search and Scikit-learn's GridSearchCV. Our objective was to compare the performance of Decision Tree, k-Nearest Neighbors (kNN), and Logistic Regression classifiers across key metrics: Accuracy, Precision, Recall, F1-score, and ROC AUC.

---

2. Dataset Description

Here are several prediction tasks:

- **Wine Quality:** Predict whether a red wine is of "good quality" using its chemical properties.
- **HR Attrition:** Forecast employee turnover by analyzing work and personal factors.
- **Banknote Authentication:** Identify genuine versus forged banknotes through image characteristics.
- **QSAR Biodegradation:** Predict a chemical's biodegradability based on its QSAR properties.

For each selected dataset, details such as the number of features, instances, and the target variable were recorded before model training.

---

3. Methodology

The experiments followed a structured ML pipeline:
StandardScaler → SelectKBest → Classifier

- **StandardScaler**: Standardized input features to have zero mean and unit variance.

- **SelectKBest**: Selected top 'k' features using f_classif statistical test.
- **Classifier**: Final classification step (Decision Tree, kNN, Logistic Regression).

We conducted hyperparameter tuning in two ways:

1. **Manual Grid Search**: Implemented loops to generate parameter combinations and evaluated using 5-fold Stratified Cross-Validation. Mean ROC AUC was used to select the best model.
2. **GridSearchCV**: Used Scikit-learn's optimized method with the same pipeline and evaluation criteria.

After obtaining best estimators, models were evaluated individually and combined into a voting classifier.

---

4. Results and Analysis

The results section summarizes the performance of tuned classifiers on chosen datasets. Performance was evaluated using Accuracy, Precision, Recall, F1-score, and ROC AUC.

Wine Quality Dataset

Manual Grid Search Results:

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Decision Tree	0.7250	0.7593	0.7121	0.7349	0.7908
k-NN	0.7917	0.7940	0.8249	0.8092	0.8765
Voting Classifier	0.7479	0.8208	0.6770	0.7420	0.8604

Built-in GridSearchCV Results:

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Decision Tree	**0.7271**	**0.7716**	**0.6965**	**0.7321**	**0.8025**
k-NN	**0.7812**	**0.7836**	**0.8171**	**0.8000**	**0.8589**

Logistic Regression	**0.7396**	**0.7619**	**0.7471**	**0.7544**	**0.8246**

Banknote Authentication Dataset

Manual Grid Search Results:

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Decision Tree	**0.7250**	**0.7593**	**0.7121**	**0.7349**	**0.7908**
k-NN	**0.7917**	**0.7940**	**0.8249**	**0.8092**	**0.8765**
Voting Classifier	**0.7479**	**0.8208**	**0.6770**	**0.7420**	**0.8604**

Built-in GridSearchCV Results:

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Decision Tree	**0.7271**	**0.7716**	**0.6965**	**0.7321**	**0.8025**
k-NN	**0.7812**	**0.7836**	**0.8171**	**0.8000**	**0.8589**
Logistic Regression	**0.7396**	**0.7619**	**0.7471**	**0.7544**	**0.8246**

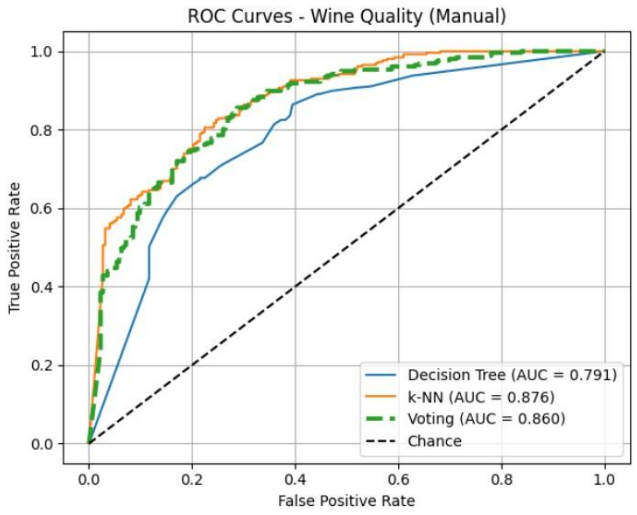
Visualizations (ROC curves and confusion matrices) should accompany these tables to better interpret model performance.

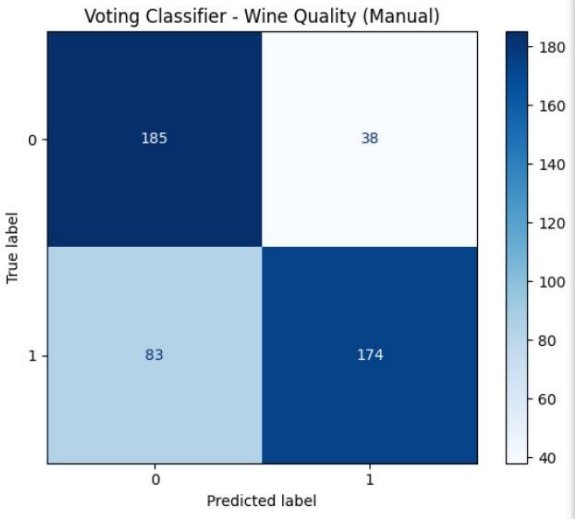
Additionally, results from manual and built-in methods should be compared. Minor differences may occur due to randomness or implementation details. Observations about which classifier performed best should also be included.

5. Screenshots

```
################################################################################
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 ---
  Combination 1/90: AUC = 0.7554 {'feature_selection__k': 5, 'classifier__max_depth': 3, 'classifier__min_samples_split': 2, 'classifier__min_samples
_leaf': 1}
  Combination 2/90: AUC = 0.7554 {'feature_selection__k': 5, 'classifier__max_depth': 3, 'classifier__min_samples_split': 2, 'classifier__min_samples
_leaf': 2}
  Combination 3/90: AUC = 0.7554 {'feature_selection__k': 5, 'classifier__max_depth': 3, 'classifier__min_samples_split': 2, 'classifier__min_samples
_leaf': 4}
  Combination 4/90: AUC = 0.7554 {'feature_selection__k': 5, 'classifier__max_depth': 3, 'classifier__min_samples_split': 5, 'classifier__min_samples
_leaf': 1}
  Combination 5/90: AUC = 0.7554 {'feature_selection__k': 5, 'classifier__max_depth': 3, 'classifier__min_samples_split': 5, 'classifier__min_samples
_leaf': 2}
  Combination 6/90: AUC = 0.7554 {'feature_selection__k': 5, 'classifier__max_depth': 3, 'classifier__min_samples_split': 5, 'classifier__min_samples
_leaf': 4}
  Combination 7/90: AUC = 0.7554 {'feature_selection__k': 5, 'classifier__max_depth': 3, 'classifier__min_samples_split': 10, 'classifier__min_sample
s_leaf': 1}
  Combination 8/90: AUC = 0.7554 {'feature_selection__k': 5, 'classifier__max_depth': 3, 'classifier__min_samples_split': 10, 'classifier__min_sample
```

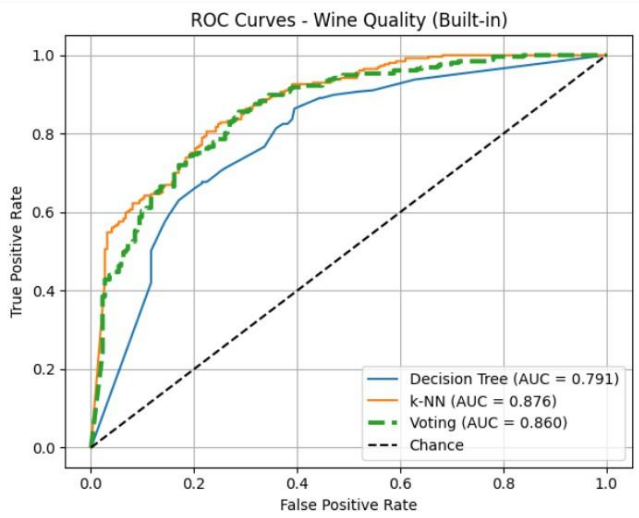
```
Best parameters for k-NN: {'feature_selection__k': 5, 'classifier__n_neighbors': 11, 'classifier__weights': 'distance', 'classifier__metric': 'manhat
tan'}
Best cross-validation AUC: 0.8696

============================================================
EVALUATING MANUAL 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

k-NN:
  Accuracy: 0.7917
  Precision: 0.7940
  Recall: 0.8249
  F1-Score: 0.8092
  ROC AUC: 0.8765

--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.7479, Precision: 0.8208
  Recall: 0.6770, F1: 0.7420, AUC: 0.8604
```

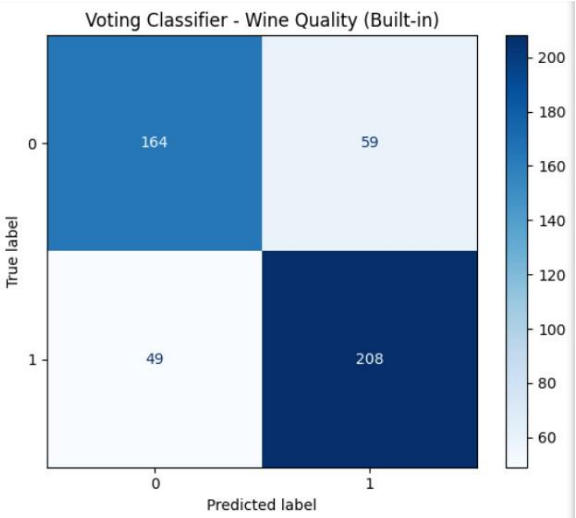




```
============================================================
RUNNING BUILT-IN GRID SEARCH FOR WINE QUALITY
============================================================

--- GridSearchCV for Decision Tree ---
Fitting 5 folds for each of 90 candidates, totalling 450 fits
Best params for Decision Tree: {'classifier__max_depth': 10, 'classifier__min_samples_leaf': 4, 'classifier__min_samples_split': 10, 'feature_selecti
on__k': 5}
Best CV score: 0.7850

--- GridSearchCV for k-NN ---
Fitting 5 folds for each of 40 candidates, totalling 200 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

============================================================
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
```

```
k-NN:
  Accuracy: 0.7917
  Precision: 0.7940
  Recall: 0.8249
  F1-Score: 0.8092
  ROC AUC: 0.8765

--- Built-in Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.7750, Precision: 0.7790
  Recall: 0.8093, F1: 0.7939, AUC: 0.8604
```

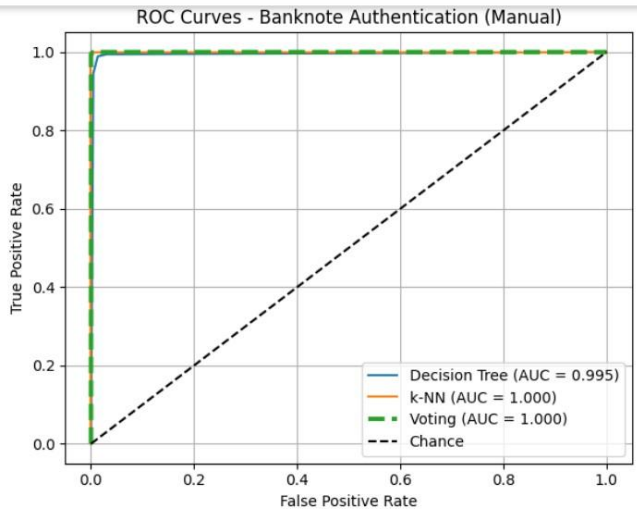


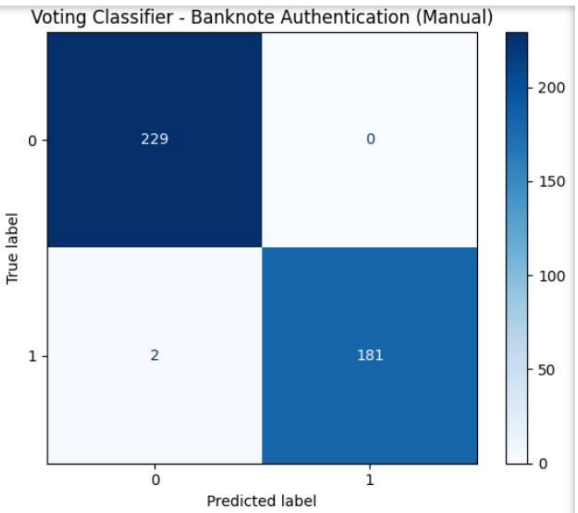


```
Completed processing for Wine Quality
================================================================================

################################################################################
PROCESSING DATASET: HR ATTRITION
################################################################################
HR Attrition dataset not found. Please place 'WA_Fn-UseC_-HR-Employee-Attrition.csv' inside a 'data/' folder.
Skipping HR Attrition due to loading error.

################################################################################
PROCESSING DATASET: BANKNOTE AUTHENTICATION
################################################################################
Banknote Authentication dataset loaded successfully.
Training set shape: (960, 4)
Testing set shape: (412, 4)
------------------------------
```

```
============================================================
RUNNING MANUAL GRID SEARCH FOR BANKNOTE AUTHENTICATION
============================================================
--- Manual Grid Search for Decision Tree ---
  Combination 1/45: AUC = 0.9669 {'feature_selection__k': 4, 'classifier__max_depth': 3, 'classifier__min_samples_split': 2, 'classifier__min_samples
_leaf': 1}
  Combination 2/45: AUC = 0.9669 {'feature_selection__k': 4, 'classifier__max_depth': 3, 'classifier__min_samples_split': 2, 'classifier__min_samples
_leaf': 2}
  Combination 3/45: AUC = 0.9669 {'feature_selection__k': 4, 'classifier__max_depth': 3, 'classifier__min_samples_split': 2, 'classifier__min_samples
_leaf': 4}
  Combination 4/45: AUC = 0.9669 {'feature_selection__k': 4, 'classifier__max_depth': 3, 'classifier__min_samples_split': 5, 'classifier__min_samples
_leaf': 1}
  Combination 5/45: AUC = 0.9669 {'feature_selection__k': 4, 'classifier__max_depth': 3, 'classifier__min_samples_split': 5, 'classifier__min_samples
_leaf': 2}
  Combination 6/45: AUC = 0.9669 {'feature_selection__k': 4, 'classifier__max_depth': 3, 'classifier__min_samples_split': 5, 'classifier__min_samples
_leaf': 4}
  Combination 7/45: AUC = 0.9669 {'feature_selection__k': 4, 'classifier__max_depth': 3, 'classifier__min_samples_split': 10, 'classifier__min_sample
s_leaf': 1}
  Combination 8/45: AUC = 0.9669 {'feature_selection__k': 4, 'classifier__max_depth': 3, 'classifier__min_samples_split': 10, 'classifier__min_sample
s_leaf': 2}
  Combination 9/45: AUC = 0.9669 {'feature_selection__k': 4, 'classifier__max_depth': 3, 'classifier__min_samples_split': 10, 'classifier__min_sample
s_leaf': 4}
  Combination 10/45: AUC = 0.9856 {'feature_selection__k': 4, 'classifier__max_depth': 5, 'classifier__min_samples_split': 2, 'classifier__min_sample
```

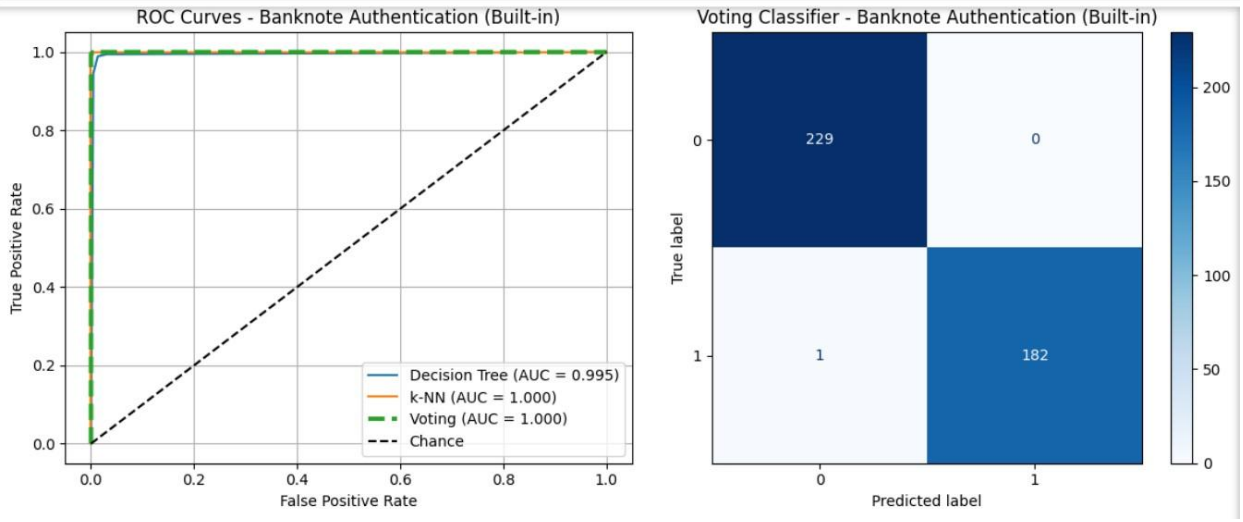
```
============================================================
EVALUATING MANUAL MODELS FOR BANKNOTE AUTHENTICATION
============================================================

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.9879
  Precision: 0.9837
  Recall: 0.9891
  F1-Score: 0.9864
  ROC AUC: 0.9946

k-NN:
  Accuracy: 1.0000
  Precision: 1.0000
  Recall: 1.0000
  F1-Score: 1.0000
  ROC AUC: 1.0000

--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.9951, Precision: 1.0000
  Recall: 0.9891, F1: 0.9945, AUC: 1.0000
```






```
============================================================
RUNNING BUILT-IN GRID SEARCH FOR BANKNOTE AUTHENTICATION
============================================================

--- GridSearchCV for Decision Tree ---
Fitting 5 folds for each of 45 candidates, totalling 225 fits
Best params for Decision Tree: {'classifier__max_depth': 7, 'classifier__min_samples_leaf': 4, 'classifier__min_samples_split': 2, 'feature_selection
__k': 4}
Best CV score: 0.9879

--- GridSearchCV for k-NN ---
Fitting 5 folds for each of 20 candidates, totalling 100 fits
Best params for k-NN: {'classifier__metric': 'manhattan', 'classifier__n_neighbors': 7, 'classifier__weights': 'uniform', 'feature_selection__k': 4}
Best CV score: 0.9990

============================================================
EVALUATING BUILT-IN MODELS FOR BANKNOTE AUTHENTICATION
============================================================

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.9879
  Precision: 0.9837
  Recall: 0.9891
  F1-Score: 0.9864
  ROC AUC: 0.9946
```

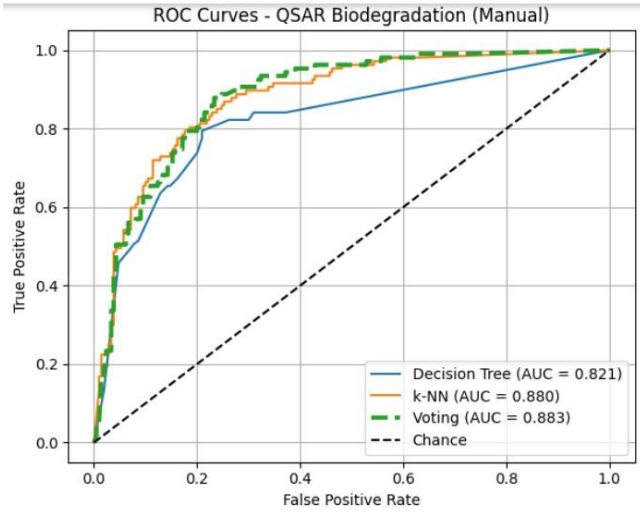
```
k-NN:
  Accuracy: 1.0000
  Precision: 1.0000
  Recall: 1.0000
  F1-Score: 1.0000
  ROC AUC: 1.0000

--- Built-in Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.9976, Precision: 1.0000
  Recall: 0.9945, F1: 0.9973, AUC: 1.0000
```

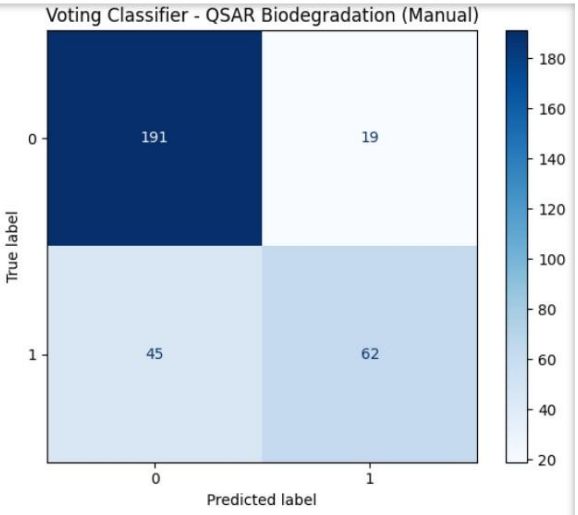



```
Completed processing for Banknote Authentication
================================================================================

################################################################################
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 ---
  Combination 1/180: AUC = 0.7902 {'feature_selection__k': 5, 'classifier__max_depth': 3, 'classifier__min_samples_split': 2, 'classifier__min_sample
s_leaf': 1}
  Combination 2/180: AUC = 0.7902 {'feature_selection__k': 5, 'classifier__max_depth': 3, 'classifier__min_samples_split': 2, 'classifier__min_sample
s_leaf': 2}
  Combination 3/180: AUC = 0.7938 {'feature_selection__k': 5, 'classifier__max_depth': 3, 'classifier__min_samples_split': 2, 'classifier__min_sample
s_leaf': 4}
  Combination 6/180: AUC = 0.7938 {'feature_selection__k': 5, 'classifier__max_depth': 3, 'classifier__min_samples_split': 5, 'classifier__min_sample
s_leaf': 4}
  Combination 9/180: AUC = 0.7938 {'feature_selection__k': 5, 'classifier__max_depth': 3, 'classifier__min_samples_split': 10, 'classifier__min_sampl
es_leaf': 4}
  Combination 10/180: AUC = 0.7979 {'feature_selection__k': 5, 'classifier__max_depth': 5, 'classifier__min_samples_split': 2, 'classifier__min_sampl
es_leaf': 1}
  Combination 11/180: AUC = 0.7996 {'feature_selection__k': 5, 'classifier__max_depth': 5, 'classifier__min_samples_split': 2, 'classifier__min_sampl
es_leaf': 2}
  Combination 12/180: AUC = 0.8016 {'feature_selection__k': 5, 'classifier__max_depth': 5, 'classifier__min_samples_split': 2, 'classifier__min_sampl
es_leaf': 4}
  Combination 15/180: AUC = 0.8016 {'feature_selection__k': 5, 'classifier__max_depth': 5, 'classifier__min_samples_split': 5, 'classifier__min_sampl
es_leaf': 4}
  Combination 17/180: AUC = 0.8021 {'feature_selection__k': 5, 'classifier__max_depth': 5, 'classifier__min_samples_split': 10, 'classifier__min_samp
les_leaf': 2}
```

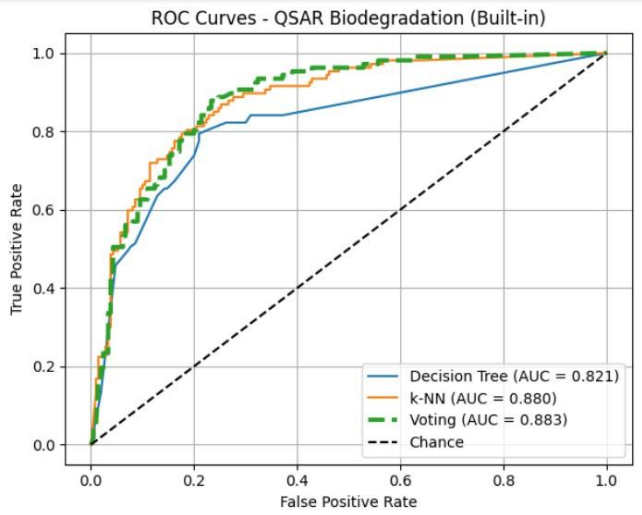
```
============================================================
EVALUATING MANUAL MODELS FOR QSAR BIODEGRADATION
============================================================

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.7823
  Precision: 0.6792
  Recall: 0.6729
  F1-Score: 0.6761
  ROC AUC: 0.8211

k-NN:
  Accuracy: 0.8202
  Precision: 0.7358
  Recall: 0.7290
  F1-Score: 0.7324
  ROC AUC: 0.8799

--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.7981, Precision: 0.7654
  Recall: 0.5794, F1: 0.6596, AUC: 0.8827
```

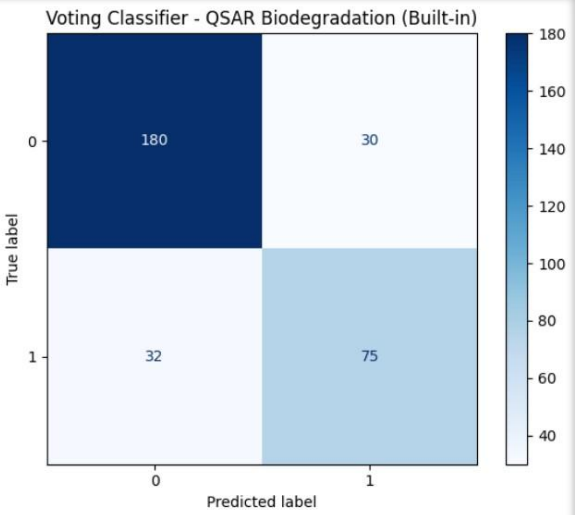




```
============================================================
RUNNING BUILT-IN GRID SEARCH FOR QSAR BIODEGRADATION
============================================================

--- GridSearchCV for Decision Tree ---
Fitting 5 folds for each of 180 candidates, totalling 900 fits
Best params for Decision Tree: {'classifier__max_depth': 7, 'classifier__min_samples_leaf': 4, 'classifier__min_samples_split': 10, 'feature_selectio
n__k': 20}
Best CV score: 0.8536

--- GridSearchCV for k-NN ---
Fitting 5 folds for each of 80 candidates, totalling 400 fits
Best params for k-NN: {'classifier__metric': 'manhattan', 'classifier__n_neighbors': 11, 'classifier__weights': 'distance', 'feature_selection__k': 2
0}
Best CV score: 0.8977

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

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.7823
  Precision: 0.6792
  Recall: 0.6729
  F1-Score: 0.6761
  ROC AUC: 0.8211
```

```
k-NN:
  Accuracy: 0.8202
  Precision: 0.7358
  Recall: 0.7290
  F1-Score: 0.7324
  ROC AUC: 0.8799

--- Built-in Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.8044, Precision: 0.7143
  Recall: 0.7009, F1: 0.7075, AUC: 0.8827
```






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

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

6. Conclusion

This lab provided a comprehensive exploration of hyperparameter tuning and model selection in machine learning. We gained a foundational understanding of the process through manual grid search and observed the enhanced efficiency offered by
GridSearchCV. Our analysis of diverse datasets allowed us to discern the relative strengths and weaknesses of various classifiers. The application of a voting classifier highlighted the significant performance gains achievable through ensemble methods. Ultimately, this lab solidified our grasp of both the theoretical principles and practical steps involved in constructing a robust machine learning pipeline.