ML: Lab 4 Report

NAME: C YOGESH REDDY

SRN: PES2UG23CS159

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

Data Preparation

- The HR dataset was loaded from CSV and checked for missing values.
- The target variable, 'Attrition', was encoded as binary (Yes = 1, No = 0).
- All categorical features were transformed using one-hot encoding.
- The feature set was stripped of unique identifiers and split into training (70%) and testing (30%) subsets using stratified sampling to ensure the class distribution was preserved.

1) Model Pipeline

- The classification pipeline consisted of:
  StandardScaler → SelectKBest (ANOVA F-test) → Classifier
- Three classifiers were systematically evaluated:

- Decision Tree

    - k-Nearest Neighbors (kNN)
    - Logistic Regression
- For each classifier, relevant hyperparameters (criterion, neighbors, regularization, etc.) and feature selection parameter k were explored.

1) Hyperparameter Tuning

Manual Grid Search

- For each classifier and combination of hyperparameters, 5-fold stratified cross-validation was performed on the training set.
- For each fold/setting, ROC-AUC scores were computed using the predicted probabilities.
- The best hyperparameters for each model were selected by average cross-validated ROC-AUC.

Automated Grid Search (GridSearchCV)

- The same parameter grid was provided to scikit-learn’s GridSearchCV along with a pipeline setup.
- 5-fold CV and ROC-AUC were used as the evaluation/scoring metric.
- The estimator with the highest cross-validated ROC-AUC was selected.

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 shape: (441, 46)
------------------------------
```

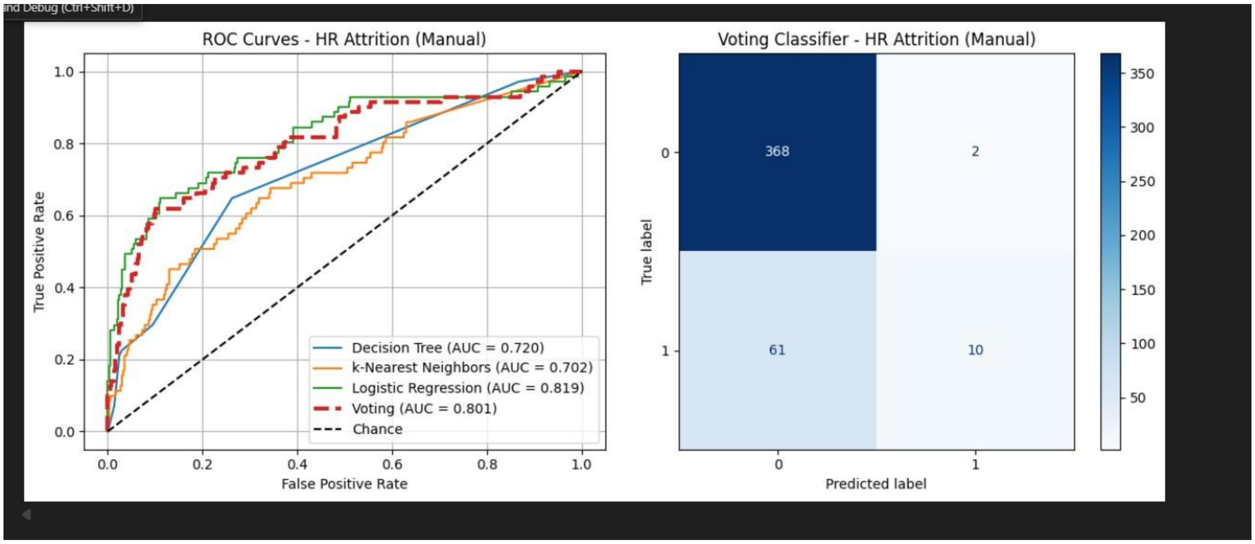
```
============================================================
RUNNING MANUAL GRID SEARCH FOR HR ATTRITION
============================================================
--- Manual Grid Search for Decision Tree ---
Best params for Decision Tree: {'classifier__criterion': 'entropy', 'classifier__max_depth': 3, 'classifier__min_samples_split': 2, 'classifier__min_samples_leaf': 2, 'feature_selection__k': 30}
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
```

```
============================================================
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".




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

--- GridSearchCV for Decision Tree ---
Best params for Decision Tree: {'classifier__criterion': 'entropy', 'classifier__max_depth': 3, 'classifier__min_samples_leaf': 1, 'classifier__min_samples_split': 2, 'feature_selection__k': 30}
Best CV score: 0.7261

--- GridSearchCV for k-Nearest Neighbors ---
Best params for k-Nearest Neighbors: {'classifier__n_neighbors': 11, 'classifier__p': 1, 'classifier__weights': 'distance', 'feature_selection__k': 46}
Best CV score: 0.7305

--- GridSearchCV for Logistic Regression ---
Best params for Logistic Regression: {'classifier__C': 0.1, 'classifier__penalty': 'l2', 'classifier__solver': 'lbfgs', 'feature_selection__k': 46}
Best CV score: 0.8329
```

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.