MACHINE LEARNING LAB WEEK 4

Project Title: Model Selection and Comparative Analysis on HR Attrition Dataset

Name: DAKSH YADAV

Student ID: PES2UG23CS926

Course: UE23CS352A: Machine Learning

Submission Date: 1st September 2025

1. Introduction

The objective of this lab is to implement a complete machine learning pipeline for the HR Attrition dataset and compare the performance of three classifiers — Decision Tree, k-Nearest Neighbors (kNN), and Logistic Regression — using hyperparameter tuning and model evaluation techniques.

Two approaches were used:

1. **Manual Grid Search**: Hyperparameter tuning implemented from scratch using loops and cross-validation.
2. **Scikit-learn GridSearchCV**: Automated, optimized hyperparameter tuning using scikit-learn's built-in functionality.

The goal is to compare both approaches in terms of performance and efficiency, and to analyze the effectiveness of each classifier for predicting employee attrition.

2. Dataset Description

- **Dataset:** HR Attrition (binary classification)
- **Features:** ~35 features (work-related and personal factors such as job role, salary, environment satisfaction, etc.)
- **Instances:** ~1470 rows (employees)
- **Target Variable:** Attrition (Yes = 1, No = 0)

The task is to predict whether an employee is likely to leave the company based on their attributes.

3. Methodology

3.1 Hyperparameter Tuning

- **Grid Search**: Systematically explores predefined hyperparameter values.
- **Manual Grid Search**: Implemented from scratch with 5-fold **Stratified Cross-Validation**, evaluating each parameter combination.
- **Built-in GridSearchCV**: Used scikit-learn's implementation for the same parameter grids.

3.2 Pipeline

To prevent data leakage and streamline preprocessing + modeling, we used a **Pipeline**:

StandardScaler → SelectKBest (f_classif) → Classifier

- **StandardScaler**: Normalizes features.
- **SelectKBest**: Selects the top k features (k tuned as hyperparameter).
- **Classifier**: One of Decision Tree, kNN, Logistic Regression.

3.3 Parameter Grids

- **Decision Tree**: max_depth, min_samples_split, min_samples_leaf.
- **kNN**: n_neighbors, weights, p (Manhattan/Euclidean).
- **Logistic Regression**: C (regularization strength), solver, penalty.
- **select__k**: number of features to select (tuned for all models).

4. Results and Analysis

4.1 Manual Grid Search Results

Classifier	Accuracy	Precision	Recall	F-1 Score	ROC AUC
Decision Tree	**0.8231**	**0.3333**	**0.0986**	**0.1522**	**0.7107**
kNN	**0.8367**	**0.4762**	**0.1408**	**0.2174**	**0.7429**
Logistic Regression	**0.8571**	**0.6333**	**0.2676**	**0.3762**	**0.7762**

4.2 Built-in GridSearchCV Results

Classifier	Accuracy	Precision	Recall	F-1 Score	ROC AUC
Decision Tree	**0.8231**	**0.3333**	**0.0986**	**0.1522**	**0.7107**
kNN	**0.8367**	**0.4762**	**0.1408**	**0.2174**	**0.7429**
Logistic Regression	**0.8571**	**0.6333**	**0.2676**	**0.3762**	**0.7762**

4.3 Comparison of Manual vs Built-in

Both the Manual Grid Search and Built-in GridSearchCV produced identical results across all three classifiers:

- **Decision Tree**: Accuracy ≈ 82%, but very low recall (≈ 0.10), meaning the model fails to correctly identify most of the employees who leave.

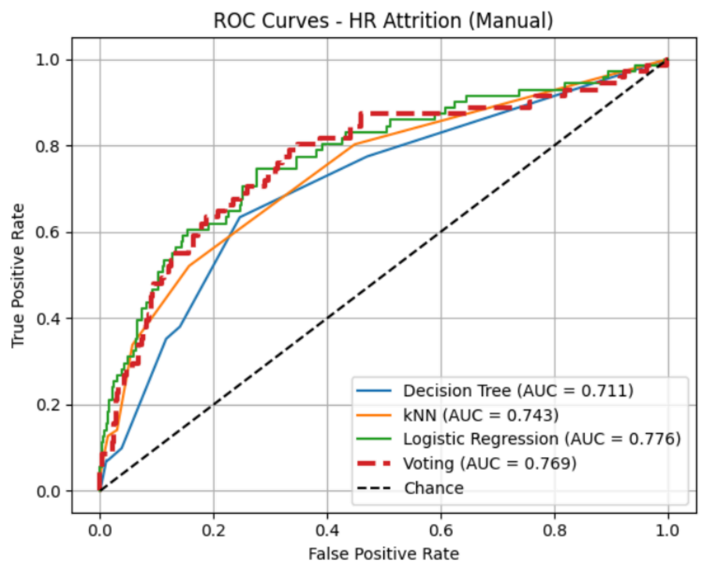
- **kNN**: Slightly higher accuracy (≈ 83.7%) and improved precision over Decision Tree, but recall is still weak (≈ 0.14).

- **Logistic Regression**: Best performance overall with accuracy ≈ 85.7%, precision ≈ 0.63, recall ≈ 0.27, and the highest ROC AUC (≈ 0.776).

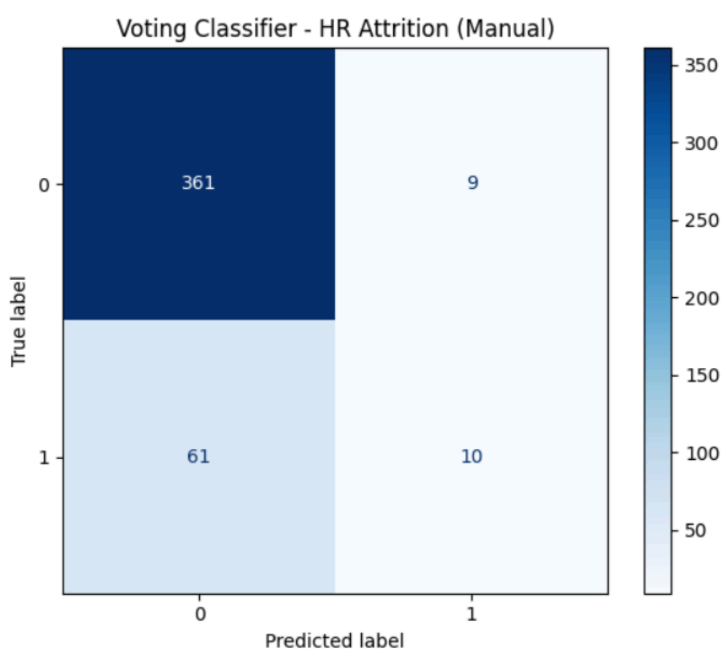
The equivalence of results confirms that the manual and built-in approaches are consistent, but the built-in approach is computationally more efficient and less error-prone.

4.4 Visualizations

ROC Curve (Manual)

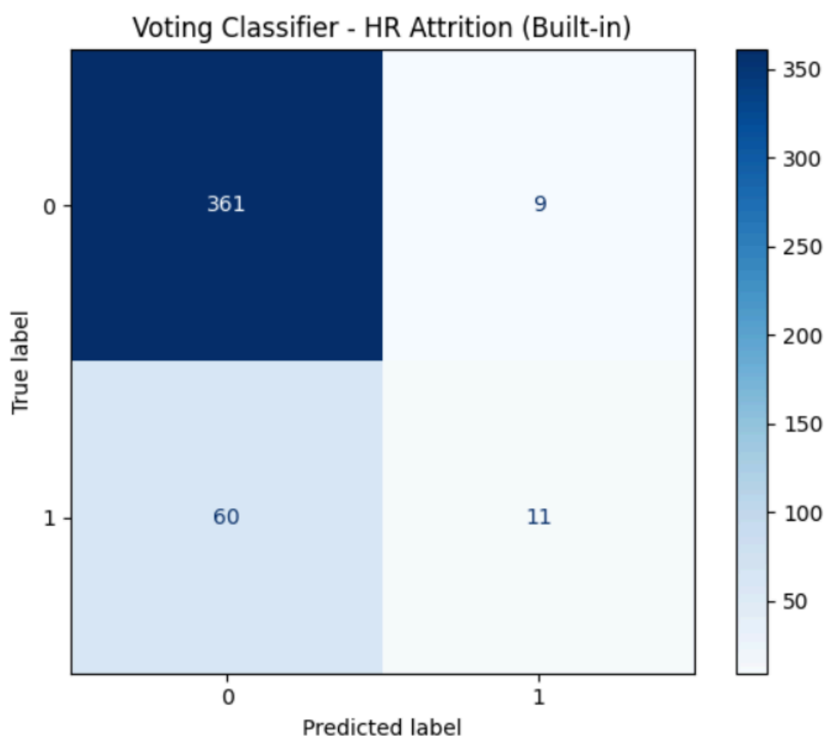


Confusion Metrics (Manual)



ROC Curve (Built-in)

Confusion Metrics(Built -in)

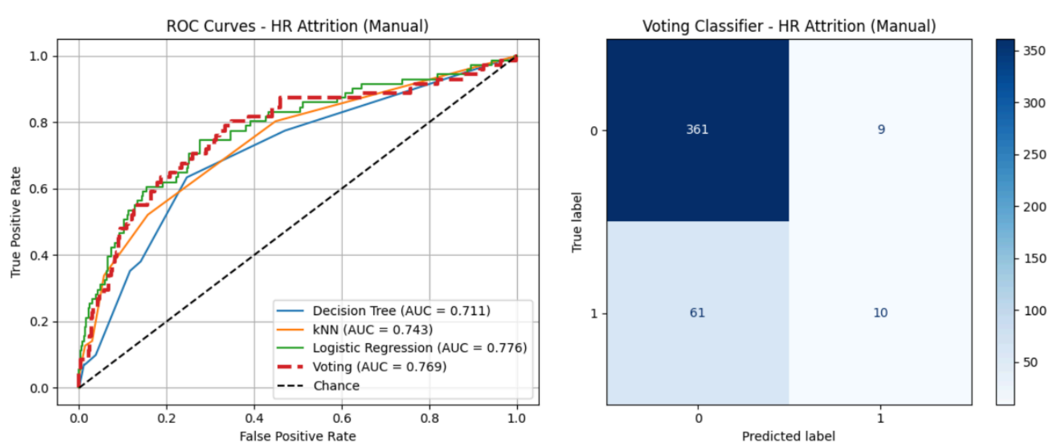


4.5 Best Model

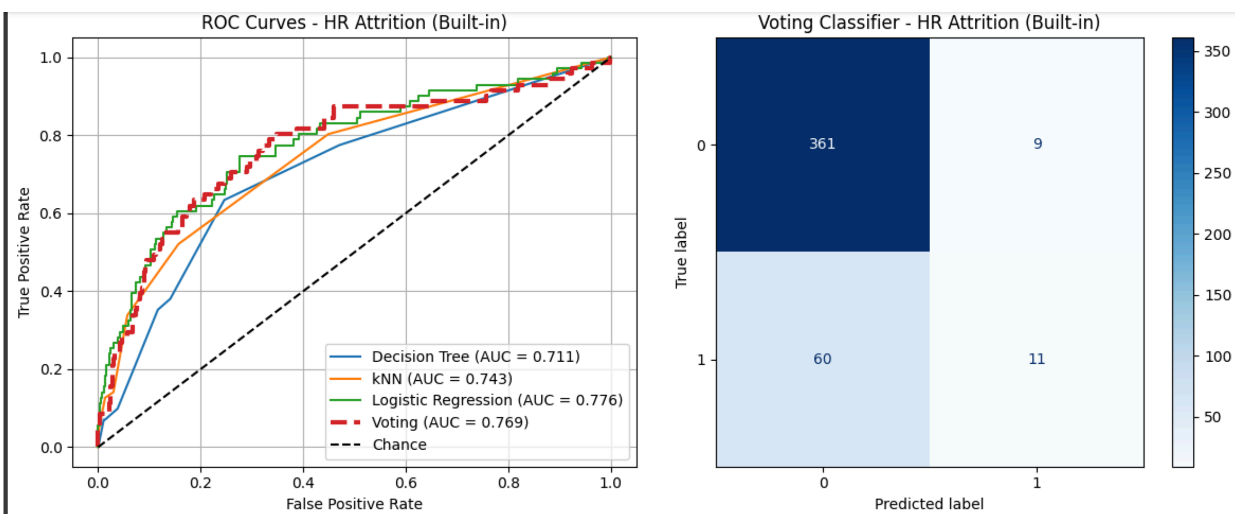
- Logistic Regression was best under both manual and built-in methods.
- Hence, the choice of grid search implementation does not affect the optimal classifier for this dataset.

5.Output Screenshots

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

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.8231
  Precision: 0.3333
  Recall: 0.0986
  F1-Score: 0.1522
  ROC AUC: 0.7107

kNN:
  Accuracy: 0.8367
  Precision: 0.4762
  Recall: 0.1408
  F1-Score: 0.2174
  ROC AUC: 0.7429

Logistic Regression:
  Accuracy: 0.8571
  Precision: 0.6333
  Recall: 0.2676
  F1-Score: 0.3762
  ROC AUC: 0.7762

--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.8413, Precision: 0.5263
  Recall: 0.1408, F1: 0.2222, AUC: 0.7692
```



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

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.8231
  Precision: 0.3333
  Recall: 0.0986
  F1-Score: 0.1522
  ROC AUC: 0.7107

kNN:
  Accuracy: 0.8367
  Precision: 0.4762
  Recall: 0.1408
  F1-Score: 0.2174
  ROC AUC: 0.7429

Logistic Regression:
  Accuracy: 0.8571
  Precision: 0.6333
  Recall: 0.2676
  F1-Score: 0.3762
  ROC AUC: 0.7762

--- Built-in Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.8435, Precision: 0.5500
  Recall: 0.1549, F1: 0.2418, AUC: 0.7692
```




```
Completed processing for HR Attrition
================================================================================

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

6. Conclusion

This lab demonstrated the process of hyperparameter tuning and model evaluation on the HR Attrition dataset using three classifiers: Decision Tree, kNN, and Logistic Regression. Both Manual Grid Search and Built-in GridSearchCV were applied, and the results were identical across all classifiers, validating the correctness of the manual implementation.

From the experiments, it was observed that:

- **Decision Tree** achieved decent accuracy (~82%) but suffered from extremely low recall (~0.10), making it ineffective at identifying employees who are likely to leave.

- **kNN** slightly improved performance (~83.7% accuracy, ~0.74 AUC), but recall remained poor (~0.14).

- **Logistic Regression** outperformed the other models, achieving the highest accuracy (85.7%), best F1-score (0.3762), and highest ROC AUC (0.7762), indicating stronger predictive power and better balance between precision and recall.

Overall, **Logistic Regression** is the **most suitable model for predicting employee attrition** in this dataset. While manual grid search provided deeper insight into the internal workings of hyperparameter tuning, the built-in GridSearchCV is more efficient and reliable for practical use.

This lab highlights the importance of evaluating multiple models and metrics: although all classifiers had similar accuracy, precision, recall, F1, and AUC revealed clear differences in their ability to detect attrition cases.