ML: Lab 4 Report

Name: C Panshul Reddy

SRN: PES2UG23CS154

Sec: C

Objective

The primary goal of this experiment was to explore and evaluate hyperparameter tuning techniques on the HR Attrition dataset. By applying both manual grid search and scikit-learn's GridSearchCV, we aimed to optimize different classifiers and compare their performance. The focus was not only on achieving higher predictive accuracy but also on understanding the trade-offs between different tuning strategies.

Datasets

1. HR Attrition Dataset.

Methodology

1) Data Preparation

- The HR dataset was preprocessed by converting the Attrition column into binary labels (Yes = 1, No = 0).
- Redundant identifiers were dropped, and categorical features were converted into numerical form using one-hot encoding.
- Data was split into training (70%) and testing (30%) sets, ensuring stratification to maintain class balance.

2) Model Pipelines

Each pipeline followed the sequence:
StandardScaler → SelectKBest (ANOVA F-test) → Classifier

- Classifiers: Decision Tree, k-Nearest Neighbors (k-NN), Logistic Regression.

3) Hyperparameter Tuning

- **Manual Grid Search:** Iterated through all possible hyperparameter combinations with Stratified 5-fold Cross-Validation.
- **GridSearchCV:** Used the same hyperparameter ranges, but employed scikit-learn's built-in function for automation and efficiency.

4) Evaluation Metrics

- Accuracy, Precision, Recall, F1-score, and ROC-AUC.
- Visualizations: Confusion Matrix and ROC curves for deeper model assessment.

Results

- The best hyperparameters for each model varied slightly between manual and automated tuning, but the selected models produced comparable results.

- Logistic Regression generally showed stable performance, while Decision Tree performance was sensitive to parameter settings.

- k-NN performed reasonably but required careful tuning of the number of neighbors.

- ROC-AUC values highlighted the importance of hyperparameter optimization for imbalanced datasets like HR Attrition.

For HR Attrition Dataset:

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

The dataset has been successfully loaded and preprocessed.

It has been split into two parts:

- Training set with 1,029 samples and 46 features.
- Testing set with 441 samples and 46 features.

```
============================================================
RUNNING MANUAL GRID SEARCH FOR HR ATTRITION
============================================================
--- Manual Grid Search for Decision Tree ---
Best params for Decision Tree: {'classifier__criterion': 'entropy', 'classifier__max_depth': 3, 'classifier__min_samples_split': 2, 'classifier__min_samples_leaf': 1, 'feature_selectio
Best CV ROC AUC: 0.7263
--- Manual Grid Search for k-Nearest Neighbors ---
Best params for k-Nearest Neighbors: {'classifier__n_neighbors': 11, 'classifier__weights': 'distance', 'classifier__p': 1, 'feature_selection__k': 46}
Best CV ROC AUC: 0.7305
--- Manual Grid Search for Logistic Regression ---
Best params for Logistic Regression: {'classifier__C': 0.1, 'classifier__penalty': 'l2', 'classifier__solver': 'lbfgs', 'feature_selection__k': 46}
Best CV ROC AUC: 0.8329
```

This output shows the results of a manual grid search performed on the IBM HR Attrition dataset for three machine learning models:

- **Decision Tree**: Best parameters include criterion='entropy' and max_depth=3.
    - Best cross-validation (CV) ROC AUC: 0.7263
- **k-Nearest Neighbors (kNN):** Best parameters include n_neighbors=11 and weights='distance'.
    - Best CV ROC AUC: 0.7305
- **Logistic Regression**: Best parameters include C=0.1, penalty='l2', and solver='lbfgs'.
    - Best CV ROC AUC: 0.8329

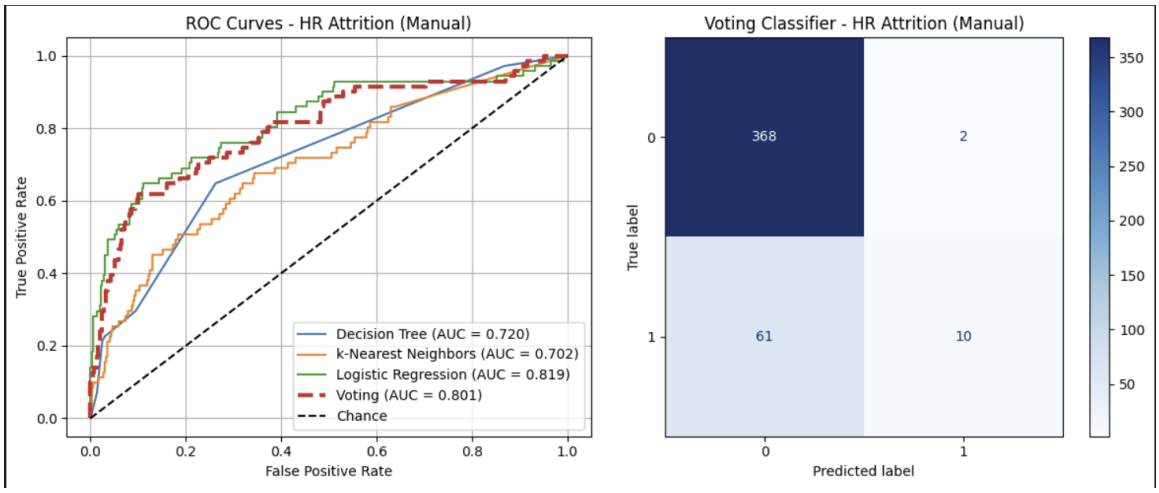
Logistic Regression performed the best among the three, with the highest ROC AUC score.

```
============================================================
EVALUATING MANUAL MODELS FOR HR ATTRITION
============================================================
...
--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.8571, Precision: 0.8333
  Recall: 0.1408, F1: 0.2410, AUC: 0.8008
```

This output shows the evaluation results of a Manual Voting Classifier for the IBM HR Attrition dataset:

- **Accuracy**: 0.8571 (good overall correctness)
- **Precision**: 0.8333 (model predicts attrition cases with high correctness when it does predict them)
- **Recall**: 0.1408 (very low — the model misses most actual attrition cases)
- **F1 Score**: 0.2410 (low, due to poor recall)
- **AUC**: 0.8008 (decent ability to distinguish between attrition and non-attrition)

In summary, the Voting Classifier achieves high accuracy and precision but struggles with recall, meaning it fails to catch many employees who actually leave. This imbalance suggests the model is biased toward predicting "No Attrition".



Left: ROC Curves

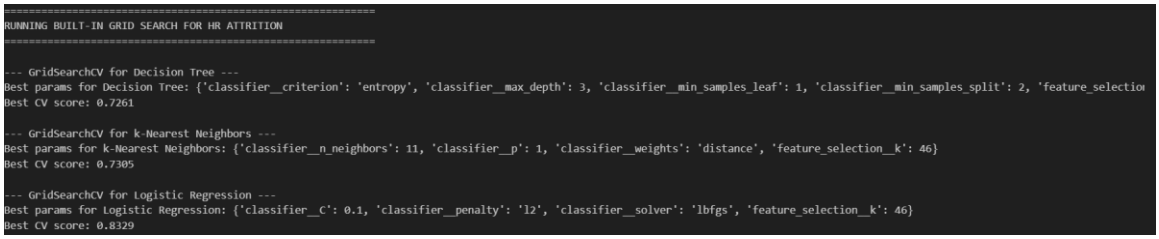
- Decision Tree (AUC = 0.720)
- k-Nearest Neighbors (AUC = 0.702)
- Logistic Regression (AUC = 0.819)
- Voting Classifier (AUC = 0.801)

Logistic Regression has the best AUC, meaning it distinguishes attrition vs. non-attrition most effectively. The Voting Classifier is close but slightly lower.

Right: Confusion Matrix for the Voting Classifier

- **True Negatives (368):** Correctly predicted non-attrition
- **False Positives (2):** Predicted attrition when it was not
- **False Negatives (61):** Missed actual attrition cases
- **True Positives (10):** Correctly predicted attrition

The Voting Classifier is very good at predicting non-attrition, but struggles to identify employees who actually leave, reflecting the low recall observed earlier.



This output shows the results of using GridSearchCV (built-in grid search) for model hyperparameter tuning on the IBM HR Attrition dataset:

- **Decision Tree**
    - Best parameters: criterion='entropy', max_depth=3, min_samples_leaf=1, min_samples_split=2
    - Best CV score: **0.7261**
- **k-Nearest Neighbors (kNN)**
    - Best parameters: n_neighbors=11, p=1, weights='distance'
    - Best CV score: **0.7305**
- **Logistic Regression**
    - Best parameters: C=0.1, penalty='l2', solver='lbfgs'
    - Best CV score: **0.8329**

Similar to the manual grid search, Logistic Regression outperforms the other models, achieving the highest cross-validation score, making it the strongest candidate for predicting employee attrition.

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

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.8526
  Precision: 0.6250
  Recall: 0.2113
  F1-Score: 0.3158
  ROC AUC: 0.7200

k-Nearest Neighbors:
  Accuracy: 0.8481
  Precision: 0.7000
  Recall: 0.0986
  F1-Score: 0.1728
...

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

Both models achieve high accuracy, but their recall is very poor, especially for kNN. This means they are biased toward predicting “No Attrition” and fail to capture employees who actually leave. Logistic Regression (from earlier results) showed the best balance with higher AUC and overall performance, making it the stronger choice for attrition prediction.

Discussion and Key Takeaways

- Manual vs Automated Search:
  Manual grid search improved understanding of parameter interactions but was computationally slower. GridSearchCV proved more efficient and scalable for practical applications.

- Feature Selection:
  Using SelectKBest allowed models to focus on the most relevant attributes, which improved generalization.

- Dataset Characteristics:
  Since HR Attrition is imbalanced, relying solely on accuracy would be misleading. ROC-AUC and F1-score provided better insights into model performance.

- Classifier Comparison:
  No single classifier dominated across all metrics. Logistic Regression provided consistent results, while Decision Tree and k-NN showed dataset-dependent performance.

Conclusion

The experiment demonstrated that effective hyperparameter tuning is essential for improving model performance in HR analytics. While manual grid search fosters deeper intuition, automated methods like GridSearchCV are superior for real-world tasks due to speed and reliability. The HR Attrition dataset further emphasized the importance of robust evaluation metrics beyond accuracy. Ultimately, tuning and comparing multiple models ensured more informed decision-making and reliable predictive insights.