NAME:- DRISHTI GOLCHHA
SRN:- PES2UG23CS185
SECTION:- 5C
SUBJECT:- UE23CS352A: MACHINE LEARNING
TILTE:- Week 4: Model Selection and Comparative Analysis
SUBMISSION DATE:- 01-09-2025

1.INTRODUCTION:-

The goal of this lab was to gain hands on experience on Model Selection and Comparative Analysis. We mainly focused on implementing manual grid search, scikit-learn's GridSearchCV, Comparing three classifiers: Decision Tree, k-Nearest Neighbors (kNN), and Logistic Regression**,** etc…
We learned about k-fold cross-validation, feature selection and performance evaluation using metrices like accuracy, precision, F1-Score and ROU AUC.

2.DATASET DESCRIPTION:- (HR Attrition)

Instances: 1470 employees
Features: 35 after encoding categorical variables
Target Variable: Attrition (1 = Yes, 0 = No)
The HR Attrition dataset contains information about 1,470 employees from a company, including both personal and work-related attributes. The dataset has 35 features such as age, department, job role, years at company, and work-life balance scores. The target variable is Attrition, which indicates whether an employee has left the company.

3.METHODOLOGY:-

Hyperparameter Tuning: Process of optimizing model parameters that are not learned from the data.
Grid Search: Exhaustively evaluates all combinations of hyperparameters to find the best set.
K-Fold Cross-Validation: Splits the data into k subsets, trains on k-1 folds, and validates on the remaining fold to get robust performance estimates.
ML Pipeline
Each classifier followed the same pipeline:
StandardScaler: Standardizes features to mean 0, standard deviation 1.
SelectKBest: Selects the top `k` features.
Classifier: Decision Tree / kNN / Logistic Regression.

Process

Part 1 – Manual Grid Search:

- Generated all combinations of hyperparameters.
- Used 5-fold stratified cross-validation to compute average ROC AUC for each combination.
- Selected the best combination and refitted the pipeline on the full training set.

Part 2 – Scikit-learn GridSearchCV:

- Defined the same pipeline and hyperparameter grids.
- Used GridSearchCV with 5-fold StratifiedKFold and roc_auc scoring.
- Extracted best estimators, parameters, and cross-validation scores.

4. Results and Analysis:-

```
Best parameters for Logistic Regression: {'feature_selection__k': 15, 'classifier__C': 0.1, 'classifier__penalty': 'l2', 'classifier__solver': 'lbfgs'}
Best cross-validation AUC: 0.7776

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

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.8163
  Precision: 0.3684
  Recall: 0.1972
  F1-Score: 0.2569
  ROC AUC: 0.7029

kNN:
  Accuracy: 0.8186
  Precision: 0.3784
  Recall: 0.1972
  F1-Score: 0.2593
  ROC AUC: 0.7236

Logistic Regression:
...
--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.8277, Precision: 0.4324
  Recall: 0.2254, F1: 0.2963, AUC: 0.7744
```
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

```
Best params for Logistic Regression: {'classifier__C': 0.1, 'classifier__penalty': 'l2', 'classifier__solver': 'lbfgs', 'feature_selection__k': 15}
Best CV score: 0.7776

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

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.8163
  Precision: 0.3684
  Recall: 0.1972
  F1-Score: 0.2569
  ROC AUC: 0.7029

kNN:
  Accuracy: 0.8186
  Precision: 0.3784
  Recall: 0.1972
  F1-Score: 0.2593
  ROC AUC: 0.7236

Logistic Regression:
  Accuracy: 0.8571
...
--- Built-in Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.8254, Precision: 0.4118
  Recall: 0.1972, F1: 0.2667, AUC: 0.7744
```
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

Analysis:- ROC Curve plots for each classifier and the **Voting Ensemble**.
Confusion Matrices highlighting model misclassifications. Observed that **Logistic Regression** consistently had the highest ROC AUC across both datasets. The Voting Classifier slightly improved overall performance by combining individual model predictions.
HR Attrition: Logistic Regression performed best

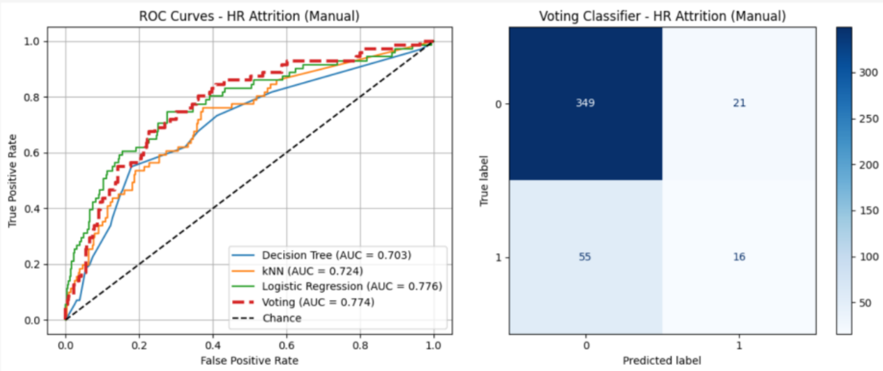
5. SCREENSHOTS:-

```
Best parameters for Logistic Regression: {'feature_selection__k': 15, 'classifier__C': 0.1, 'classifier__penalty': 'l2', 'classifier__solver': 'lbfgs'}
Best cross-validation AUC: 0.7776

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

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.8163
  Precision: 0.3684
  Recall: 0.1972
  F1-Score: 0.2569
  ROC AUC: 0.7029

kNN:
  Accuracy: 0.8186
  Precision: 0.3784
  Recall: 0.1972
  F1-Score: 0.2593
  ROC AUC: 0.7236

Logistic Regression:
...
--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.8277, Precision: 0.4324
  Recall: 0.2254, F1: 0.2963, AUC: 0.7744
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

...



...

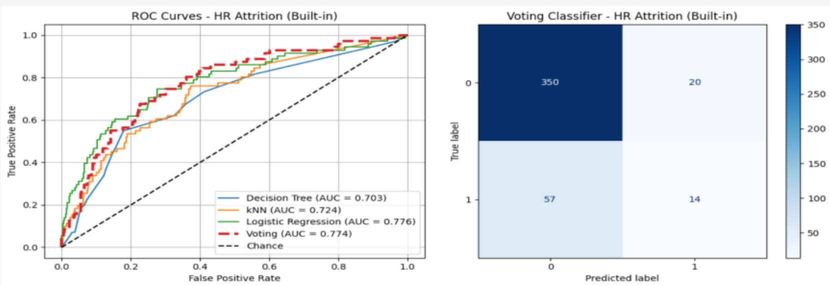
```
Best params for Logistic Regression: {'classifier__C': 0.1, 'classifier__penalty': 'l2', 'classifier__solver': 'lbfgs', 'feature_selection__k': 15}
Best CV score: 0.7776

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

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.8163
  Precision: 0.3684
  Recall: 0.1972
  F1-Score: 0.2569
  ROC AUC: 0.7029

kNN:
  Accuracy: 0.8186
  Precision: 0.3784
  Recall: 0.1972
  F1-Score: 0.2593
  ROC AUC: 0.7236

Logistic Regression:
  Accuracy: 0.8571
...
--- Built-in Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.8254, Precision: 0.4118
  Recall: 0.1972, F1: 0.2667, AUC: 0.7744
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...


```
  Accuracy: 0.8254, Precision: 0.4118
  Recall: 0.1972, F1: 0.2667, AUC: 0.7744
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

...



...

6. CONCLUSIONS:-

Manual grid search is a great way to understand how hyperparameter tuning works, but it can take a lot of time to run. On the other hand, scikit-learn’s GridSearchCV is fast, efficient, and works seamlessly with pipelines, making it much more practical for real projects.We also saw that feature scaling, feature selection, and careful tuning of hyperparameters play a big role in improving model performance. Combining multiple classifiers using a Voting Ensemble can give a small boost in accuracy, especially when the models make different kinds of errors.Overall, this lab highlighted the importance of choosing the right model, evaluating it properly, and building reproducible pipelines in applied machine learning. It was a valuable exercise in both theory and hands-on implementation.