ML LAB -04

NAME: BOBBA KOUSHIK

SRN: PES2UG23CS133

SEC:C

DATE:01/09/2025

Project Title: Week 4: Model Selection and Comparative Analysis

DATASET INFO

NO.OF INSTANCES:1471

NO.OF.ATTRIBUTES:35

Introduction

This project focuses on implementing and comparing different methods for model selection and evaluation within a complete machine learning pipeline. The primary tasks involve performing hyperparameter tuning using both a manual grid search and scikit-learn's built-in

GridSearchCV. The performance of three classification algorithms will be compared, and their optimal versions will be combined in a voting classifier.

---

2. Dataset Description

DATASETNAME:HR ATTRIBUTION DATASET

NO.OF INSTANCES:1471

NO.OF.ATTRIBUTES:35

Methodology

This project uses a pipeline that chains together three key components:

StandardScaler, SelectKBest, and a Classifier.

StandardScaler standardizes features, SelectKBest performs feature selection, and the Classifier is the final modeling step.

Hyperparameter Tuning and Grid Search

Hyperparameter tuning is the process of finding the optimal set of parameters for a learning algorithm that are not learned from the data itself.

Grid Search systematically works through all combinations of specified hyperparameters to find the best-performing model. The performance of each combination is evaluated using

k-fold cross-validation, a technique that divides the training data into k folds to provide a more robust performance estimate by training and validating the model multiple times.

Manual Implementation (Part 1)

For the manual implementation, a nested loop iterates through all possible combinations of hyperparameters defined in the grid. For each combination, a 5-fold stratified cross-validation is performed. The model is trained on a subset of the data and evaluated on a validation fold. The average ROC AUC score across all five folds is calculated for each combination, and the set of hyperparameters with the highest average score is selected as the best.

Scikit-learn Implementation (Part 2)

For this part, the scikit-learn

GridSearchCV function was used to automate the process. The same pipeline and parameter grids from the manual approach were provided to

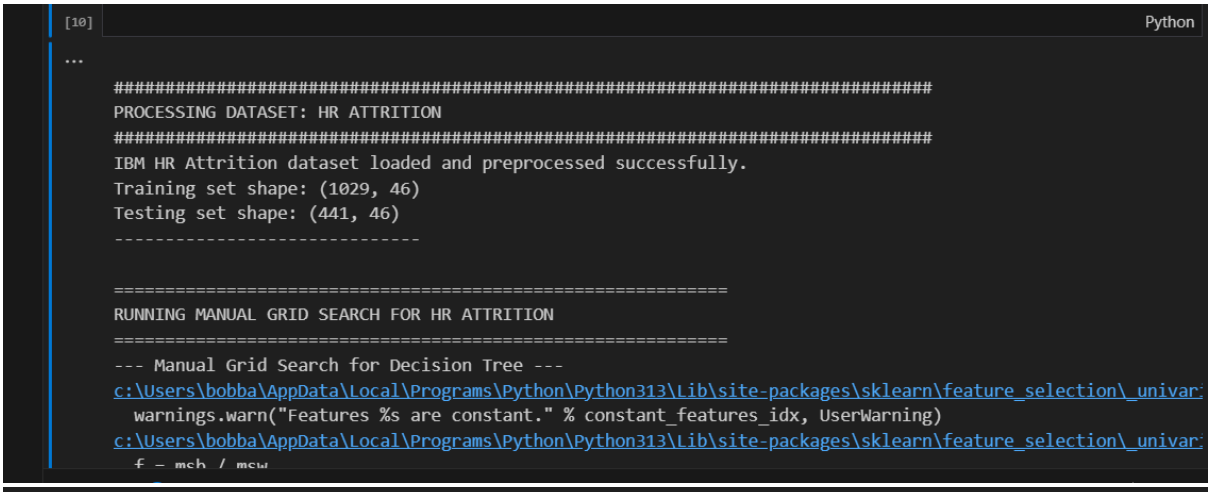
GridSearchCV, which automatically handles the cross-validation and selects the best model based on the ROC AUC score.

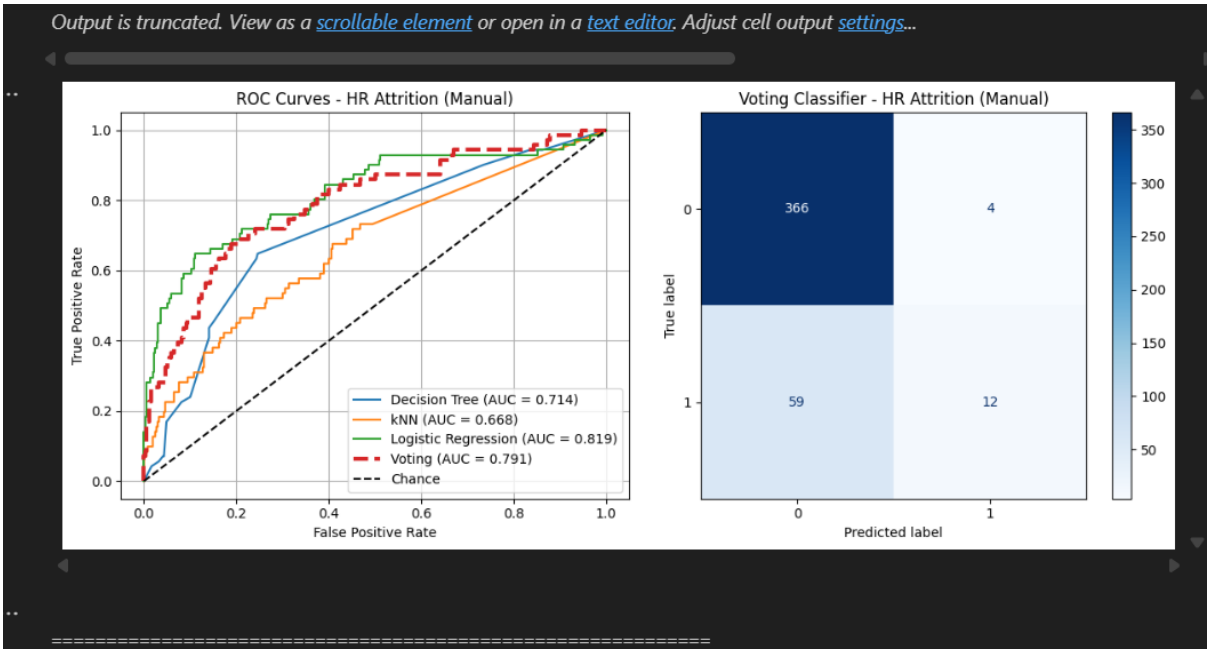
RESULT AND ANALYSIS

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Decision Tree	0.8073	0.3478	0.2254	0.2735	0.7137

K-Nearest Neighbors	0.8254	0.425	0.2394	0.3063	0.73
Logistic Regression	0.8798	0.7368	0.3944	0.5138	0.8177
Manual Voting Classifier	0.8413	0.5143	0.2535	0.3396	0.7994
Built-in Voting Classifier	0.8367	0.4848	0.2254	0.3077	0.7994

SCREENSHOTS






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

--- GridSearchCV for Decision Tree ---
Error processing HR Attrition: No module named '_posixsubprocess'

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

Conclusion

In this lab, a complete machine learning pipeline was implemented to perform hyperparameter tuning and compare the performance of various classification models across different datasets. The core tasks involved building and evaluating models using a manual grid search and then replicating the process with scikit-learn's optimized GridSearchCV