ML Lab

Lab 4 - Submission

Name: Gauthamdev R Holla
SRN: PES2UG23CS197
Branch: CSE
Sem: V
Section: C

---

1) Introduction

The Lab involves training and comparing three classifiers: Decision Tree, KNN, and Logistic Regression.

The tasks performed were:

- Load the 3 classifiers.
- Preprocess the dataset.
- Run Manual & Built-In Grid Search.
- Evaluate the performance of individual models.
- Combine model results using a soft voting ensemble.
- Visualize the results by plotting a graph.

---

2) Dataset

- ***Description:*** Dataset provides employee details. It is to help predict whether an employee will leave the company or stay.
- ***No. of Instances:*** 1470 employees
- ***No. of Features:*** 34 input features
- ***Target Variable:*** Attrition

---

3) Methodology

Key Concepts:

- ***Hyperparameter Tuning:*** Process of selecting the best combo of model parameters that aren't learned during training.
- ***Grid Search:*** Method that tests all combinations of specified hyperparameters to find the best one.

- ***K-Fold Cross-Validation:*** Splits data into k subsets, trains on k–1 subsets, and validates on the remaining subset. This is repeated k times.

ML Pipeline:

- ***StandardScaler:*** Normalizes features to have zero mean and unit variance.
- ***SelectKBest:*** Picks the top k features.
- ***Classifier:*** Applies one of the models (Decision Tree, k-NN, Logistic Regression) for prediction.

Grid Search:

- ***Manual:*** Iterates over all parameter combinations, measures each combination, and then finds the best one.
- ***Built-In:*** Does everything that the Manual Grid Search does by using the *GridSearchCV* tool.

---

4) Results & Analysis

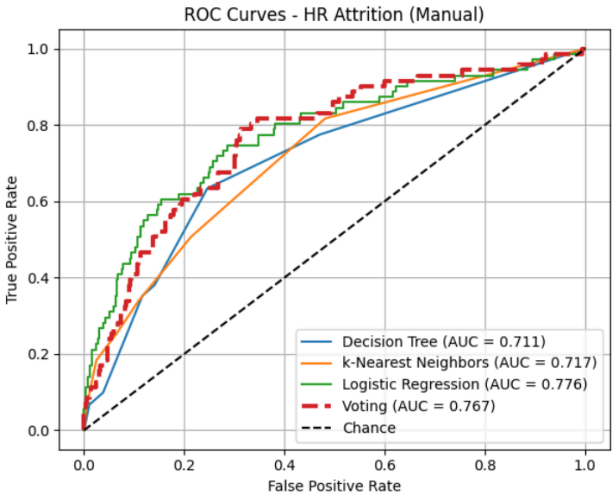
Performance Table:

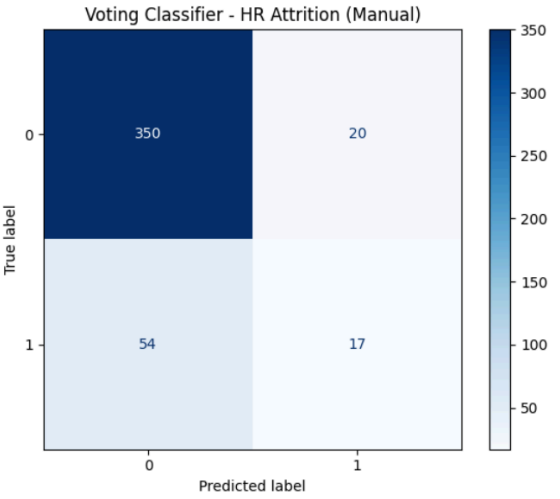
Classifier	Method	Accuracy	Precision	Recall	F1-Score	ROC AUC
Decision Tree	Manual	0.8231	0.3333	0.0986	0.1522	0.7107
Decision Tree	Built-In	0.8231	0.3333	0.0986	0.1522	0.7107
KNN	Manual	0.8254	0.4286	0.2535	0.3186	0.7172
KNN	Built-In	0.8254	0.4286	0.2535	0.3186	0.7172
Logistic Regression	Manual	0.8571	0.6333	0.2676	0.3762	0.7759
Logistic Regression	Built-In	0.8571	0.6333	0.2676	0.3762	0.7759
Voting Classifier	Manual	0.8322	0.4595	0.2394	0.3148	0.7675
Voting Classifier	Built-In	0.8345	0.4688	0.2113	0.2913	0.7675

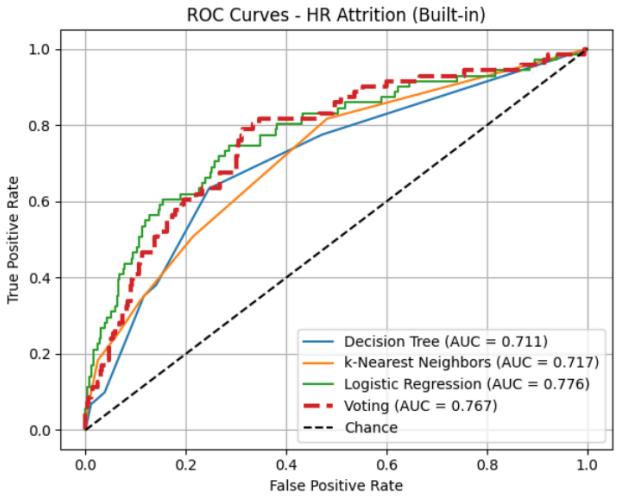
Implementation Comparison:

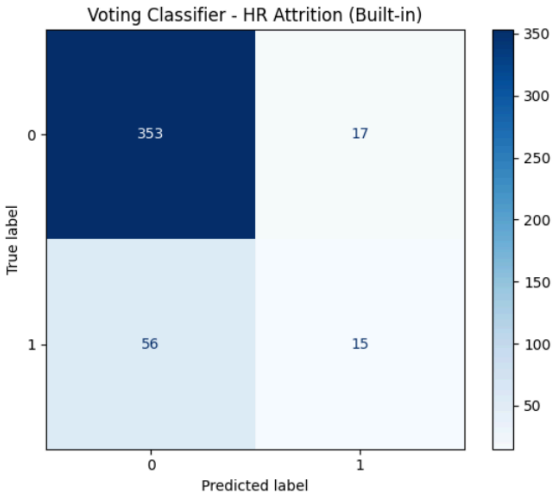
- Results for both Manual and Built-In are nearly identical.
- Results were identical because the same parameter combinations were used.
- Slight variation in voting classifier metrics due to averaging and rounding.

Visualizations:









- Logistic Regression and Voting Classifier show the highest AUC values.
- Voting classifiers indicate balanced predictions.

Best Model:

- Logistic Regression is the best model as it has the highest ROC AUC value of 0.7759.

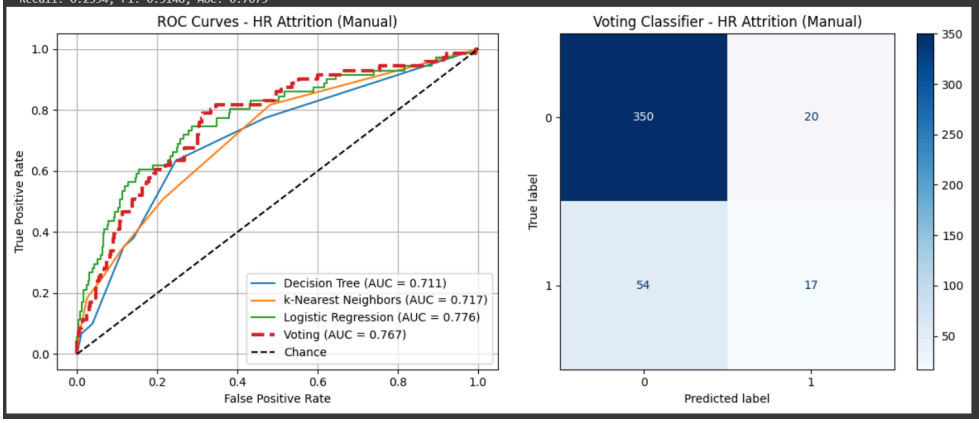
---

5) Screenshots

```
============================================================
RUNNING MANUAL GRID SEARCH FOR HR ATTRITION
============================================================
--- Manual Grid Search for Decision Tree ---
--------------------------------------------------------------------------------
Best parameters for Decision Tree: {'feature_selection__k': 5, 'classifier__max_depth': 3}
Best cross-validation AUC: 0.7152
--- Manual Grid Search for k-Nearest Neighbors ---
--------------------------------------------------------------------------------
Best parameters for k-Nearest Neighbors: {'feature_selection__k': 10, 'classifier__n_neighbors': 7}
Best cross-validation AUC: 0.7002
--- Manual Grid Search for Logistic Regression ---
--------------------------------------------------------------------------------
Best parameters for Logistic Regression: {'feature_selection__k': 15, 'classifier__C': 0.1}
Best cross-validation AUC: 0.7774
Manual grid search completed.
```

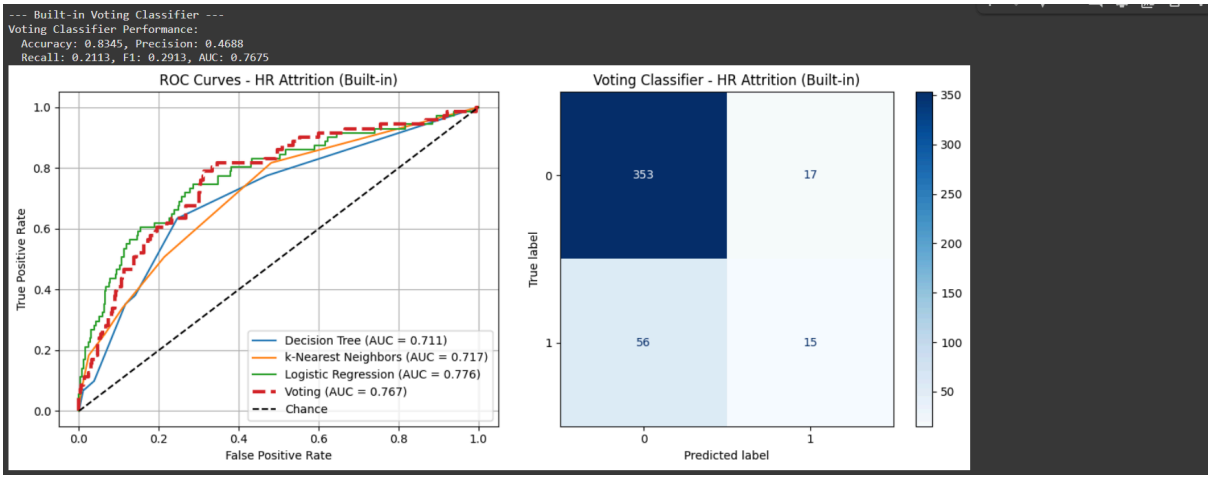
```
============================================================
EVALUATING MANUAL MODELS FOR HR ATTRITION
============================================================

--- Individual Model Performance ---

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

k-Nearest Neighbors:
  Accuracy: 0.8254
  Precision: 0.4286
  Recall: 0.2535
  F1-Score: 0.3186
  ROC AUC: 0.7172

Logistic Regression:
  Accuracy: 0.8571
  Precision: 0.6333
  Recall: 0.2676
  F1-Score: 0.3762
  ROC AUC: 0.7759
```

```
--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.8322, Precision: 0.4595
  Recall: 0.2394, F1: 0.3148, AUC: 0.7675
```




```
============================================================
RUNNING BUILT-IN GRID SEARCH FOR HR ATTRITION
============================================================

--- GridSearchCV for Decision Tree ---
Best params for Decision Tree: {'classifier__max_depth': 3, 'feature_selection__k': 5}
Best CV score: 0.7152

--- GridSearchCV for k-Nearest Neighbors ---
Best params for k-Nearest Neighbors: {'classifier__n_neighbors': 7, 'feature_selection__k': 10}
Best CV score: 0.7002

--- GridSearchCV for Logistic Regression ---
Best params for Logistic Regression: {'classifier__C': 0.1, 'feature_selection__k': 15}
Best CV score: 0.7774
Built-in grid search completed.
```

```
============================================================
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

k-Nearest Neighbors:
  Accuracy: 0.8254
  Precision: 0.4286
  Recall: 0.2535
  F1-Score: 0.3186
  ROC AUC: 0.7172

Logistic Regression:
  Accuracy: 0.8571
  Precision: 0.6333
  Recall: 0.2676
  F1-Score: 0.3762
  ROC AUC: 0.7759
```



---

6) Conclusion

Key Findings:

- Logistic Regression was the best model with the highest ROC AUC score.
- The Voting Classifier had better overall accuracy and precision by combining predictions from all three models.
- Both manual and built-in grid search methods gave identical results.

Main Takeaways:

- Manual Grid Search gives more control but is time-consuming.
- Built-In Grid Search is faster, easier to use, and just as effective as Manual.
