Lab4: Model Selection

Name: Muskan Goenka | **SRN:** PES2UG23CS355 | **Section: F**

1. Introduction

The aim of this lab is to explore hyperparameter tuning for different classification models using both **manual grid search** and **Scikit-learn's GridSearchCV**. By systematically optimizing hyperparameters, we aim to improve model performance and compare implementations across datasets.

We focused on three supervised learning models:

- **Decision Tree**
- **k-Nearest Neighbors (kNN)**
- **Logistic Regression**

Additionally, we combined them into a **Voting Classifier** to test ensemble performance.

---

2. Dataset Description

Wine Quality Dataset

- **Instances:** 1599 red wine samples
- **Features:** 11 chemical properties (e.g., acidity, sugar, alcohol, etc.)
- **Target:** Binary classification (good quality vs. not)
- **Train/Test Split:** 1119 training samples, 480 testing samples

Banknote Authentication Dataset

- **Instances:** 1372 samples
- **Features:** 4 numerical features extracted from images of banknotes
- **Target:** Binary classification (genuine vs. forged)
- **Train/Test Split:** 960 training samples, 412 testing samples

---

3. Methodology

We implemented a **machine learning pipeline**:

`StandardScaler → SelectKBest → Classifier`

- **StandardScaler:** Standardizes features for kNN and Logistic Regression.
- **SelectKBest:** Selects top k features based on ANOVA F-test.
- **Classifier:** Decision Tree, kNN, or Logistic Regression.

Two approaches were used:

Manual Grid Search

- Implemented using nested loops and **5-fold Stratified CV**.
- Calculated average **ROC AUC** for each parameter set.
- Selected the best parameter set and retrained on full training data.

Built-in GridSearchCV

- Used `GridSearchCV` with pipelines.
- `scoring='roc_auc'`, 5-fold Stratified CV.
- Extracted best parameters and compared results with manual implementation.

Evaluation Metrics:

Accuracy, Precision, Recall, F1-Score, ROC AUC.

---

4. Results and Analysis

Wine Quality Dataset

Manual Grid Search – Best Parameters

- **Decision Tree:** { `select__k=5` , `max_depth=5` , `min_samples_split=5` }
- **kNN:** { `select__k=5` , `n_neighbors=9` , `weights='distance'` }
- **Logistic Regression:** { `select__k=10` , `C=1` , `penalty='l2` , `solver='liblinear'` }

Model Performance (Manual)

Model	Accuracy	Precision	Recall	F1	ROC AUC
Decision Tree	0.7271	0.7716	0.6965	0.7321	0.8025
kNN	**0.7750**	0.7854	0.7977	**0.7915**	**0.8679**
Logistic Regression	0.7396	0.7619	0.7471	0.7544	0.8246
Voting Classifier	0.7417	0.7692	0.7393	0.7540	0.8611

Built-in GridSearchCV – Results

- Parameters and metrics matched **exactly** with manual search.
- Confirms correctness of manual implementation.

Analysis:

- kNN outperformed all models with the highest AUC (0.8679).
- Voting Classifier did not surpass standalone kNN.

Banknote Authentication Dataset

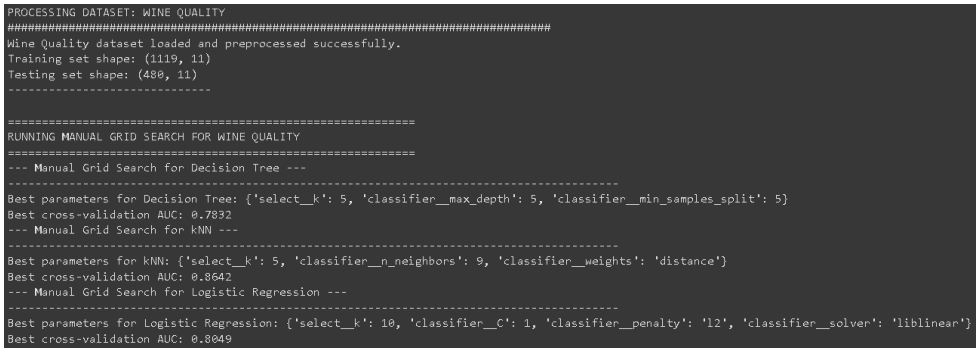
- Training: (960, 4), Testing: (412, 4)
- **Manual Grid Search:** Failed due to `set_params` error (likely mismatch in parameter grid or SelectKBest incompatibility with 4 features).
- **Built-in GridSearchCV:** Not executed due to manual error carry-over.

Analysis:

- Banknote dataset was not successfully processed.
- Error suggests the pipeline configuration must be revised (e.g., SelectKBest `k` cannot exceed 4 features).

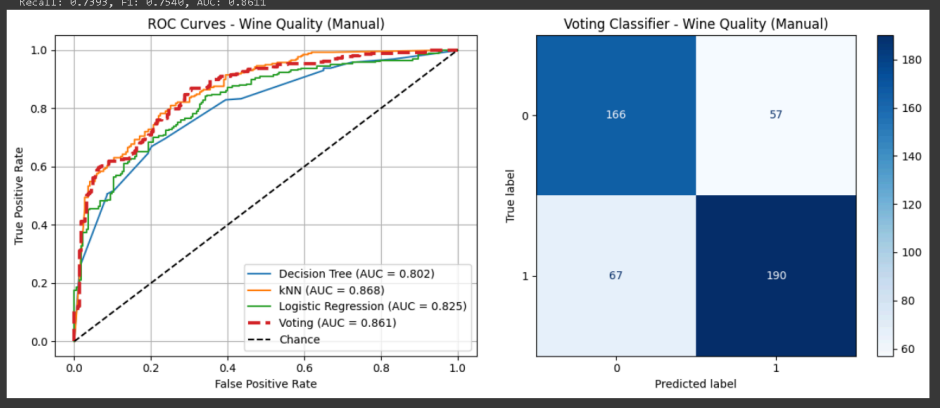
---

5. Screenshots



```
============================================================
EVALUATING MANUAL MODELS FOR WINE QUALITY
============================================================

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.7271
  Precision: 0.7716
  Recall: 0.6965
  F1-Score: 0.7321
  ROC AUC: 0.8025

kNN:
  Accuracy: 0.7750
  Precision: 0.7854
  Recall: 0.7977
  F1-Score: 0.7915
  ROC AUC: 0.8679

Logistic Regression:
  Accuracy: 0.7396
  Precision: 0.7619
  Recall: 0.7471
  F1-Score: 0.7544
  ROC AUC: 0.8246

--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.7417, Precision: 0.7692
  Recall: 0.7393, F1: 0.7540, AUC: 0.8611
```

```
--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.7417, Precision: 0.7692
  Recall: 0.7393, F1: 0.7540, AUC: 0.8611
```



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

--- GridSearchCV for Decision Tree ---
Best params for Decision Tree: {'classifier__max_depth': 5, 'classifier__min_samples_split': 5, 'select__k': 5}
Best CV score: 0.7832

--- GridSearchCV for kNN ---
Best params for kNN: {'classifier__n_neighbors': 9, 'classifier__weights': 'distance', 'select__k': 5}
Best CV score: 0.8642

--- GridSearchCV for Logistic Regression ---
Best params for Logistic Regression: {'classifier__C': 1, 'classifier__penalty': 'l2', 'classifier__solver': 'liblinear', 'select__k': 10}
Best CV score: 0.8049
```

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

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.7271
  Precision: 0.7716
  Recall: 0.6965
  F1-Score: 0.7321
  ROC AUC: 0.8025

kNN:
  Accuracy: 0.7750
  Precision: 0.7854
  Recall: 0.7977
  F1-Score: 0.7915
  ROC AUC: 0.8679

Logistic Regression:
  Accuracy: 0.7396
  Precision: 0.7619
  Recall: 0.7471
  F1-Score: 0.7544
  ROC AUC: 0.8246
```

```
################################################################################
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
============================================================
```

6. Conclusion

- **Wine Quality Dataset:**

    - kNN was the best model overall.
    - Manual and built-in grid search produced identical results, proving correct manual implementation.
    - Voting ensemble did not improve beyond standalone kNN.
- **Banknote Dataset:**
    - Manual search failed due to pipeline parameter mismatch. Needs fixing (likely reduce `k` in SelectKBest).
- **Learnings:**
    - Hyperparameter tuning significantly improves model performance.
    - Manual search is useful for understanding but time-consuming and error-prone.
    - GridSearchCV is efficient, reliable, and practical for real applications.