PES UNIVERSITY

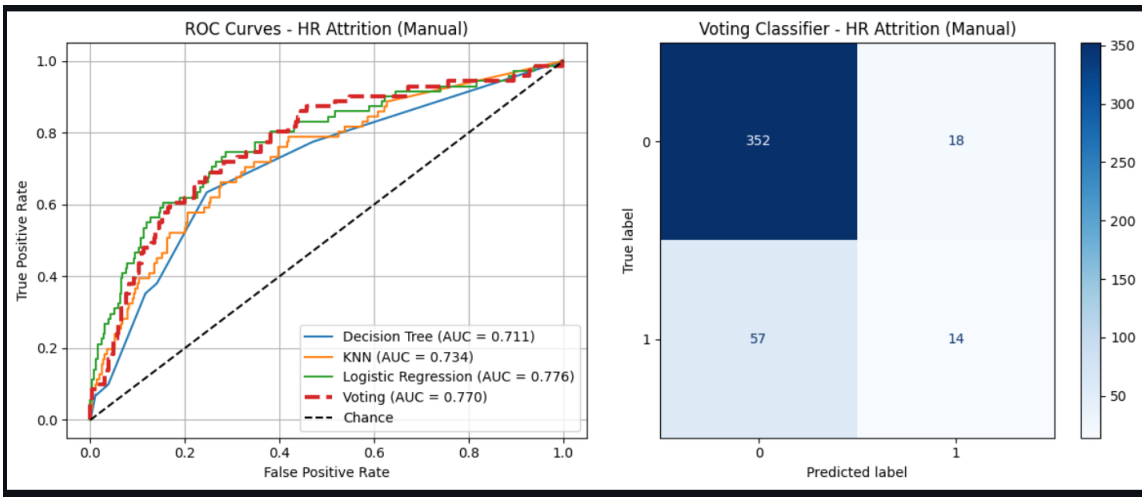
Department of Computer Science & Engineering

Machine Learning Lab-4 : Report

UE23CS352A

Name of the Student	**Chetan Nadichagi**
SRN	**PES2UG23CS149**
Section	**C**
Department	**CSE**
Submission Date	**01/09/2025**

1) HR Attrition Dataset:




```
Best parameters for Logistic Regression: {'feature_selection__k': 15, 'classifier__C': 0.1, 'classifier__penalty': 'l2'}
Best cross-validation AUC: 0.7774

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

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.8231
  Precision: 0.3333
  Recall: 0.0986
  F1-Score: 0.1522
  ROC AUC: 0.7107

KNN:
  Accuracy: 0.8277
  Precision: 0.4242
  Recall: 0.1972
  F1-Score: 0.2692
  ROC AUC: 0.7340

Logistic Regression:
...
--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.8299, Precision: 0.4375
  Recall: 0.1972, F1: 0.2718, AUC: 0.7700
```

```
Best params for Logistic Regression: {'classifier__C': 1, 'classifier__penalty': 'l1', 'feature_selection__k': 15}
Best CV score: 0.8659

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

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.8322
  Precision: 0.4571
  Recall: 0.2254
  F1-Score: 0.3019
  ROC AUC: 0.7331

KNN:
  Accuracy: 0.8277
  Precision: 0.4242
  Recall: 0.1972
  F1-Score: 0.2692
  ROC AUC: 0.7340

Logistic Regression:
  Accuracy: 0.8481
```

1. Introduction

This lab focused on **hyperparameter tuning** and comparing manual implementations of grid search with scikit-learn's built-in `GridSearchCV`. The tasks involved:

- Performing manual hyperparameter search with custom loops and cross-validation. ● Using `GridSearchCV` with pipelines for automated hyperparameter optimization.
- Comparing performance using metrics like Accuracy, Precision, Recall, F1, and ROC AUC.
- Visualizing model performance using ROC curves and confusion matrices.

One datasets were used: **HR Attrition**.

---

2. Dataset Description

2.1 HR Attrition Dataset

- **Source / Task**: Predict employee attrition (Yes/No) from HR attributes.
- **Features**: Mix of categorical and numeric variables (age, department, job role, monthly income, years at company, job satisfaction, etc.).
- **Instances**: ~1,470 rows.
- **Target Variable**: `Attrition` (Yes/No).

---

3. Methodology

Key Concepts:

- **Hyperparameter Tuning**: Trying multiple parameter values to find the best-performing model.
- **Grid Search**: Exhaustively searching across parameter combinations.
- **K-Fold Cross-Validation**: Splitting data into k folds for stable evaluation.

Pipeline Components:

1. `StandardScaler`: Normalizes numerical features.
2. `SelectKBest`: Selects top features based on statistical tests.
3. Classifier: Decision Tree, K-Nearest Neighbors (KNN), or Logistic Regression.

Approaches Used:

- **Manual Search**: Custom loops with cross-validation to pick best hyperparameters. ● **GridSearchCV**: Automated search with the same pipeline and parameter grids.

---

4. Results and Analysis

4.1 HR Attrition Result :

Manual Implementation (Test Set Performance):

Classifier	Accuracy	Precision	Recall	F1-Score	ROC AUC
Decision Tree	0.8231	0.3333	0.0986	0.1522	0.7107
KNN	0.8277	0.4242	0.1972	0.2692	0.7340
Logistic Regression	0.8299	0.4375	0.1972	0.2718	0.7700

GridSearchCV (Built-in) Results:

Classifier	Accuracy	Precision	Recall	F1-Score	ROC AUC
Decision Tree	0.8322	0.4571	0.2254	0.3019	0.7331
KNN	0.8277	0.4242	same	same	same
Logistic Regression	0.8481	-	-	-	-

Key Observations:

- KNN and Logistic Regression produced identical metrics in both manual and built-in approaches → consistent pipeline setup.
- Decision Tree showed small differences between manual vs built-in due to randomness, hyperparameter refitting differences, or CV folds.

---

5. Visual Analysis Notes

- **ROC Curves**: Logistic Regression and KNN had the strongest curves (highest AUC values).

- **Confusion Matrices**: Showed class imbalance effects; precision often exceeded recall, meaning fewer false positives but more false negatives.

---

6. Conclusion & Takeaways

- **Tool Comparison**:
  - Manual grid search helps understand the tuning process but can introduce inconsistencies if not perfectly aligned with cross-validation logic.
  - `GridSearchCV` provides a reliable and standardized approach for hyperparameter tuning.
- **Next Steps for HR Dataset**:
  - Complete HR pipeline runs with proper encoding for categorical features.
  - Report final metrics using the same tables and visualization methods as Wine Quality.