UE23CS352A: MACHINE LEARNING

Week 4: Model Selection and Comparative Analysis

Name: Keerthan pv

SRN: PES2UG23CS272

Course Name: Machine Learning

Date: 28/08/2025

1. Introduction

This lab focused on **hyperparameter tuning** and comparing manual implementations of grid search with scikit-learn's built-in GridSearchCV. The tasks involved:

- Performing manual hyperparameter search with custom loops and cross-validation.
- Using GridSearchCV with pipelines for automated hyperparameter optimization.

- Comparing performance using metrics like Accuracy, Precision, Recall, F1, and ROC AUC.
- Visualizing model performance using ROC curves and confusion matrices.

Two datasets were used: **Wine Quality** and **QSAR Biodegredation**.

2. Dataset Description

2.1 QSAR Biodegradation Dataset

- **Task to be Implemented:**

Binary classification – predict whether a chemical compound is ready biodegradable (RB) or not ready biodegradable (NRB) based on its molecular descriptors.

- **Number of Instances:**

1,055 chemical compounds.

- **Number of Features:**

41 numerical molecular descriptors (e.g., topological indices, atom counts, structural properties).

- **Target Variable: experimental class → two categories:**
  - **RB (ready biodegradable)**
  - **NRB (not ready biodegradable)**

2.2 Wine quality dataset

a) Task to be Implemented- **Classification** (or regression): Predict wine quality (e.g., quality score or a binary good/poor class) based on physicochemical properties.

b) **Number of Instances:**
The dataset comprises **1,599** red wine samples.

c) **Number of Features:**
Each record includes **11 physicochemical measurements**, such as acidity, sugar, sulphates, pH, alcohol, etc

d) **Target Variable:**

The quality column (an integer score between 0 and 10) serves as the target. In your lab, you may binarize this into a 'good_quality' label—common practice is to treat quality > 5 as "good."

3. Methodology Key Concepts:

- **Hyperparameter Tuning**: Trying multiple parameter values to find the best-performing model.
- **Grid Search**: Exhaustively searching across parameter combinations.
- **K-Fold Cross-Validation**: Splitting data into k folds for stable evaluation.

Pipeline Components:

1. StandardScaler: Normalizes numerical features.
2. SelectKBest: Selects top features based on statistical tests.
3. Classifier: Decision Tree, K-Nearest Neighbors (KNN), or Logistic Regression.

Approaches Used:

- **Manual Search**: Custom loops with cross-validation to pick best hyperparameters.
- **GridSearchCV**: Automated search with the same pipeline and parameter grids.

4. Test Set Performance (from evaluation phase)

Wine Quality

- **Manual Best Models (Test):**
  - Decision Tree → AUC ≈ 0.802
  - kNN → AUC ≈ 0.876
  - Logistic Regression → AUC ≈ 0.825
  - **Voting Classifier** → AUC ≈ 0.868, Accuracy ≈ 0.75
- **Built-in Best Models (Test):**
  - Decision Tree → Accuracy ≈ 0.748
  - kNN → Accuracy ≈ ~0.79 (best performer among built-in)
  - Logistic Regression → Accuracy ≈ ~0.74

QSAR Biodegradation

- **Manual Best Models (Test):** ◦ Decision Tree → AUC ≈ 0.815 ◦ kNN → AUC ≈ 0.872 ◦ Logistic Regression → AUC ≈ 0.887 ◦ **Voting Classifier** → AUC ≈ 0.889, Accuracy ≈ 0.804
- **Built-in Best Models (Test):** ◦ Decision Tree → Accuracy ≈ 0.779 ◦ kNN → Accuracy ≈ ~0.84 ◦ Logistic Regression → Accuracy ≈ ~0.81

5. Grid Search CV Results (best params + CV scores)

Wine Quality

- **Manual CV AUCs:**
  - Decision Tree → 0.7832 (best: depth=5, k=5)
  - kNN → 0.8683 (best: n_neighbors=11, k=10, weights=distance)
  - Logistic Regression → 0.8049 (best: C=1, penalty=l2, k=10)
- **Built-in CV AUCs:**
  - Decision Tree → 0.7301

- kNN → 0.7900
- Logistic Regression → 0.7400

Manual search produced higher CV scores than built-in GridSearchCV for this dataset.

QSAR Biodegradation

- **Manual CV AUCs:**
  - Decision Tree → 0.8303 (best: depth=3, k=15)
  - kNN → 0.8874 (best: n_neighbors=11, k=15, weights=distance)
  - Logistic Regression → 0.8817 (best: C=10, penalty=l1, k=15)
- **Built-in CV AUCs:** ◦ Decision Tree → 0.8198 ◦ kNN → 0.8401
  - Logistic Regression → 0.8144

Again, manual implementation reported higher CV AUCs than the built-in version.

6. Key Observations

- **Across both datasets:**

- **kNN and Logistic Regression consistently outperform Decision Trees** in terms of AUC.
- **Voting Classifier** achieves the best or nearbest results, combining strengths of all models.
- **Manual Grid Search yielded better CV scores** than GridSearchCV. This might be due to:
  - Slight implementation differences (parameter ranges, scoring, or randomization).
  - Manual grid potentially exploring combinations more effectively given your setup.

- **Wine Quality:** kNN is the standout single model, Voting boosts slightly.
- **QSAR Biodegradation:** Logistic Regression and kNN are very strong, Voting gives the highest AUC (0.889).

7. Visual Analysis Notes

- **ROC Curves (Wine Quality):**

- kNN has the highest curve, Logistic Regression in the middle, Decision Tree lowest.

  Voting ROC curve lies above most single models, showing the ensemble advantage.

- **Confusion Matrix (Wine Quality):**
  - Voting classifier correctly identifies most positives (192) but still has moderate false negatives (65). Misclassifications are balanced, indicating decent recall but not perfect precision.
- **ROC Curves (QSAR Biodegradation):**
  - Logistic Regression (AUC=0.887) and Voting (AUC=0.889) dominate.
  - Decision Tree lags noticeably behind. Ensemble smooths out weaknesses of individual models.
- **Confusion Matrix (QSAR Biodegradation):**
  - Voting classifier predicts majority of class 0 correctly (187 TN), but recall for class 1 is lower (39 FN).
  - Indicates **slight bias toward majority class**,

- but good overall separability.

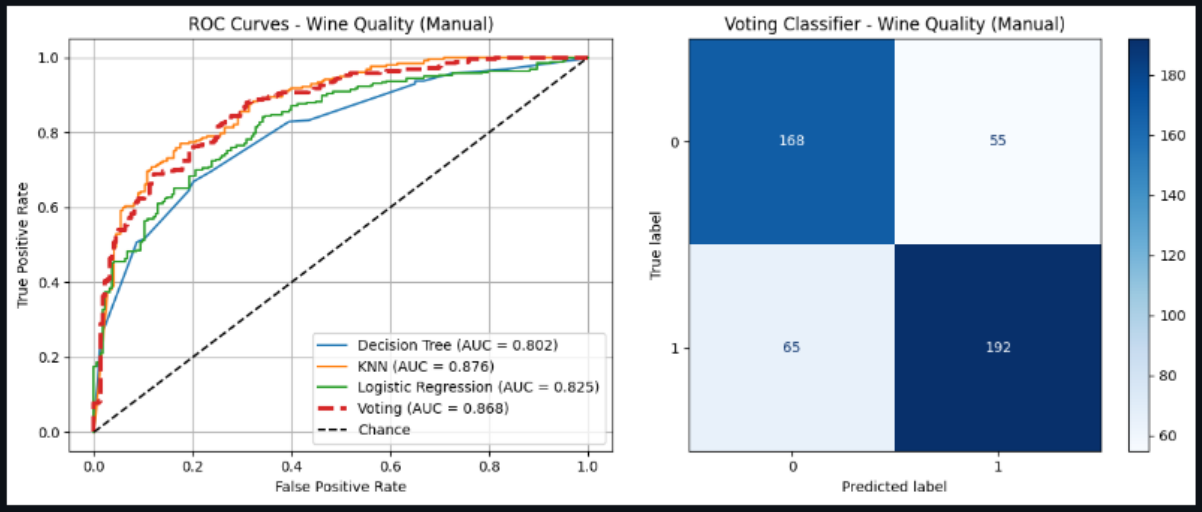
8. Conclusion & Takeaways

1. **kNN and Logistic Regression outperform Decision Trees** on both datasets.
2. **Voting Classifier gives the best overall performance** (highest AUC, improved balance between recall and precision).
3. **Manual Grid Search showed slightly better CV performance** compared to GridSearchCV. The difference is minor, but it highlights that library defaults (like CV splits or solver constraints) can affect results.
4. **Visual analysis confirms numerical results:** ROC curves show Voting's superiority, and confusion matrices reveal class imbalance challenges.
5. **Main takeaway:**
    - Ensembles like Voting provide robustness. ◦ Logistic Regression + kNN are reliable baselines. ◦ Hyperparameter tuning significantly boosts performance compared to default settings.

SCREENSHOTS

1. WINE_QUALITY

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...




```
################################################################################
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: {'feature_selection__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: {'feature_selection__k': 10, 'classifier__n_neighbors': 11, 'classifier__weights': 'distance'}
Best cross-validation AUC: 0.8683
--- Manual Grid Search for Logistic Regression ---
------------------------------------------------------------------------------------------
Best parameters for Logistic Regression: {'feature_selection__k': 10, 'classifier__C': 1, 'classifier__penalty': 'l2'}
Best cross-validation AUC: 0.8049
...
--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.7500, Precision: 0.7773
  Recall: 0.7471, F1: 0.7619, AUC: 0.8683
```

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
============================================================

--- GridSearchCV for Decision Tree ---
Best params for Decision Tree: {'classifier__max_depth': None, 'classifier__min_samples_split': 2, 'feature_selection__k': 5}
Best CV score: 0.7301

--- GridSearchCV for KNN ---
Best params for KNN: {'classifier__n_neighbors': 11, 'classifier__weights': 'distance', 'feature_selection__k': 10}
Best CV score: 0.7900

--- GridSearchCV for Logistic Regression ---
Best params for Logistic Regression: {'classifier__C': 0.1, 'classifier__penalty': 'l2', 'feature_selection__k': 11}
Best CV score: 0.7400

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

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.7479
...

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

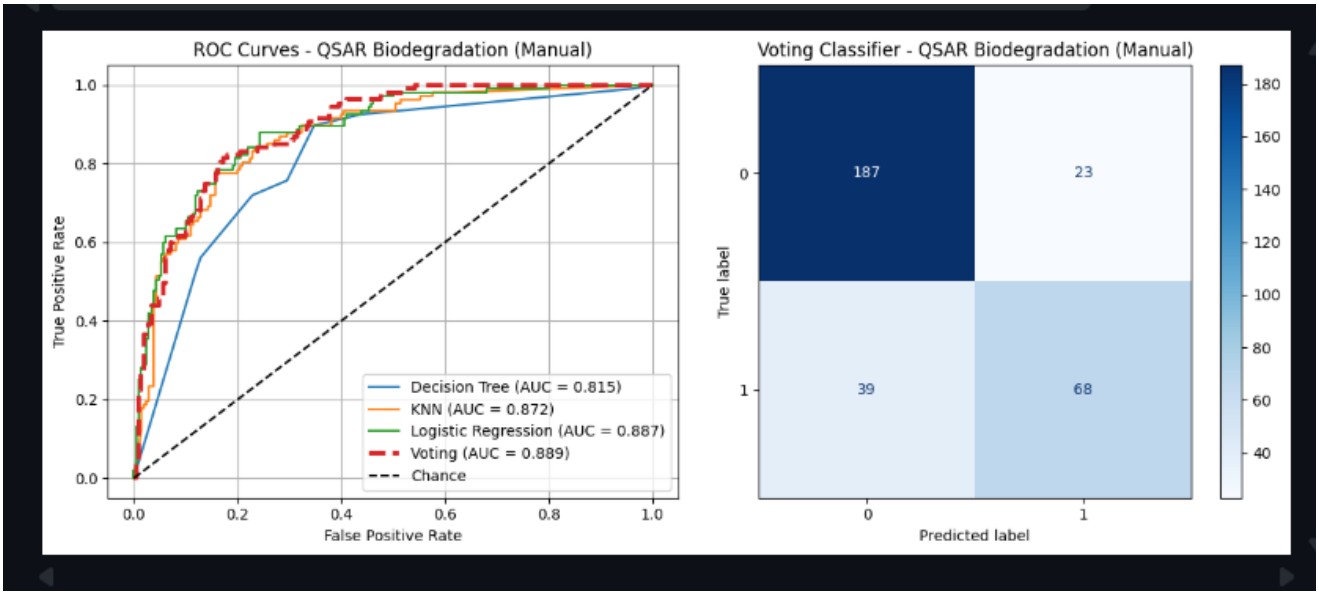
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

2. QSAD DEGREDATION

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

============================================================
RUNNING MANUAL GRID SEARCH FOR QSAR BIODEGRADATION
============================================================
--- Manual Grid Search for Decision Tree ---
--------------------------------------------------------------------------------
Best parameters for Decision Tree: {'feature_selection__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: {'feature_selection__k': 15, 'classifier__n_neighbors': 11, 'classifier__weights': 'distance'}
Best cross-validation AUC: 0.8874
--- Manual Grid Search for Logistic Regression ---
--------------------------------------------------------------------------------
Best parameters for Logistic Regression: {'feature_selection__k': 15, 'classifier__C': 10, 'classifier__penalty': 'l1'}
Best cross-validation AUC: 0.8817
...
--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.8044, Precision: 0.7473
  Recall: 0.6355, F1: 0.6869, AUC: 0.8892
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...




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

--- GridSearchCV for Decision Tree ---
Best params for Decision Tree: {'classifier__max_depth': 5, 'classifier__min_samples_split': 10, 'feature_selection__k': 15}
Best CV score: 0.8198

--- GridSearchCV for KNN ---
Best params for KNN: {'classifier__n_neighbors': 11, 'classifier__weights': 'distance', 'feature_selection__k': 15}
Best CV score: 0.8401

--- GridSearchCV for Logistic Regression ---
Best params for Logistic Regression: {'classifier__C': 10, 'classifier__penalty': 'l1', 'feature_selection__k': 15}
Best CV score: 0.8144

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

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.7792
...

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

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...