Model Selection and Comparative Analysis

NAME: CHERUKURI VENKATA KARTIK

SRN: PES2UG23CS148

COURSE TITLE: MACHINE LEARNING

DATE : 01.09.2025

1.OVERVIEW

In this lab we are tuning the hyperparameters that is we are choosing the best hyperparameters among the ones given, we are using , manual and built-in grids here to choose the best hyperparameters. We have been given multiple models that is **KNN**,**Logistic Regression**,**Decision Tree** and choosing the the best model in these three.

2.DATASET FEATURES

- **Number of Features**:There are 35 features in this dataset
- **Number of instances**:The dataset has 1470 instances that is the number of rows, it gives us the employee details
- **Target Variable**:The target variable for the dataset is "attrition" , which means , if the employee has left the company or not .
  - **YES**: The employees that have left the company are around 240
  - **NO**: The employees that have not left the company are around 1230

3.METHODOLOGY

- HYPERPARARAMETERS TUNING : Hyperparameters are set so that we get the best performance from models. It can be described as the process of finding the best possible combination of hyperparameters to obtain maximum/best performance.
- GRID SEARCH: It is a method of hyperparameter tuning where a loop is used to calculate all possible combination of hyperparameters available and choosing the best

one among them, this method is suitable for less number of variables/hyperparameters but a lengthy one when the number of hyperparameters is relatively large.

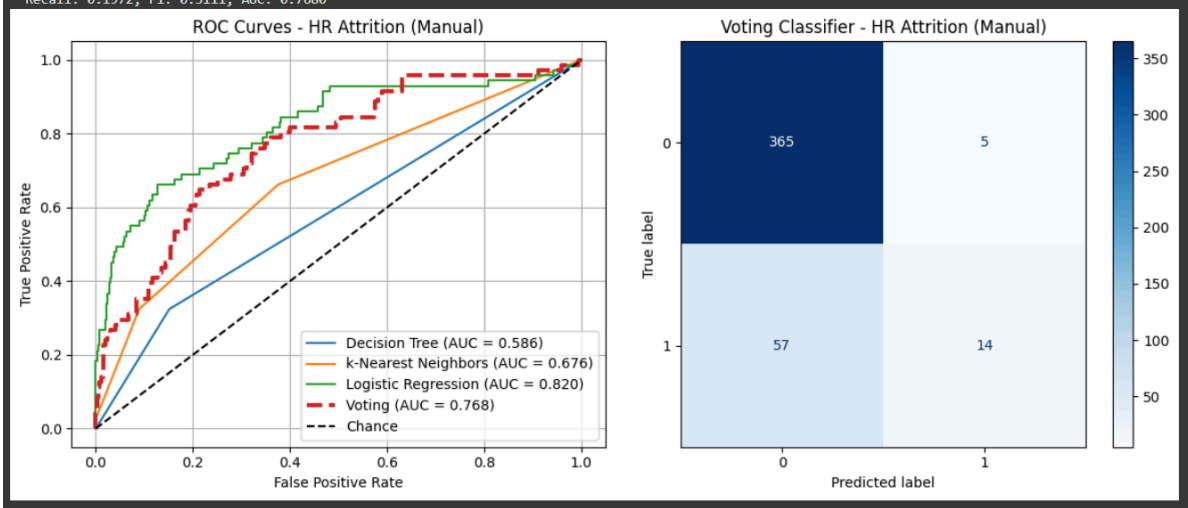
- K-Fold Cross Validation: In this method of performance evaluation , the dataset is divided into k subsets then the model is trained for k times, where one fold is used as testing set and the remaining k-1 folds are used as training set.
- StandardScaler: It is a preprocessing step where standard deviation is set as 1 and mean is set as 0.
- SelectKBest: It is statistical test which calculates score of each feature and chooses the K best features with high scores
- Manual_Grid_Search: In this function we are checking for all possible hyperparameter combinations and then trying to choose the best one for each knn, logistic regression and decision tree.
- run_builtin_grid_search: This function is uses the built-in scikit function to automated the implemented function . It runs all the processes parallelly hence reducing the execution time.

4.Result and Analysis

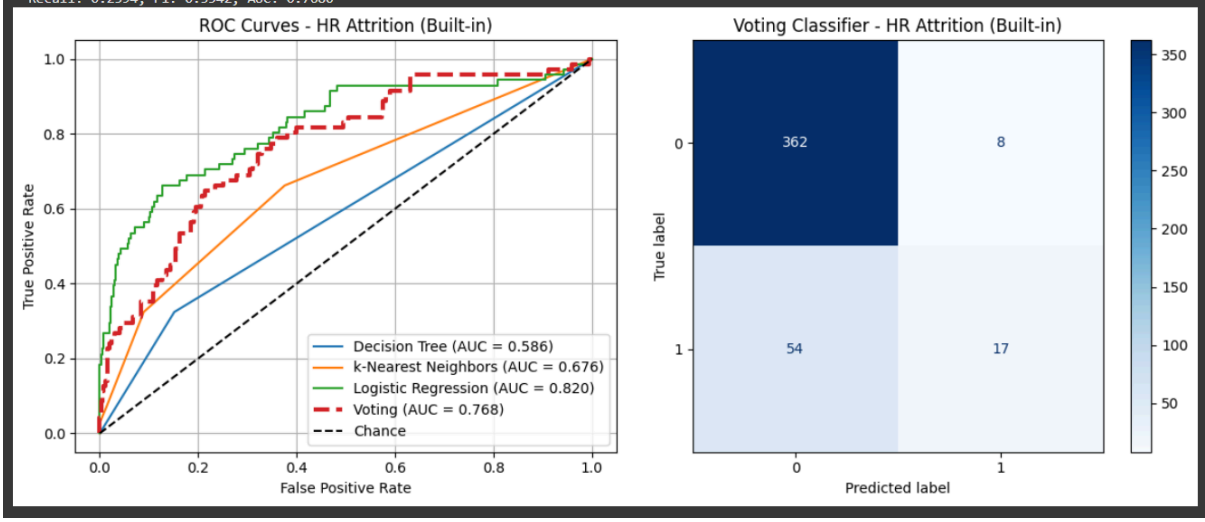
Classifier	Part	Accuracy	Precision	Recall	F1-Score	ROC AUC
Decision Tree	Part 1	0.7642	0.2911	0.3239	0.3067	0.5863
	Part 2	0.7642	0.2911	0.3239	0.3067	0.5863
Logistic Regression	Part 1	0.8798	0.7250	0.4085	0.5225	0.8204
	Part 2	0.8798	0.7250	0.4085	0.5225	0.8204
K-Nearest Neighbors	Part 2	0.8390	0.5000	0.0704	0.1235	0.6756
	Part 2	0.8390	0.5000	0.0704	0.1235	0.6756

The values are exactly the same because the basic functionality of both the functions is the same it is just that the built-in function is more efficient . Both the function use the same methods to find the best possible values of hyperparameters , hence end up having exactly same values.

```
--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.8594, Precision: 0.7368
  Recall: 0.1972, F1: 0.3111, AUC: 0.7680
```




```
Voting Classifier Performance:
  Accuracy: 0.8594, Precision: 0.6800
  Recall: 0.2394, F1: 0.3542, AUC: 0.7680
```



Logistic regression turns out to be the best model out of all showing good precision , recall values and the ROC AUC Score is also better than the others.

---

5. OUTPUT OF CODE:

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

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.7642
  Precision: 0.2911
  Recall: 0.3239
  F1-Score: 0.3067
  ROC AUC: 0.5863

k-Nearest Neighbors:
  Accuracy: 0.8390
  Precision: 0.5000
  Recall: 0.0704
  F1-Score: 0.1235
  ROC AUC: 0.6756

Logistic Regression:
  Accuracy: 0.8798
  Precision: 0.7250
  Recall: 0.4085
  F1-Score: 0.5225
  ROC AUC: 0.8204

--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.8594, Precision: 0.7368
  Recall: 0.1972, F1: 0.3111, AUC: 0.7680
```

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

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.7642
  Precision: 0.2911
  Recall: 0.3239
  F1-Score: 0.3067
  ROC AUC: 0.5863

k-Nearest Neighbors:
  Accuracy: 0.8390
  Precision: 0.5000
  Recall: 0.0704
  F1-Score: 0.1235
  ROC AUC: 0.6756

Logistic Regression:
  Accuracy: 0.8798
  Precision: 0.7250
  Recall: 0.4085
  F1-Score: 0.5225
  ROC AUC: 0.8204
```

```
################################################################################
PROCESSING DATASET: HR ATTRITION
################################################################################
IBM HR Attrition dataset loaded and preprocessed successfully.
Training set shape: (1029, 46)
Testing set shape: (441, 46)
------------------------------
```

The Screenshots display the performance matrices , screenshots for the curve and confusion matrix have been given in the previous section.

6.Conclusion

- Logistic Regression turns out to be the best model out of all with an accuracy of 0.8798, precision of 0.7250,recall of 0.4085, F1-Score of 0.5225 and ROC AUC value of 0.8204
- We learnt two ways of finding the best hyperparameters , one manually and other using scikit functions. Both the function give same output but the built-in function is more efficient
- We also learnt about hyperparameters and explored different ways to train a model