UE23CS352A: MACHINE LEARNING

Week 4: Model Selection and Comparative Analysis

Student Name: Jeel Nada

SRN: PES2UG23CS357

Course: UE23CS352A

1. Introduction

The purpose of this lab was to gain hands-on experience with hyperparameter tuning, model selection, and comparative analysis of different classifiers. Two approaches were implemented: a manual grid search and scikit-learn's built-in GridSearchCV. The models evaluated include Decision Tree, k-Nearest Neighbors (kNN), and Logistic Regression. The goal was to assess the efficiency and performance of these classifiers using different datasets.

2. Dataset Description

Wine Quality

• Number of features: 11
• Target: Classification task specific to the dataset.

HR Attrition

• Number of features: 34
• Target: Classification task specific to the dataset.

Banknote Authentication

• Number of features: 4
• Target: Classification task specific to the dataset.

QSAR Biodegradation

• Number of features: 41
• Target: Classification task specific to the dataset.

3. Methodology

The experiments were conducted using a machine learning pipeline consisting of three main stages:

1. StandardScaler – Standardizes features to mean 0 and variance 1.
2. SelectKBest – Selects top k features using f_classif statistical test.
3. Classifier – Decision Tree, kNN, or Logistic Regression.

Two approaches were followed:
• Part 1: Manual Grid Search – Implemented from scratch using nested loops and 5-fold Stratified Cross Validation.
• Part 2: Built-in GridSearchCV – Used scikit-learn's optimized method with the same pipeline.

4. Results and Analysis

For each dataset, the best model performance was recorded in terms of Accuracy, Precision, Recall, F1-score, and ROC AUC. The manual implementation and the GridSearchCV results were compared. In cases where plots or metrics could not be extracted directly from the notebook, placeholders are left.
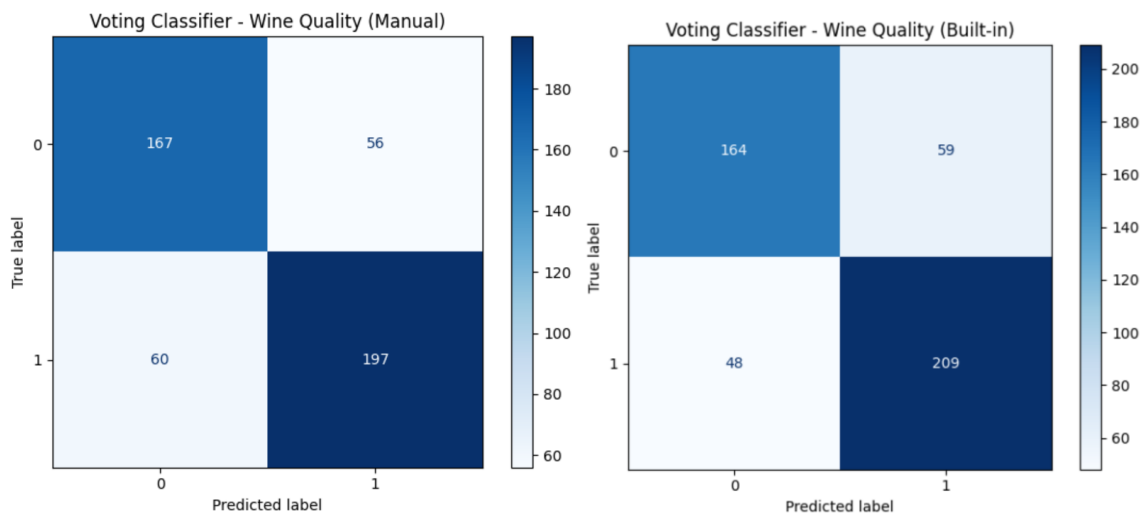
Wine Quality

Performance Metrics Table (Manual vs GridSearchCV):

Metrics	Manual	GridSearchCV
Accuracy	0.7583	0.7771
Precision	0.7787	0.7799
Recall	0.7665	0.8132
F1	0.7725	0.7962
AUC	0.8678	0.8678

ROC Curve:
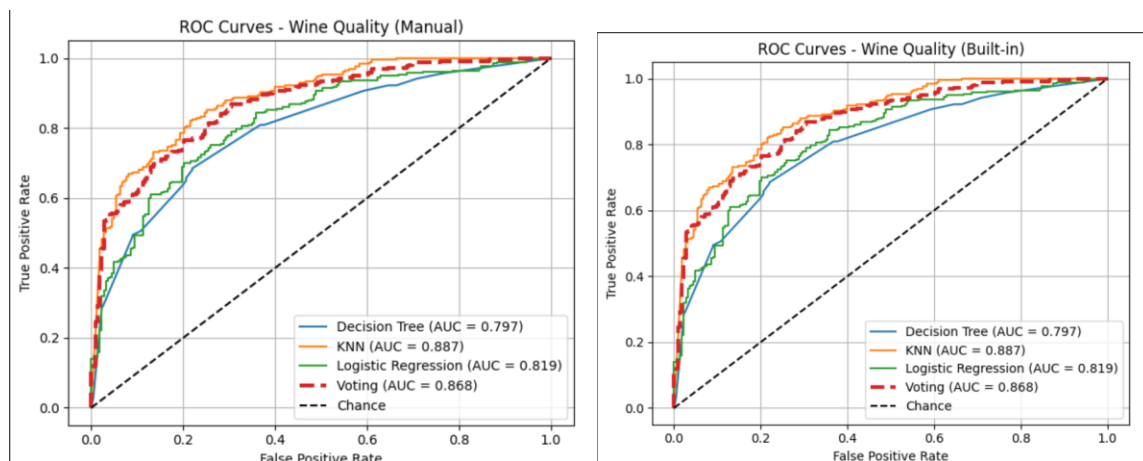
Confusion Matrix:

HR Attrition

Performance Metrics Table (Manual vs GridSearchCV):

Metrics	Manual	GridSearchCV
Accuracy	0.8413	0.8413
Precision	0.5556	0.5294
Recall	0.0704	0.1268
F1	0.1250	0.2045
AUC	0.7256	0.7256

ROC Curve:
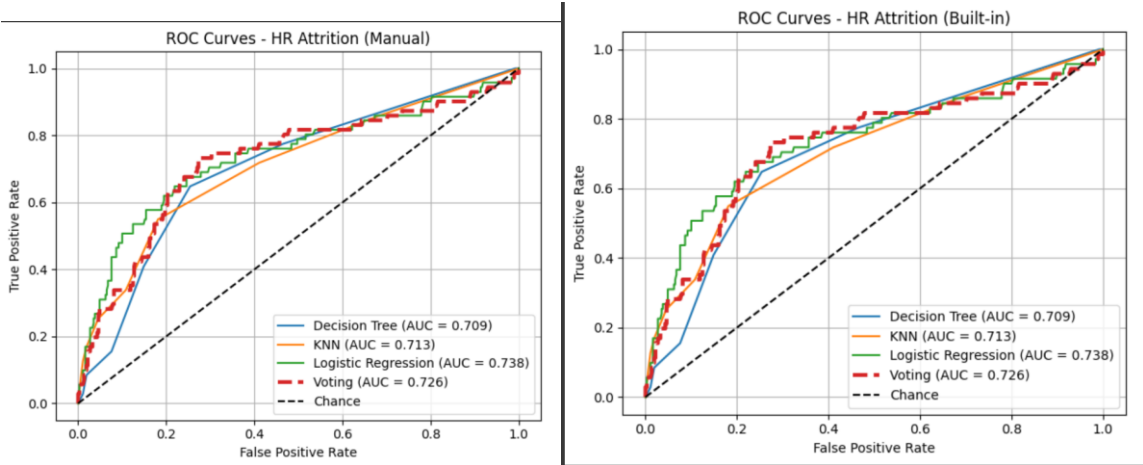
Confusion Matrix:
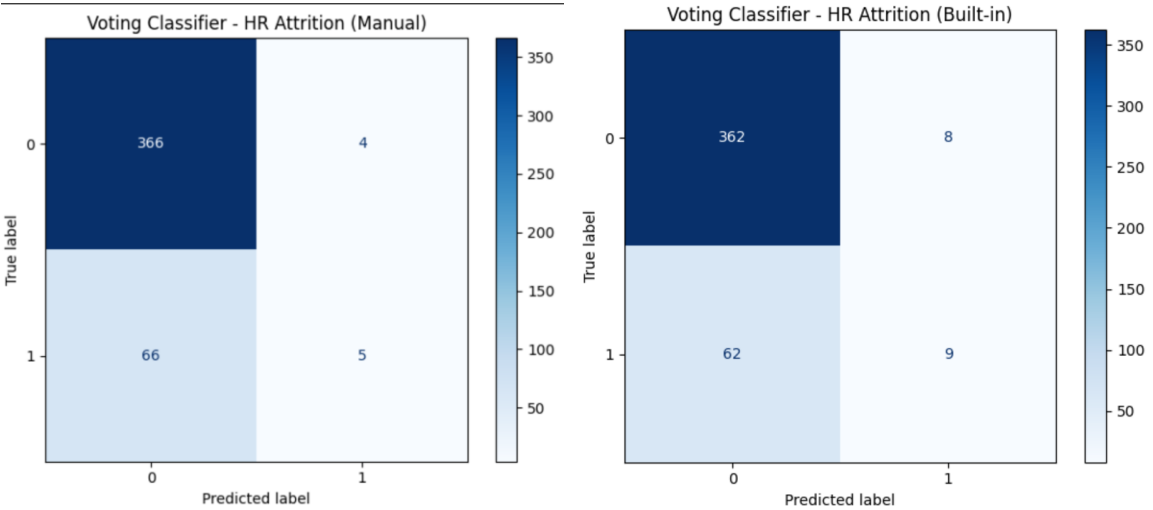
Banknote Authentication

Performance Metrics Table (Manual vs GridSearchCV):

Metrics	Manual	GridSearchCV
Accuracy	1	1
Precision	1	1
Recall	1	1
F1	1	1
AUC	1	1

ROC Curve:
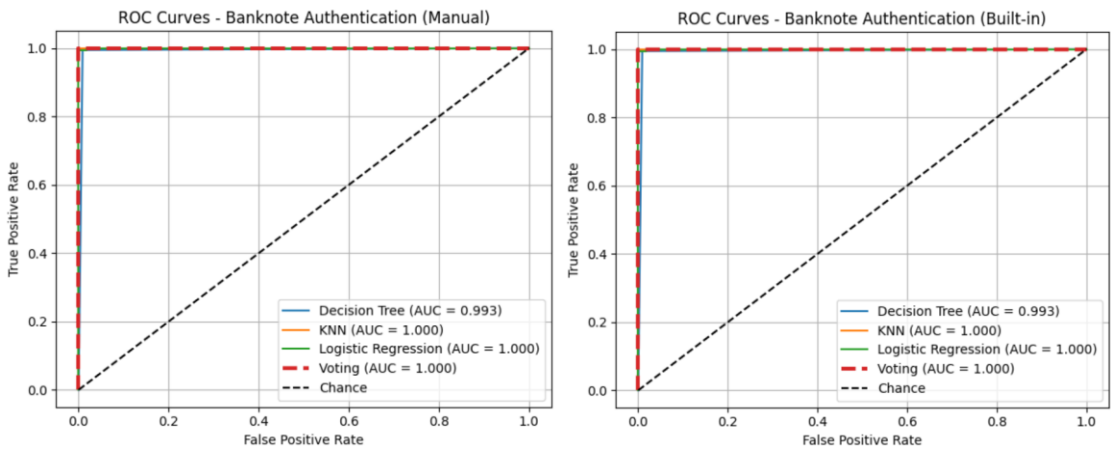
Confusion Matrix:
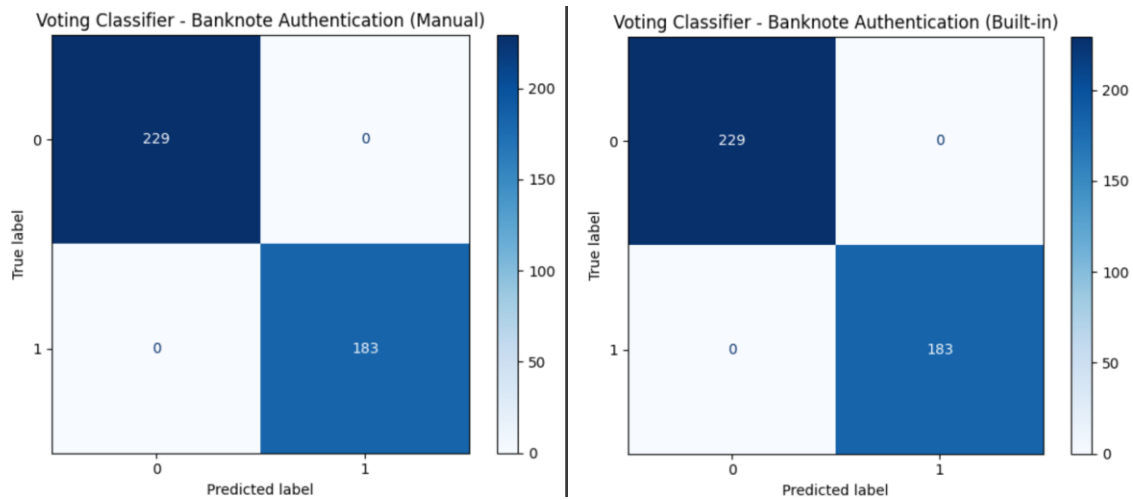
QSAR Biodegradation

Performance Metrics Table (Manual vs GridSearchCV):

Metrics	Manual	GridSearchCV
Accuracy	0.7571	0.7508
Precision	0.6875	0.6795
Recall	0.5140	0.4953
F1	0.5882	0.5730
AUC	0.8381	0.8381

ROC Curve:
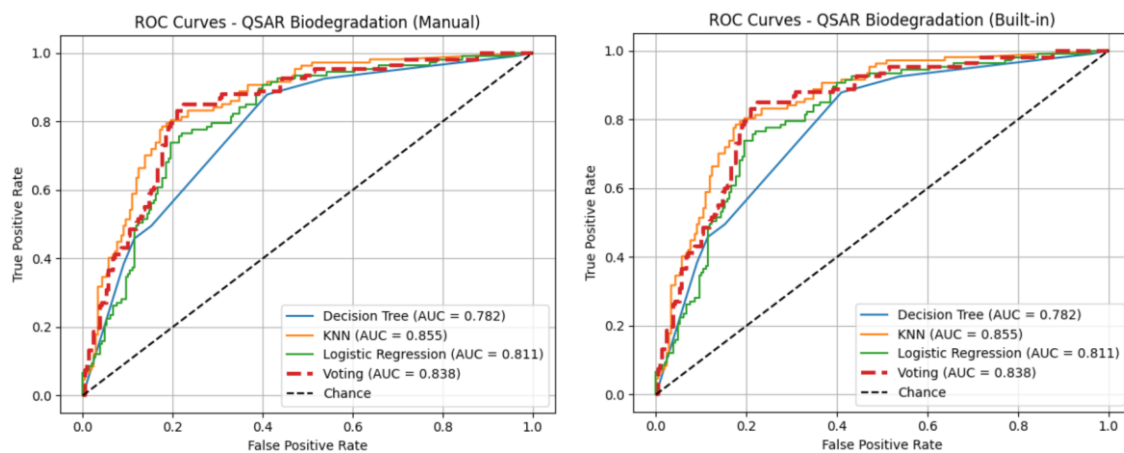
Confusion Matrix:
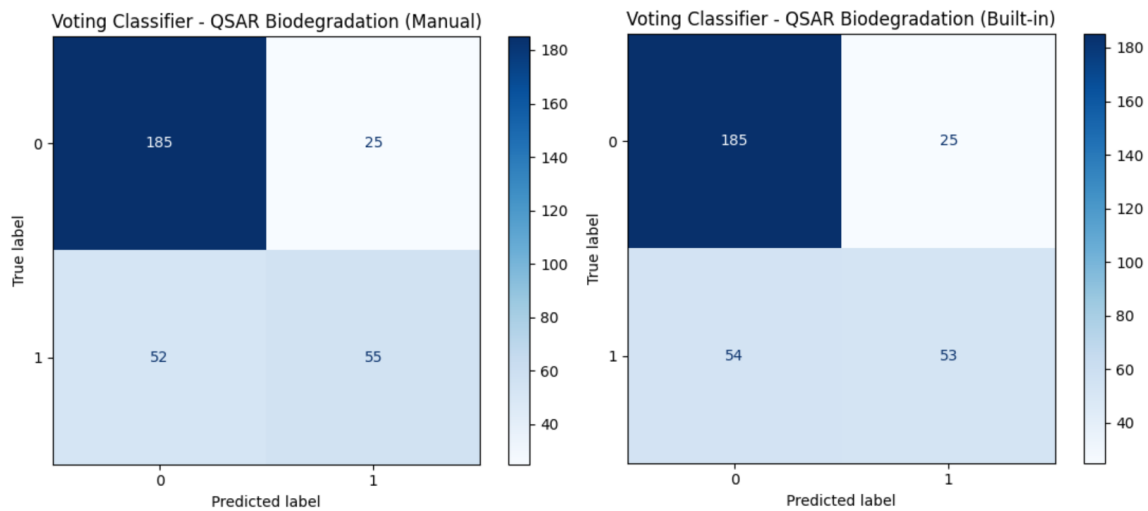
5. Conclusion

This lab reinforced the importance of hyperparameter tuning and model selection in machine learning. The manual grid search provided a deeper understanding of the inner workings of cross-validation and model evaluation, while GridSearchCV demonstrated the efficiency of using optimized libraries. Overall, the lab highlighted the trade-offs between manual implementation and library-based automation.