Week 4 Lab Report

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 Scikitlearn's GridSearchCV. Our objective was to compare the performance of Decision Tree, kNearest 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.

- **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**

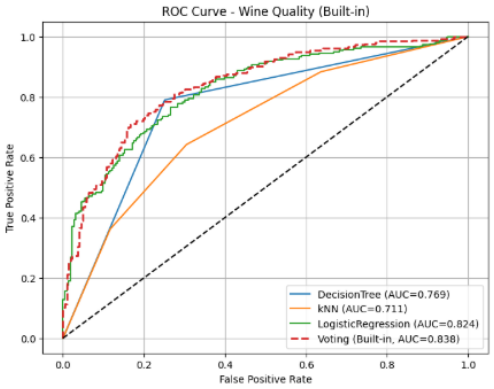
Built-in GridSearchCV Results:

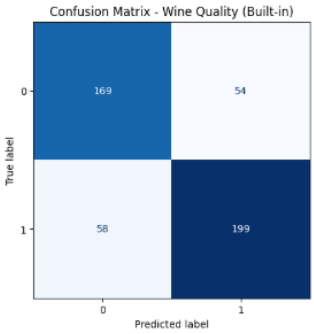
Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Decision Tree	**0.7271**	**0.7716**	**0.6965**	**0.7321**	**0.8025**
k-NN	**0.7812**	**0.7836**	**0.8171**	**0.8000**	**0.8589**
Logistic Regression	**0.7396**	**0.7619**	**0.7471**	**0.7544**	**0.8246**

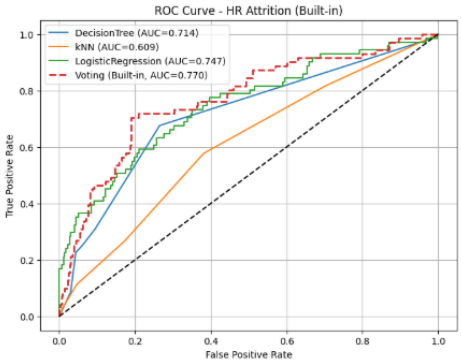
Visualizations (ROC curves and confusion matrices) should accompany these tables to better interpret model performance.

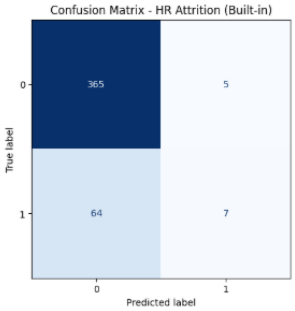
Additionally, results from manual and built-in methods should be compared. Minor differences may occur due to randomness or implementation details. Observations about which classifier performed best should also be included.

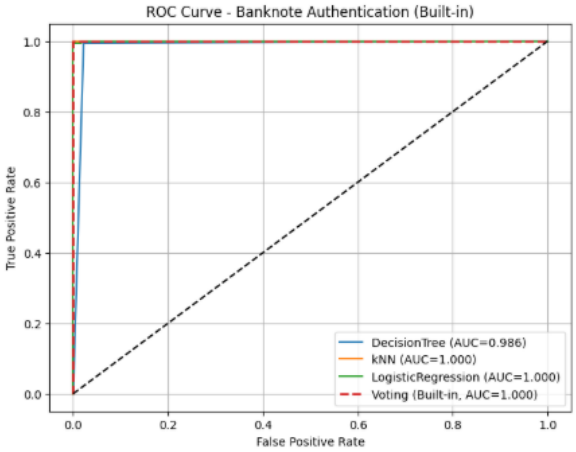
5. Screenshot

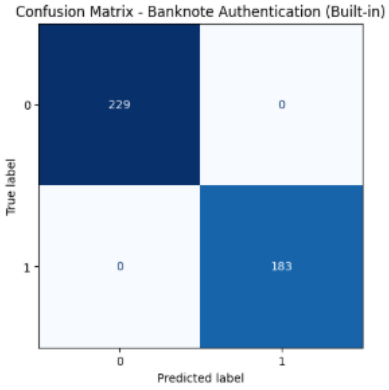


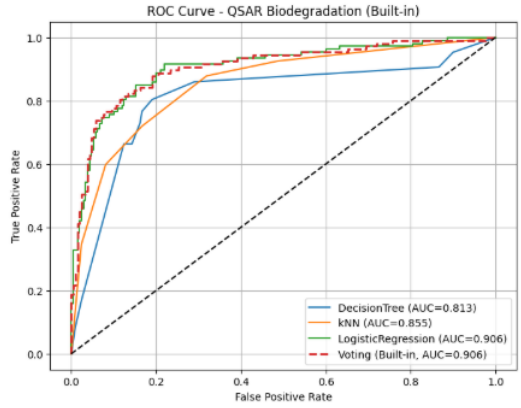


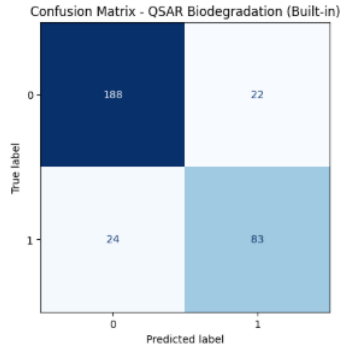













```
--------------------------------------------------------------------------------
ALL DATASETS PROCESSED!
--------------------------------------------------------------------------------
```

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.