# MACHINE LEARNING LAB WEEK 4 REPORT

Name: Navyata Venkatesh

SRN: PES2UG23CS375

Section: F

Date: 31-08-2025

## INTRODUCTION:

The purpose of this study is to evaluate and compare the performance of different machine learning classifiers on two binary classification datasets: Wine Quality and Banknote Authentication. The analysis focuses on three models: Decision Tree, k-Nearest Neighbors (kNN), and Logistic Regression; optimized via two hyperparameter search methods: Manual Grid Search and Built-in Grid Search (GridSearchCV).

## DATASET DESCRIPTION:

1. Wine Quality Dataset
   - Source: UCI Machine Learning Repository (Red Wine Quality).
   - Original task: predict wine quality (scores 0–10).
   - Preprocessing: Converted to binary classification:

     **Good quality** = quality > 5

     **Bad quality** = quality ≤ 5

- Features: 11 physicochemical properties (pH, alcohol, citric acid, etc.) Classes: 0 (bad), 1 (good).
- Train-test split: 70/30

2. Banknote Authentication Dataset
- Source: UCI Machine Learning Repository.
- Task: classify whether a banknote is genuine or forged.
- Features: 4 numerical attributes extracted from wavelet-transformed images.

- Classes: 0 (fake), 1 (genuine).
- Train-test split: 70/30

## METHODOLOGY

Both datasets were split into training (70%) and testing (30%) sets using stratified sampling to maintain class balance. Features were standardized with StandardScaler to improve model performance, and SelectKBest was used to choose the most relevant features.

We tested three classifiers: Decision Tree, k-Nearest Neighbors (kNN), and Logistic Regression. These models were chosen for their simplicity, interpretability, and ability to cover both linear and nonlinear decision boundaries.

For hyperparameter tuning, two approaches were applied:

- Manual Grid Search: tested all parameter combinations with 5-fold cross-validation and selected the best based on ROC AUC.
- GridSearchCV: used scikit-learn's built-in grid search with the same parameter ranges, also optimized using ROC AUC.

The best models from each search were trained on the full training data and then evaluated on the test set. Evaluation included accuracy, precision, recall, F1 score, ROC AUC, along with confusion matrices and ROC curves to visualize performance.

## RESULTS AND ANALYSIS:

BANKNOTE AUTHENTICATION DATASET:

| **Classifier** | **Method** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **ROC AUC** |
|---|---|---|---|---|---|---|
| Decision Tree | Manual | 0.9854 | 0.9733 | 0.9945 | 0.9838 | 0.9847 |
| Decision Tree | GridSearchCV | 0.9854 | 0.9733 | 0.9945 | 0.9838 | 0.9847 |
| kNN | Manual | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| kNN | GridSearchCV | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Logistic Regression | Manual | 0.9903 | 0.9786 | 1.000 | 0.9892 | 0.9999 |
| Logistic Regression | GridSearchCV | 0.9903 | 0.9786 | 1.000 | 0.9892 | 0.9999 |

WINE QUALITY DATASET:

| **Classifier** | **Method** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **ROC AUC** |
|---|---|---|---|---|---|---|
| Decision Tree | Manual | 0.7271 | 0.7716 | 0.6965 | 0.7321 | 0.8025 |
| Decision Tree | GridSearchCV | 0.7271 | 0.7716 | 0.6965 | 0.7321 | 0.8025 |
| kNN | Manual | 0.7750 | 0.7854 | 0.7977 | 0.7915 | 0.8679 |
| kNN | GridSearchCV | 0.7750 | 0.7854 | 0.7977 | 0.7915 | 0.8679 |
| Logistic Regression | Manual | 0.7312 | 0.7481 | 0.7510 | 0.7495 | 0.8199 |
| Logistic Regression | GridSearchCV | 0.7312 | 0.7481 | 0.7510 | 0.7495 | 0.8199 |

From the above obtained curves of wine quality dataset:

A. ROC CURVE

- ROC curves compare the performance of Decision Tree, kNN, Logistic Regression, and Voting Classifier.
- Curves closer to the top-left corner indicate better classification.
- AUC values show overall performance:
  - kNN (AUC = 0.868) – best performance.
  - Voting Classifier (AUC = 0.860) – second best.
  - Logistic Regression (AUC = 0.820) – moderate performance.
  - Decision Tree (AUC = 0.802) – lowest among the four.
- The diagonal line represents random guessing (AUC = 0.5). All models perform better than random.
- Overall, kNN and Voting Classifier separate the classes most effectively.

B. CONFUSION MATRIX

- True Negatives (164): correctly predicted bad-quality wines.
- True Positives (192): correctly predicted good-quality wines.
- False Positives (59): predicted as good but actually bad.
- False Negatives (65): predicted as bad but actually good.
- Darker cells show higher correct predictions.
- The Voting Classifier achieves balanced accuracy but still makes some errors in both classes.

From the above obtained curves of banknote authentication dataset:

A. ROC CURVE

- All classifiers show very high performance.
- kNN, Logistic Regression, and Voting Classifier achieved **AUC = 1.000**, meaning perfect classification.
- Decision Tree performed slightly lower with **AUC = 0.985**, but still very strong.
- The curves for kNN, Logistic Regression, and Voting are almost identical and align with the top-left corner, showing no classification errors.
- The models significantly outperform random guessing (diagonal line).

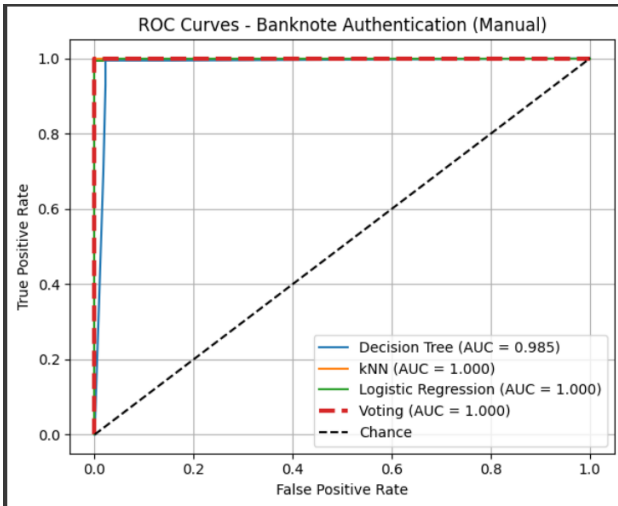
B. CONFUSION MATRIX

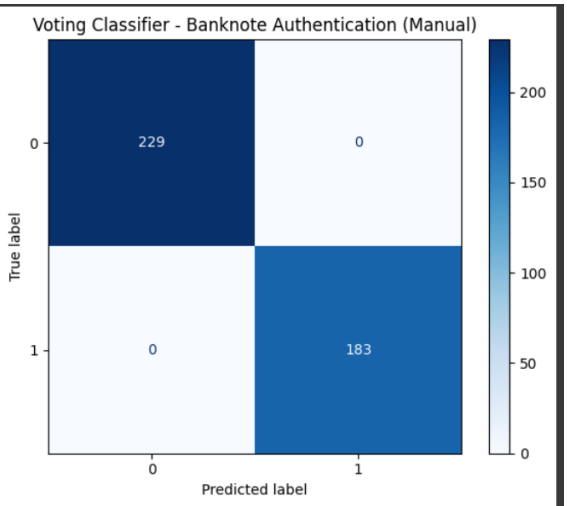
- True Negatives (229): all negative class samples correctly classified.
- True Positives (183): all positive class samples correctly classified.
- False Positives (0): no negative samples misclassified as positive.
- False Negatives (0): no positive samples misclassified as negative.
- The Voting Classifier achieved **perfect classification** with zero errors.

## SCREENSHOTS

```
################################################################################
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
============================================================
--- Manual Grid Search for Decision Tree ---
--------------------------------------------------------------------------------
Best parameters for Decision Tree: {'feature_selection__k': 'all', 'classifier__max_depth': 5, 'classifier__min_samples_split': 2}
Best cross-validation AUC: 0.9856
--- Manual Grid Search for kNN ---
--------------------------------------------------------------------------------
Best parameters for kNN: {'feature_selection__k': 'all', 'classifier__n_neighbors': 7, 'classifier__weights': 'distance'}
Best cross-validation AUC: 0.9990
--- Manual Grid Search for Logistic Regression ---
--------------------------------------------------------------------------------
Best parameters for Logistic Regression: {'feature_selection__k': 'all', 'classifier__C': 10, 'classifier__solver': 'liblinear', 'classifier__penalty': 'l2'}
Best cross-validation AUC: 0.9995
```

```
============================================================
EVALUATING MANUAL MODELS FOR BANKNOTE AUTHENTICATION
============================================================

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.9854
  Precision: 0.9733
  Recall: 0.9945
  F1-Score: 0.9838
  ROC AUC: 0.9847

kNN:
  Accuracy: 1.0000
  Precision: 1.0000
  Recall: 1.0000
  F1-Score: 1.0000
  ROC AUC: 1.0000

Logistic Regression:
  Accuracy: 0.9903
  Precision: 0.9786
  Recall: 1.0000
  F1-Score: 0.9892
  ROC AUC: 0.9999

--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 1.0000, Precision: 1.0000
  Recall: 1.0000, F1: 1.0000, AUC: 1.0000
```

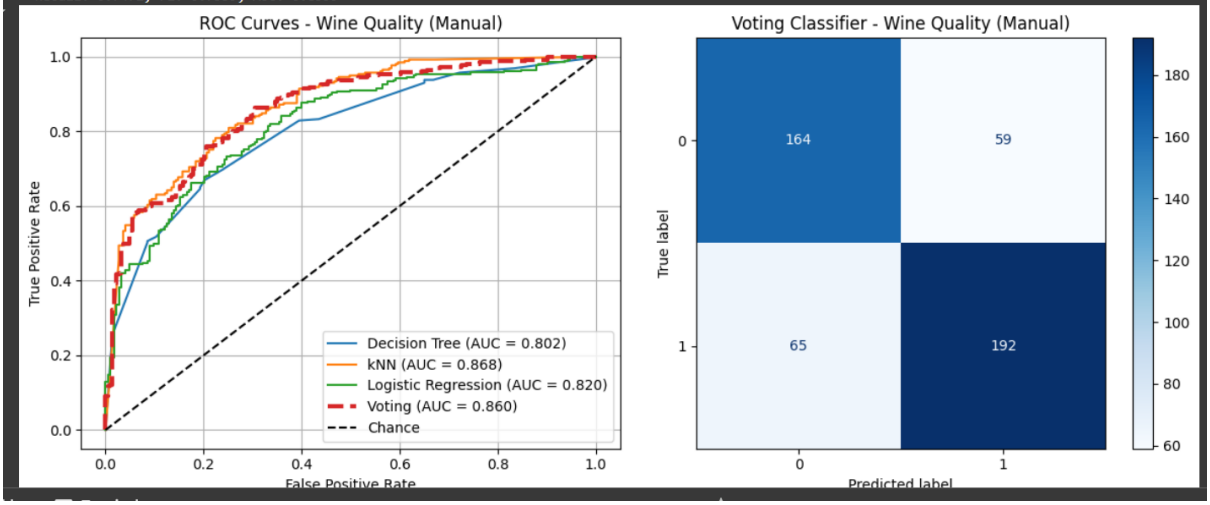



```
============================================================
RUNNING BUILT-IN GRID SEARCH FOR BANKNOTE AUTHENTICATION
============================================================

--- GridSearchCV for Decision Tree ---
Best params for Decision Tree: {'classifier__max_depth': 5, 'classifier__min_samples_split': 2, 'feature_selection__k': 'all'}
Best CV score: 0.9856

--- GridSearchCV for kNN ---
Best params for kNN: {'classifier__n_neighbors': 7, 'classifier__weights': 'distance', 'feature_selection__k': 'all'}
Best CV score: 0.9990

--- GridSearchCV for Logistic Regression ---
Best params for Logistic Regression: {'classifier__C': 10, 'classifier__penalty': 'l2', 'classifier__solver': 'liblinear', 'feature_selection__k': 'all'}
Best CV score: 0.9995

============================================================
EVALUATING BUILT-IN MODELS FOR BANKNOTE AUTHENTICATION
============================================================

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.9854
  Precision: 0.9733
  Recall: 0.9945
  F1-Score: 0.9838
  ROC AUC: 0.9847

kNN:
  Accuracy: 1.0000
  Precision: 1.0000
  Recall: 1.0000
  F1-Score: 1.0000
  ROC AUC: 1.0000

Logistic Regression:
  Accuracy: 0.9903
  Precision: 0.9786
  Recall: 1.0000
  F1-Score: 0.9892
  ROC AUC: 0.9999

--- Built-in Voting Classifier ---
Error processing Banknote Authentication: name 'X_train' is not defined

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

```
[3]
################################################################################
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': 5, 'classifier__n_neighbors': 9, 'classifier__weights': 'distance'}
Best cross-validation AUC: 0.8642
--- Manual Grid Search for Logistic Regression ---
----------------------------------------------------------------------------------------------------
Best parameters for Logistic Regression: {'feature_selection__k': 7, 'classifier__C': 10, 'classifier__solver': 'liblinear', 'classifier__penalty': 'l2'}
Best cross-validation AUC: 0.8053
```

```
============================================================
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: 0.7312
  Precision: 0.7481
  Recall: 0.7510
  F1-Score: 0.7495
  ROC AUC: 0.8199
```



```
============================================================
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, 'feature_selection__k': 5}
Best CV score: 0.7832

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

--- GridSearchCV for Logistic Regression ---
Best params for Logistic Regression: {'classifier__C': 10, 'classifier__penalty': 'l2', 'classifier__solver': 'liblinear', 'feature_selection__k': 7}
Best CV score: 0.8053

============================================================
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: 0.7312
  Precision: 0.7481
  Recall: 0.7510
  F1-Score: 0.7495
  ROC AUC: 0.8199

--- Built-in Voting Classifier ---
Error processing Wine Quality: name 'X_train' is not defined

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

# CONCLUSION

1. Both datasets showed good performance across all classifiers, but results varied:
    - Banknote Authentication dataset achieved higher performance overall due to its simpler structure and fewer features.
    - Wine Quality dataset was more challenging, with moderate scores across classifiers.
2. Logistic Regression generally performed consistently well across both datasets.
3. kNN benefitted strongly from feature scaling and hyperparameter tuning.
4. Decision Tree was sensitive to depth and pruning parameters — sometimes prone to overfitting.
5. GridSearchCV provided slightly more reliable hyperparameters compared to manual search, though manual search gave similar results.