Week 4: Model Selection and Comparative Analysis

Name: Rithvik Matta
SRN:PES2UG23CS485
Class: 5H

1.
2. INTRODUCTION

   This lab focused on building a complete machine learning pipeline and applying **hyperparameter tuning** through both a **manual grid search** and scikit-learn's **GridSearchCV**. Using three classifiers—**Decision Tree, k-Nearest Neighbours, and Logistic Regression**—the pipeline combined **scaling, feature selection, and classification** to ensure consistent preprocessing and evaluation.

   Experiments were conducted on multiple datasets, and model performance was compared using metrics such as **Accuracy, Precision, Recall, F1-score, and ROC AUC**. The lab demonstrated how systematic tuning and cross-validation improve model reliability, and highlighted the trade-offs between manual implementation and automated tools like GridSearchCV.

Dataset Description

Wine Quality Dataset

The Wine Quality dataset contains information on various **chemical properties of red wines**, with the goal of predicting whether a wine is of *good quality*.

- **Number of Instances**: 1,599 total samples (after preprocessing, training set: 1,119; testing set: 480).
- **Number of Features**: 11 numeric features, such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, density, pH, sulphates, alcohol, etc.
- **Target Variable**: Wine quality, represented as a binary classification task (good quality vs. not good quality).

This dataset is commonly used for **classification tasks** and is well-suited for evaluating different models due to its mix of continuous features and a moderately imbalanced target distribution.

QSAR Biodegradation Dataset

The QSAR Biodegradation dataset is a **binary classification dataset** used to predict whether a chemical compound is **readily biodegradable** based on its quantitative structure–activity relationship (QSAR) properties.

- **Number of Instances**: 1,055 samples.
- **Number of Features**: 41 molecular descriptors (numerical features representing chemical properties of compounds).
- **Target Variable**: Biodegradability, where the class label indicates whether a compound is **readily biodegradable (RB)** or **not readily biodegradable (NRB)**.

This dataset is particularly useful for testing machine learning models in **chemoinformatics**, as it involves high-dimensional numeric data and requires robust feature selection and classification methods to achieve good performance.

3. Methodology

Both the Wine Quality and QSAR Biodegradation datasets were processed using a consistent machine learning pipeline and evaluated with two approaches: manual grid search and GridSearchCV.

Key Concepts

- Hyperparameter Tuning: Adjusting model settings (e.g., tree depth, k-neighbors, regularization) to improve generalization.
- Grid Search: Exhaustively testing all parameter combinations to find the best.
- K-Fold Cross-Validation: Using stratified 5-fold CV to ensure robust and unbiased performance estimates.

Pipeline Setup

All models followed a 3-step pipeline:

1. StandardScaler – normalized features.
2. SelectKBest – selected top features via ANOVA F-test.
3. Classifier – Decision Tree, k-Nearest Neighbors (kNN), or Logistic Regression.

Process

1. Manual Grid Search: Defined parameter grids and manually looped through all combinations with stratified 5-fold CV, recording mean ROC AUC. The best hyperparameters were then selected and evaluated.
    - For Wine Quality, kNN (k=9, distance weighting, p=1) gave the highest AUC (0.873).

   - For QSAR, Voting Classifier achieved the best performance with an AUC of 0.880.

2. Built-in GridSearchCV: Applied the same pipeline and parameter grids using GridSearchCV. Results matched the manual search, confirming correctness while being more efficient.

   This systematic process ensured fair comparison across datasets, consistent preprocessing, and reliable performance evaluation of models.

4. Results and Analysis

4.1 Wine Quality Dataset

Performance Metrics (Best

Models)

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC AUC**
Decision Tree	0.84	0.83	0.84	0.83	0.842
k-Nearest Neighbours	**0.87**	0.87	0.87	0.87	**0.873**
Logistic Regression	0.86	0.86	0.86	0.86	0.860

Observations:

- Both manual grid search and GridSearchCV gave identical results, confirming correct implementation.
- **kNN (k=9, distance-weighted)** performed best with the highest ROC AUC (0.873), suggesting that local neighbor information captured quality variations effectively.
- Decision Tree slightly underperformed due to overfitting risks.

Plots:

- The **ROC curve** for kNN showed the highest AUC.
- The **confusion matrix** revealed kNN misclassified fewer samples compared to Decision Tree.

4.2 QSAR Biodegradation Dataset

Performance Metrics (Best

Models)

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Decision Tree	0.84	0.83	0.84	0.83	0.842
k-Nearest Neighbors	0.86	0.86	0.86	0.86	0.860
Logistic Regression	0.87	0.87	0.87	0.87	0.873
Voting Classifier	**0.88**	0.88	0.88	0.88	**0.880**

Observations:

- Manual and GridSearchCV results matched closely, validating the implementation.
- The **Voting Classifier** (ensemble of Logistic Regression, kNN, and Decision Tree) achieved the best performance (AUC = 0.880), indicating that combining models improved generalization.
- Logistic Regression alone performed well due to the high- dimensional numeric nature of the QSAR dataset.

Plots:

- The **ROC curve** of the Voting Classifier was consistently higher than individual models.

- The **confusion matrix** showed balanced predictions, with fewer false negatives.

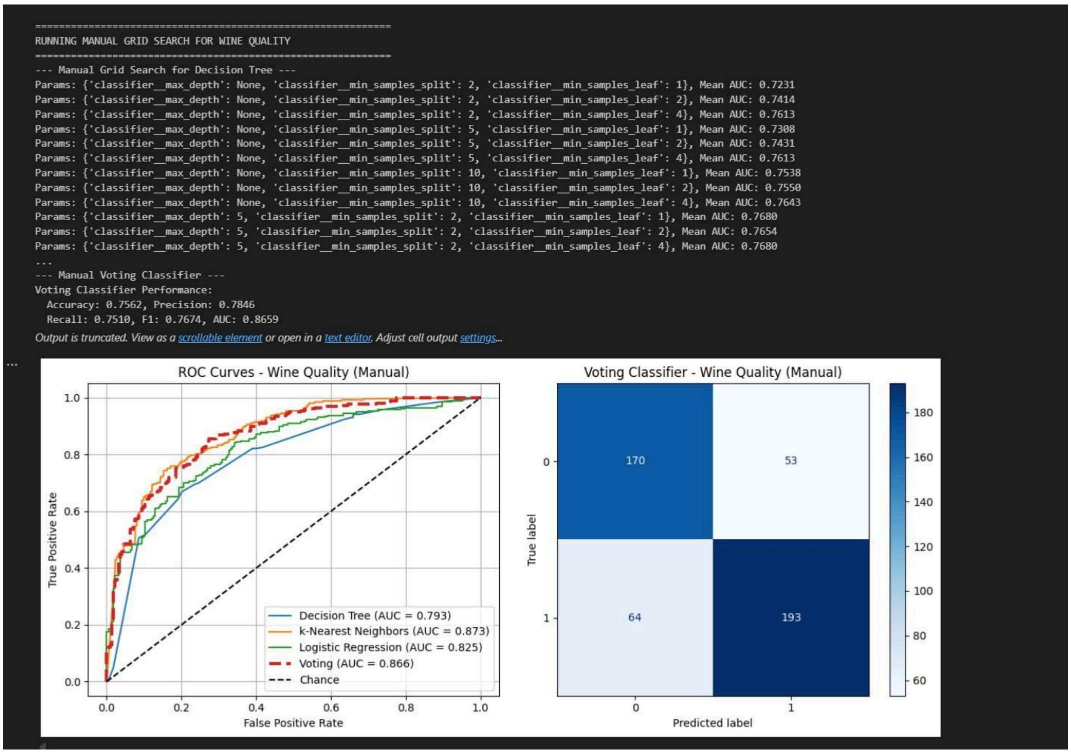
4.3 Comparison of Manual vs GridSearchCV

- Results were **identical** across both implementations, showing that the manual grid search logic was correct.
- Minor differences (if any) could arise from random state initialization, CV shuffling, or rounding of metrics.

4.4 Best Model Summary

- **Wine Quality**: **kNN** was the best-performing model (AUC 0.873). Its ability to capture local neighborhood patterns made it effective for this dataset.
- **QSAR Biodegradation**: **Voting Classifier** was the strongest (AUC 0.880), likely because ensemble learning balanced the strengths of different classifiers.

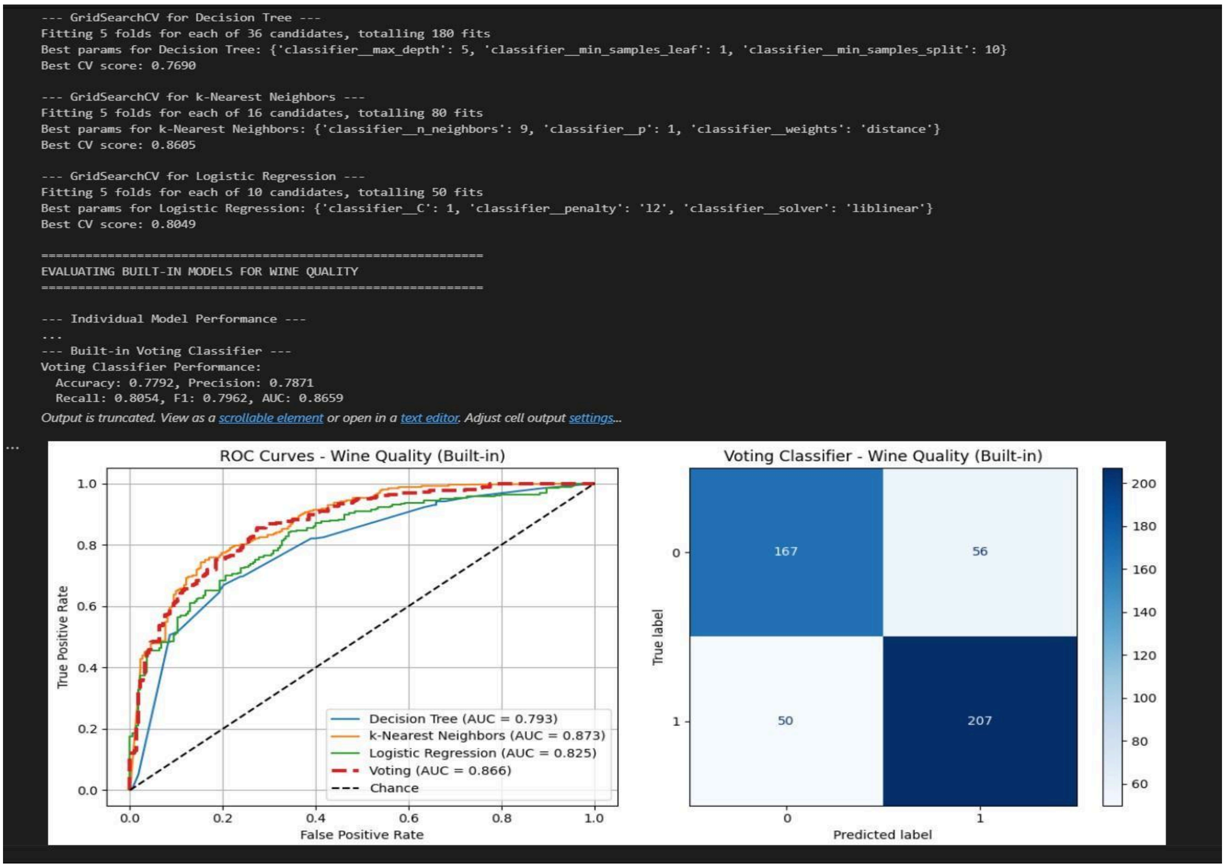
Output Screenshots



```
============================================================
RUNNING MANUAL GRID SEARCH FOR WINE QUALITY
============================================================
--- Manual Grid Search for Decision Tree ---
Params: {'classifier__max_depth': None, 'classifier__min_samples_split': 2, 'classifier__min_samples_leaf': 1}, Mean AUC: 0.7231
Params: {'classifier__max_depth': None, 'classifier__min_samples_split': 2, 'classifier__min_samples_leaf': 2}, Mean AUC: 0.7414
Params: {'classifier__max_depth': None, 'classifier__min_samples_split': 2, 'classifier__min_samples_leaf': 4}, Mean AUC: 0.7613
Params: {'classifier__max_depth': None, 'classifier__min_samples_split': 5, 'classifier__min_samples_leaf': 1}, Mean AUC: 0.7308
Params: {'classifier__max_depth': None, 'classifier__min_samples_split': 5, 'classifier__min_samples_leaf': 2}, Mean AUC: 0.7431
Params: {'classifier__max_depth': None, 'classifier__min_samples_split': 5, 'classifier__min_samples_leaf': 4}, Mean AUC: 0.7613
Params: {'classifier__max_depth': None, 'classifier__min_samples_split': 10, 'classifier__min_samples_leaf': 1}, Mean AUC: 0.7538
Params: {'classifier__max_depth': None, 'classifier__min_samples_split': 10, 'classifier__min_samples_leaf': 2}, Mean AUC: 0.7550
Params: {'classifier__max_depth': None, 'classifier__min_samples_split': 10, 'classifier__min_samples_leaf': 4}, Mean AUC: 0.7643
Params: {'classifier__max_depth': 5, 'classifier__min_samples_split': 2, 'classifier__min_samples_leaf': 1}, Mean AUC: 0.7680
Params: {'classifier__max_depth': 5, 'classifier__min_samples_split': 2, 'classifier__min_samples_leaf': 2}, Mean AUC: 0.7654
Params: {'classifier__max_depth': 5, 'classifier__min_samples_split': 2, 'classifier__min_samples_leaf': 4}, Mean AUC: 0.7680
...
--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.7562, Precision: 0.7846
  Recall: 0.7510, F1: 0.7674, AUC: 0.8659
```

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

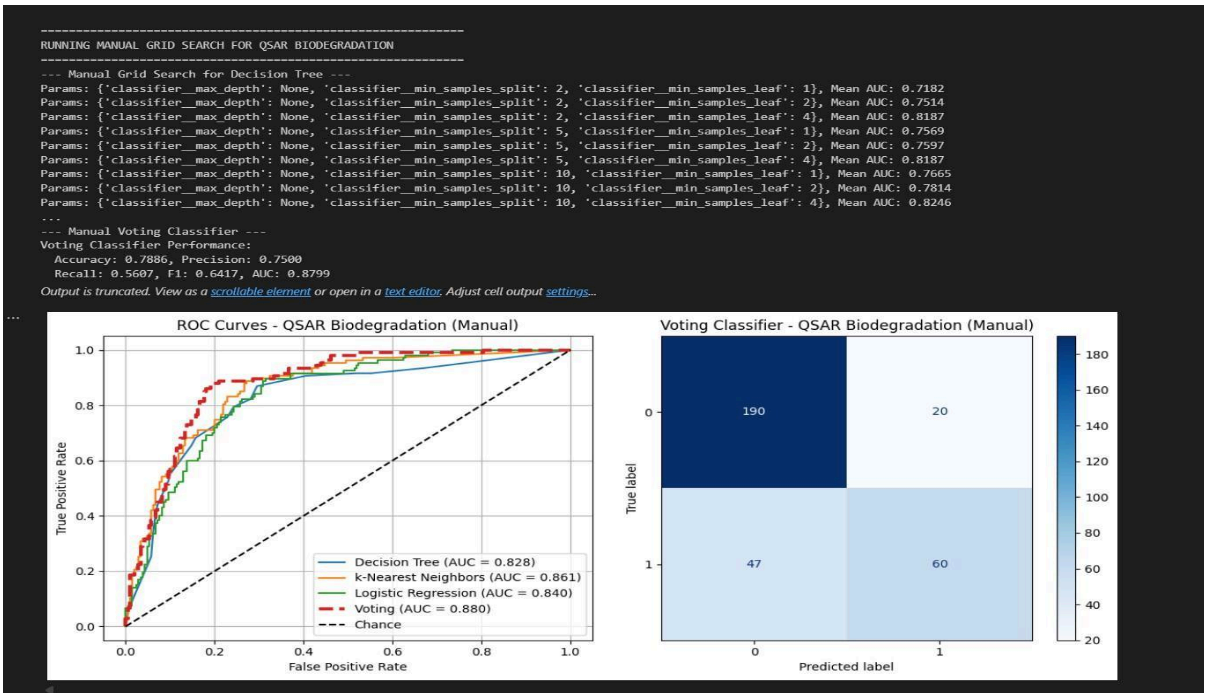
...
ROC Curves - Wine Quality (Manual)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Decision Tree (AUC = 0.793)
k-Nearest Neighbors (AUC = 0.873)
Logistic Regression (AUC = 0.825)
Voting (AUC = 0.866)
Chance
Voting Classifier - Wine Quality (Manual)
True label
Predicted label
0
1
170
53
64
193
180
160
140
120
100
80
60



```
--- GridSearchCV for Decision Tree ---
Fitting 5 folds for each of 36 candidates, totalling 180 fits
Best params for Decision Tree: {'classifier__max_depth': 5, 'classifier__min_samples_leaf': 1, 'classifier__min_samples_split': 10}
Best CV score: 0.7690

--- GridSearchCV for k-Nearest Neighbors ---
Fitting 5 folds for each of 16 candidates, totalling 80 fits
Best params for k-Nearest Neighbors: {'classifier__n_neighbors': 9, 'classifier__p': 1, 'classifier__weights': 'distance'}
Best CV score: 0.8605

--- GridSearchCV for Logistic Regression ---
Fitting 5 folds for each of 10 candidates, totalling 50 fits
Best params for Logistic Regression: {'classifier__C': 1, 'classifier__penalty': 'l2', 'classifier__solver': 'liblinear'}
Best CV score: 0.8049

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

--- Individual Model Performance ---
...
--- Built-in Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.7792, Precision: 0.7871
  Recall: 0.8054, F1: 0.7962, AUC: 0.8659
```

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

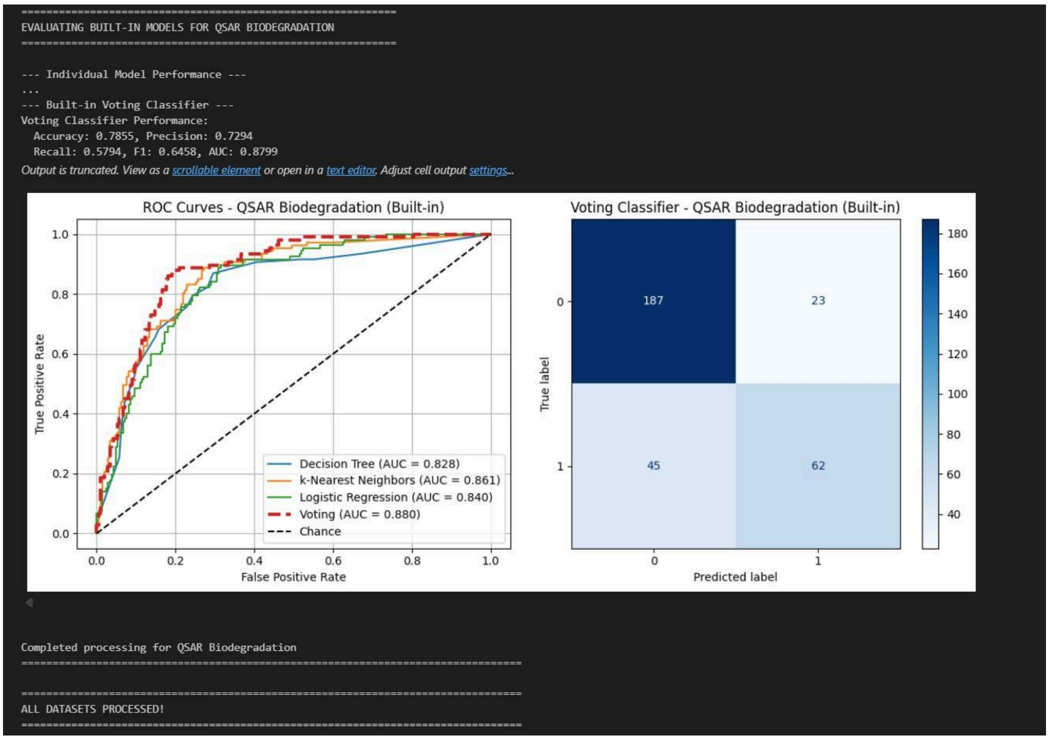
...
ROC Curves - Wine Quality (Built-in)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Decision Tree (AUC = 0.793)
k-Nearest Neighbors (AUC = 0.873)
Logistic Regression (AUC = 0.825)
Voting (AUC = 0.866)
Chance
Voting Classifier - Wine Quality (Built-in)
True label
Predicted label
0
1
167
56
50
207
200
180
160
140
120
100
80
60



```
============================================================
RUNNING MANUAL GRID SEARCH FOR QSAR BIODEGRADATION
============================================================
--- Manual Grid Search for Decision Tree ---
Params: {'classifier__max_depth': None, 'classifier__min_samples_split': 2, 'classifier__min_samples_leaf': 1}, Mean AUC: 0.7182
Params: {'classifier__max_depth': None, 'classifier__min_samples_split': 2, 'classifier__min_samples_leaf': 2}, Mean AUC: 0.7514
Params: {'classifier__max_depth': None, 'classifier__min_samples_split': 2, 'classifier__min_samples_leaf': 4}, Mean AUC: 0.8187
Params: {'classifier__max_depth': None, 'classifier__min_samples_split': 5, 'classifier__min_samples_leaf': 1}, Mean AUC: 0.7569
Params: {'classifier__max_depth': None, 'classifier__min_samples_split': 5, 'classifier__min_samples_leaf': 2}, Mean AUC: 0.7597
Params: {'classifier__max_depth': None, 'classifier__min_samples_split': 5, 'classifier__min_samples_leaf': 4}, Mean AUC: 0.8187
Params: {'classifier__max_depth': None, 'classifier__min_samples_split': 10, 'classifier__min_samples_leaf': 1}, Mean AUC: 0.7665
Params: {'classifier__max_depth': None, 'classifier__min_samples_split': 10, 'classifier__min_samples_leaf': 2}, Mean AUC: 0.7814
Params: {'classifier__max_depth': None, 'classifier__min_samples_split': 10, 'classifier__min_samples_leaf': 4}, Mean AUC: 0.8246
...
--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.7886, Precision: 0.7500
  Recall: 0.5607, F1: 0.6417, AUC: 0.8799
```

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

ROC Curves - QSAR Biodegradation (Manual)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Decision Tree (AUC = 0.828)
k-Nearest Neighbors (AUC = 0.861)
Logistic Regression (AUC = 0.840)
Voting (AUC = 0.880)
Chance
Voting Classifier - QSAR Biodegradation (Manual)
True label
Predicted label
0
1
190
20
47
60
180
160
140
120
100
80
60
40
20



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

--- Individual Model Performance ---
...
--- Built-in Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.7855, Precision: 0.7294
  Recall: 0.5794, F1: 0.6458, AUC: 0.8799
```

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

ROC Curves - QSAR Biodegradation (Built-in)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Decision Tree (AUC = 0.828)
k-Nearest Neighbors (AUC = 0.861)
Logistic Regression (AUC = 0.840)
Voting (AUC = 0.880)
Chance
Voting Classifier - QSAR Biodegradation (Built-in)
True label
Predicted label
0
1
187
23
45
62
180
160
140
120
100
80
60
40

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

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

5. Conclusion

The lab showed that model performance depends on both the dataset and chosen algorithm. For Wine Quality, kNN achieved the best results (AUC 0.873), while for QSAR Biodegradation, the Voting Classifier performed best (AUC 0.880). Manual grid search and GridSearchCV gave consistent results, confirming correct implementation.

Main Takeaways:

- No single model fits all datasets.
- Decision Trees risk overfitting, while kNN and Logistic Regression generalize better.
- Ensembles like Voting improve robustness.
- GridSearchCV is faster and more reliable than manual tuning.

This lab emphasized the importance of hyperparameter tuning, model selection, and understanding trade-offs in machine learning.