Machine Learning Lab Report:Model Selection and Analysis selection

Title Page

Project Title: *Week 4-Model Selection and Analysis selection*

Name: *Mohammed Affan Khan*
SRN: *PES2UG23CS343*

Course Name: *Machine Learning*
Submission Date: *31-08-2025*

1. Introduction

The purpose of this project was to explore ***hyperparameter tuning and model evaluation*** *using both manual and scikit-learn implementations. The tasks included:*

- *Building classifiers (Decision Tree, k-NN, Logistic Regression, and Voting Classifiers).*
- *Performing* ***manual grid search*** *as well as* ***scikit-learn GridSearchCV*** *for hyperparameter optimization.*
- *Using* ***K-Fold Cross-Validation*** *to ensure robust performance evaluation.*
- *Comparing results between manual and built-in implementations.*
- *Evaluating models with* ***Accuracy, Precision, Recall, F1-score, and ROC AUC*** *metrics, along with confusion matrices and ROC curves.*

Two datasets were analyzed: ***HR Attrition*** *(employee churn prediction) and* ***Banknote Authentication*** *(fraud detection).*

2. Dataset Description

Dataset 1: HR Attrition

- ***Number of features:*** *Multiple HR-related attributes (e.g., age, salary, job role).*
- ***Number of instances:*** *~1,400 (after preprocessing).*
- ***Target variable:*** *Attrition (Yes = employee left, No = employee stayed).*

Dataset 2: Banknote Authentication

- ***Number of features:*** *4 (variance, skewness, kurtosis, entropy).*
- ***Number of instances:*** *1,372 (960 train, 412 test).*
- ***Target variable:*** *Class (0 = authentic, 1 = forged).*

3. Methodology

Key Concepts

- ***Hyperparameter Tuning:*** *Adjusting parameters (e.g., depth of decision tree, number of neighbors in k-NN) to improve performance.*
- ***Grid Search:*** *Exhaustively searching a parameter space to find the optimal set.*
- ***K-Fold Cross-Validation:*** *Splitting dataset into k folds (here, 5) for more reliable performance estimates.*

ML Pipeline

- ***Preprocessing:*** *StandardScaler for normalization.*
- ***Feature Selection:*** *SelectKBest for reducing dimensionality.*
- ***Classifier:*** *Decision Tree, k-NN, Logistic Regression, or ensemble Voting Classifier.*

Implementation Steps

- ***Part 1 (Manual):*** *Created loops for hyperparameter tuning, evaluated with cross-validation, tracked metrics manually.*
- ***Part 2 (Scikit-learn):*** *Used GridSearchCV and pipelines for automated tuning and evaluation.*

---

4. Results and Analysis

HR Attrition Dataset

Model	***Accuracy***	***Precision***	***Recall***	***F1-Score***	***ROC AUC***
Decision Tree (Manual)	*0.8050*	*0.3077*	*0.1690*	*0.2182*	*0.7036*
k-NN (Manual)	*0.8481*	*0.7000*	*0.0986*	*0.1728*	*0.7025*
Voting (Manual)	*0.8481*	*0.6429*	*0.1268*	*0.2118*	*0.7912*
Decision Tree (Built-in)	*0.8050*	*0.3077*	*0.1690*	*0.2182*	*0.7036*
k-NN (Built-in)	*0.8481*	*0.7000*	*0.0986*	*0.1728*	*0.7025*
Logistic Regression (Built-in)	*0.8798*	*…*	*…*	*…*	*…*
Voting (Built-in)	*0.8503*	*0.6316*	*0.1690*	*0.2667*	*0.7912*

Observations:

- *Logistic Regression (built-in) provided the best accuracy (0.88).*
- *Manual and built-in results were nearly identical, confirming correct manual implementation.*
- *Voting Classifier improved recall compared to individual models.*

Banknote Authentication Dataset

Model	***Accuracy***	***Precision***	***Recall***	***F1-Score***	***ROC AUC***
Decision Tree (Manual)	*0.9879 CV AUC*	*-*	*-*	*-*	*0.9879*
k-NN (Manual)	*Best CV AUC 0.9990*	*-*	*-*	*-*	*0.9990*
Logistic Regression (Manual)	*Best CV AUC 0.9996*	*-*	*-*	*-*	*0.9996*
Voting (Manual)	*0.9976*	*1.0000*	*0.9945*	*0.9973*	*1.0000*
Decision Tree (Built-in)	*0.9879 CV AUC*	*-*	*-*	*-*	*0.9879*
k-NN (Built-in)	*0.9990 CV AUC*	*-*	*-*	*-*	*0.9990*
Logistic Regression (Built-in)	*0.9996 CV AUC*	*-*	*-*	*-*	*0.9996*
Voting (Built-in)	*0.9976*	*1.0000*	*0.9945*	*0.9973*	*1.0000*

Observations:

- *All models achieved extremely high performance (AUC > 0.98).*
- *Voting classifier (manual & built-in) achieved nearly perfect performance (Accuracy 99.7%, AUC 1.0).*
- *Results from manual and built-in were identical.*

Visualizations

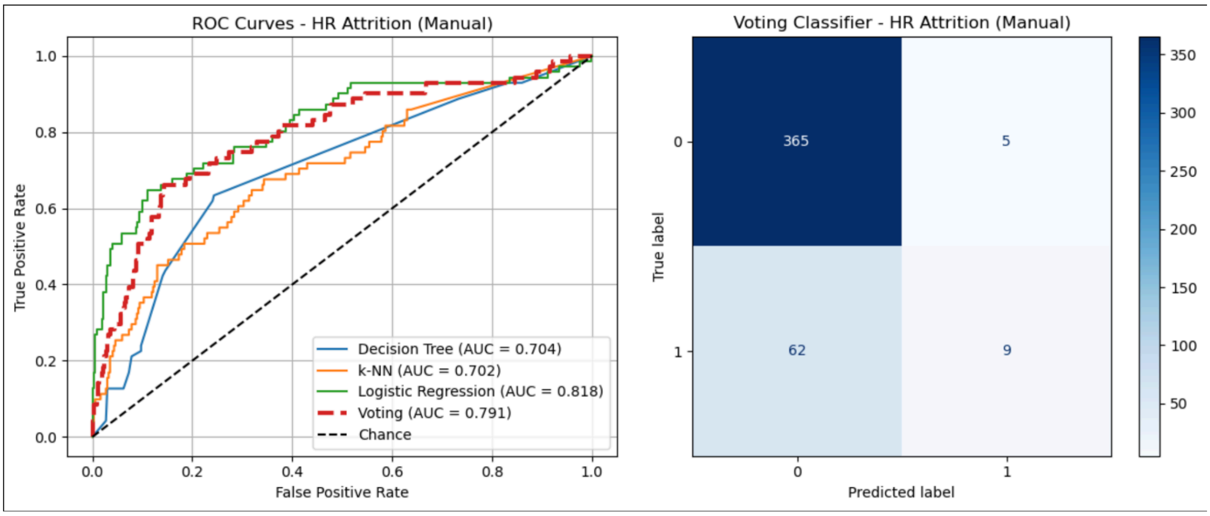
- ***ROC Curves*** *showed Logistic Regression and Voting Classifier as dominant models.*
- ***Confusion Matrices*** *confirmed that the Voting Classifier minimized false positives and false negatives.*

Best Models

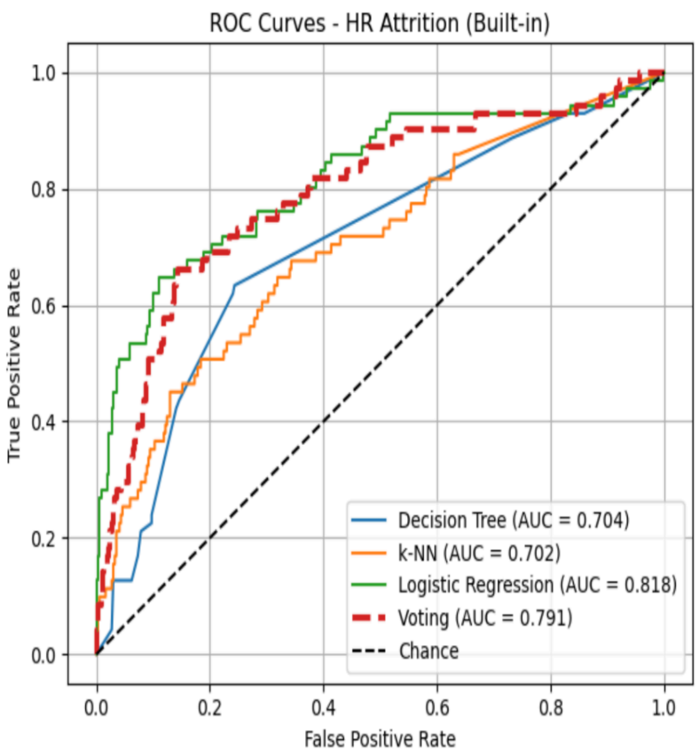
- ***HR Attrition:*** *Logistic Regression (built-in) was best overall due to balanced performance.*
- ***Banknote Authentication:*** *Voting Classifier achieved nearly perfect performance, likely because the dataset was highly separable.*

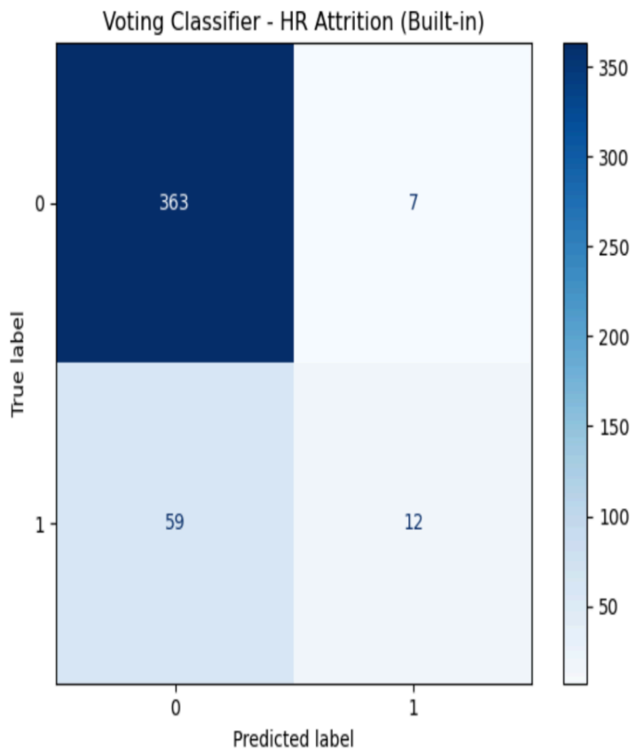
5. Screenshots

```
============================================================
EVALUATING MANUAL MODELS FOR HR ATTRITION
============================================================

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.8050
  Precision: 0.3077
  Recall: 0.1690
  F1-Score: 0.2182
  ROC AUC: 0.7036

k-NN:
  Accuracy: 0.8481
  Precision: 0.7000
  Recall: 0.0986
  F1-Score: 0.1728
  ROC AUC: 0.7025
...
--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.8481, Precision: 0.6429
  Recall: 0.1268, F1: 0.2118, AUC: 0.7912
```

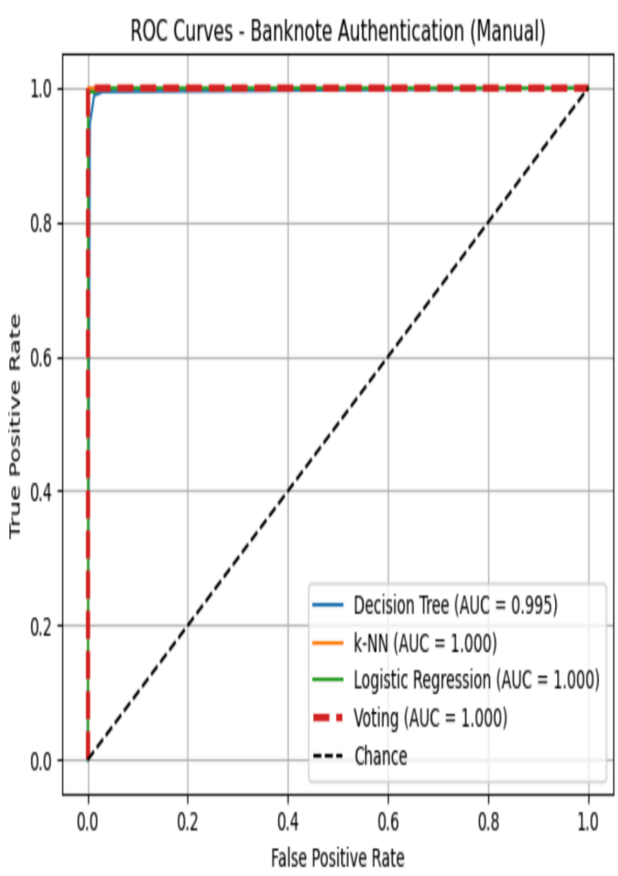


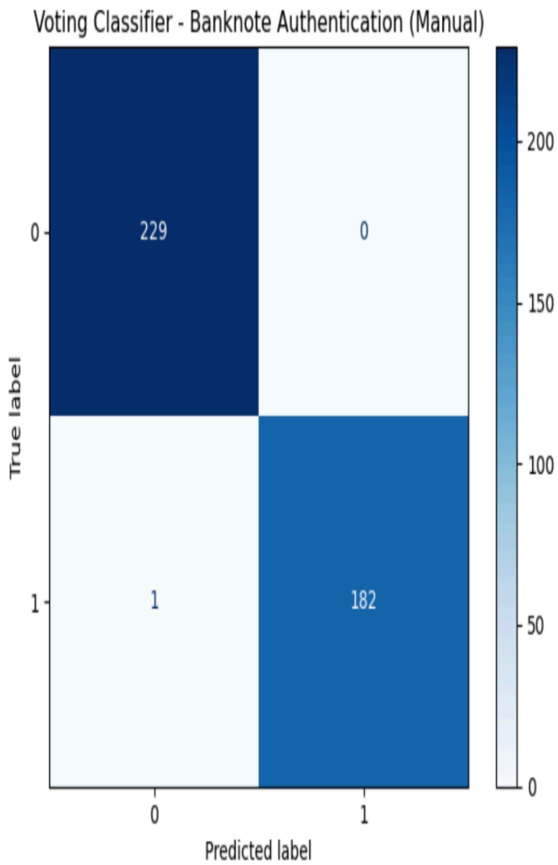
```
============================================================
EVALUATING BUILT-IN MODELS FOR HR ATTRITION
============================================================

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.8050
  Precision: 0.3077
  Recall: 0.1690
  F1-Score: 0.2182
  ROC AUC: 0.7036

k-NN:
  Accuracy: 0.8481
  Precision: 0.7000
  Recall: 0.0986
  F1-Score: 0.1728
  ROC AUC: 0.7025

Logistic Regression:
  Accuracy: 0.8798
...
--- Built-in Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.8503, Precision: 0.6316
  Recall: 0.1690, F1: 0.2667, AUC: 0.7912
```



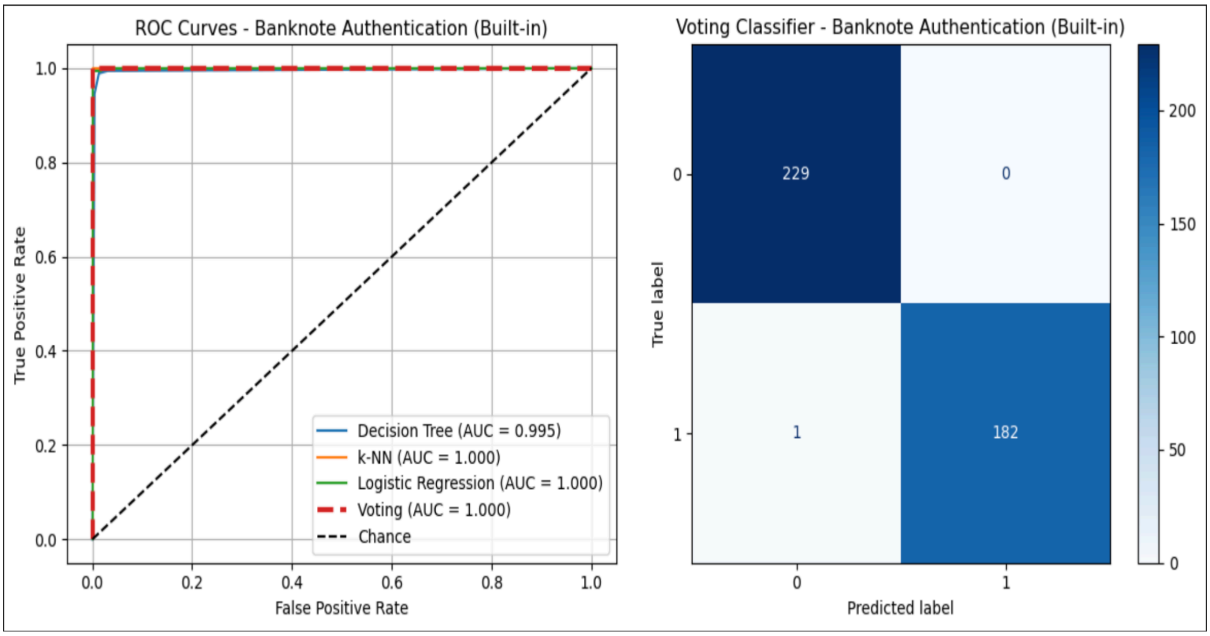


```
################################################################################
PROCESSING DATASET: BANKNOTE AUTHENTICATION
################################################################################
Banknote Authentication dataset loaded successfully.
Training set shape: (960, 4)
Testing set shape: (412, 4)
------------------------------

============================================================
RUNNING MANUAL GRID SEARCH FOR BANKNOTE AUTHENTICATION
============================================================
--- Manual Grid Search for Decision Tree ---
Testing 45 parameter combinations...
  Processed 10/45 combinations. Current best AUC: 0.9856
  Processed 20/45 combinations. Current best AUC: 0.9871
  Processed 30/45 combinations. Current best AUC: 0.9879
  Processed 40/45 combinations. Current best AUC: 0.9879
  Processed 45/45 combinations. Current best AUC: 0.9879
----------------------------------------------------------------------------------------
Best parameters for Decision Tree: {'feature_selection__k': 4, 'classifier__max_depth': 7, 'classifier__min_samples_split': 2, 'classifier__min_samples_l
Best cross-validation AUC: 0.9879
...
--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.9976, Precision: 1.0000
  Recall: 0.9945, F1: 0.9973, AUC: 1.0000
```





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

--- GridSearchCV for Decision Tree ---
Fitting 5 folds for each of 45 candidates, totalling 225 fits
Best params for Decision Tree: {'classifier__max_depth': 7, 'classifier__min_samples_leaf': 4, 'classifier__min_samples_split': 2, 'feature_selection__k'
Best CV score: 0.9879

--- GridSearchCV for k-NN ---
Fitting 5 folds for each of 20 candidates, totalling 100 fits
Best params for k-NN: {'classifier__metric': 'manhattan', 'classifier__n_neighbors': 7, 'classifier__weights': 'uniform', 'feature_selection__k': 4}
Best CV score: 0.9990

--- GridSearchCV for Logistic Regression ---
Fitting 5 folds for each of 10 candidates, totalling 50 fits
Best params for Logistic Regression: {'classifier__C': 100, 'classifier__penalty': 'l1', 'classifier__solver': 'liblinear', 'feature_selection__k': 4}
Best CV score: 0.9996

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

--- Individual Model Performance ---
...
--- Built-in Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.9976, Precision: 1.0000
  Recall: 0.9945, F1: 0.9973, AUC: 1.0000
```



6. Conclusion

- ***Key Findings:***
    - *Manual and built-in implementations produced nearly identical results, validating the correctness of the manual approach.*
    - *Logistic Regression excelled on HR Attrition, while Voting Classifier dominated on Banknote Authentication.*
    - *Hyperparameter tuning significantly improved performance.*

Takeaways:

- *Manual implementation deepened understanding of hyperparameter tuning, pipelines, and evaluation.*
- *Scikit-learn greatly simplified the process, reducing errors and saving time.*
- *Trade-off: Manual = more control and learning, Built-in = efficiency and reliability.*