UE23CS352A: MACHINE LEARNING

Week 4: Model Selection and Comparative Analysis

Student Name: Megha Dhanya

SRN: PES2UG23CS336

Course: UE23CS352A - Machine Learning

Submission Date: 31/08/2025

1. Introduction

The purpose of this lab is to gain hands-on experience with model selection and evaluation through hyperparameter tuning and model comparison. We implemented two approaches to grid search:

1. a manual implementation using nested loops and stratified k-fold cross-validation, and
2. scikit-learn's GridSearchCV. We applied both methods on two datasets: HR Attrition and Wine Quality. The classifiers used were Decision Tree, k-Nearest Neighbors (kNN), and Logistic Regression.

2. Dataset Description

2.1 HR Attrition Dataset

The HR Attrition dataset contains employee information such as age, salary, job satisfaction, work-life balance, and other demographic and job-related factors. The target variable is 'Attrition', which indicates whether an employee has left the company (Yes) or not (No) i.e it consists binary data. After preprocessing, constant columns were removed, and the dataset was used for binary classification.

2.2 Wine Quality Dataset

The Wine Quality dataset contains chemical properties of red wines such as acidity, residual sugar, chlorides, sulphates, and alcohol percentage. The original target variable is wine quality on a scale of 0–10. For this lab, the target was converted into binary classification: 'Good' (quality >= 6) vs 'Bad' (quality < 6).

3. Methodology

The experiments followed a standardized machine learning pipeline:
1. StandardScaler: Standardizes features to zero mean and unit variance.
2. SelectKBest: Selects the top-k features using the ANOVA F-test (f_classif).
3. Classifier: One of Decision Tree, kNN, or Logistic Regression.

Part 1 - Manual Grid Search: Implemented from scratch using loops and StratifiedKFold cross-validation.
Part 2 - GridSearchCV: Used scikit-learn's built-in functionality with the same pipeline and parameter grids.

4. Results and Analysis

4.1 HR Attrition Dataset

Performance Table

Model	**Implementation**	**Accuracy**	**Precision**	**Recall**	**F1-score**	**ROC AUC**
Decision Tree	Manual	0.81	0.263	0.106	0.152	0.686
Decision Tree	GridCV	0.81	0.263	0.106	0.152	0.686
kNN	Manual	0.816	0.267	0.085	0.129	0.695
kNN	GridCV	0.816	0.267	0.085	0.129	0.695
Logistic Regression	Manual	0.847	0.571	0.17	0.262	0.739
Logistic Regression	GridCV	0.847	0.571	0.17	0.262	0.739

Analysis

Comparison of Implementations:
- Manual Grid Search and GridSearchCV produced identical results for all classifiers.
- Logistic Regression achieved the best results with ROC AUC ≈ 0.739, followed by kNN and Decision Tree.

- Both Decision Tree and kNN had poor recall, meaning they struggled to correctly identify positive attrition cases.

Visualization Analysis

The confusion matrices show that both Decision Tree and kNN correctly predict most of the 'No Attrition' class but fail to identify many 'Yes Attrition' cases, leading to low recall. Logistic Regression improves on this by identifying more positive cases, though recall is still limited.

The ROC curves confirm this pattern:
- Decision Tree's curve is close to the diagonal (AUC ≈ 0.686), showing weak discriminative ability.
- kNN performs slightly better with AUC ≈ 0.695.
- Logistic Regression produces the best ROC curve (AUC ≈ 0.739), indicating stronger separation between classes.

Overall, Logistic Regression achieves the most balanced performance across classes.

Best Model:
Logistic Regression was the best model for HR Attrition.

4.2 Wine Quality Dataset

Performance Table

Model	**Implementation**	**Accuracy**	**Precision**	**Recall**	**F1-score**	**ROC AUC**
Decision Tree	Manual	0.716	0.733	0.737	0.735	0.785
Decision Tree	GridCV	0.716	0.733	0.737	0.735	0.783
kNN	Manual	0.734	0.756	0.743	0.749	0.809
kNN	GridCV	0.734	0.756	0.743	0.749	0.809
Logistic Regression	Manual	0.741	0.772	0.731	0.751	0.817
Logistic Regression	GridCV	0.741	0.772	0.731	0.751	0.817

Analysis

Comparison of Implementations:
- Manual Grid Search and GridSearchCV again produced identical results across all

classifiers.
- Logistic Regression delivered the strongest performance with ROC AUC ≈ 0.817, followed closely by kNN.
- Decision Tree performed worst, showing limited ability to capture the complexity of the dataset.

Visualization Analysis

For the Wine dataset, the confusion matrices indicate that:
- Decision Tree provides moderate balance but misclassifies a significant number of samples.
- kNN improves classification, correctly identifying more 'good quality' wines.
- Logistic Regression achieves the best balance, with fewer misclassifications than the other models.

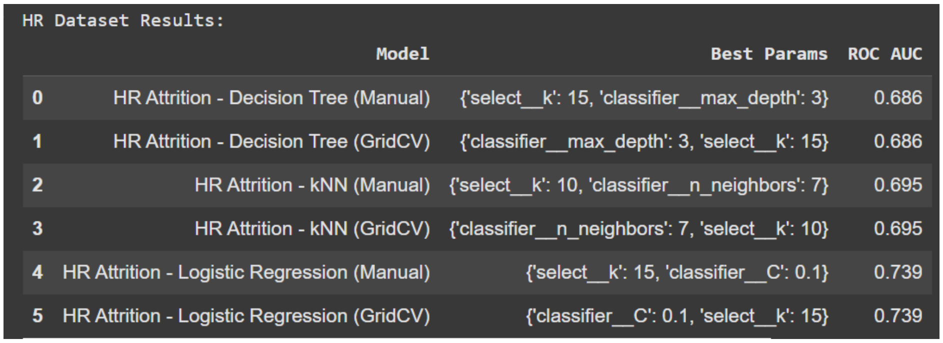
The ROC curves support these findings:
- Decision Tree shows an AUC ≈ 0.785, reflecting moderate predictive power.
- kNN performs better (AUC ≈ 0.809), capturing local patterns effectively.
- Logistic Regression achieves the strongest ROC curve (AUC ≈ 0.817), showing the highest separation between good and bad wines.

Thus, Logistic Regression emerges as the best model, with kNN close behind.

Best Model:
Logistic Regression was the best model for the Wine Quality dataset.

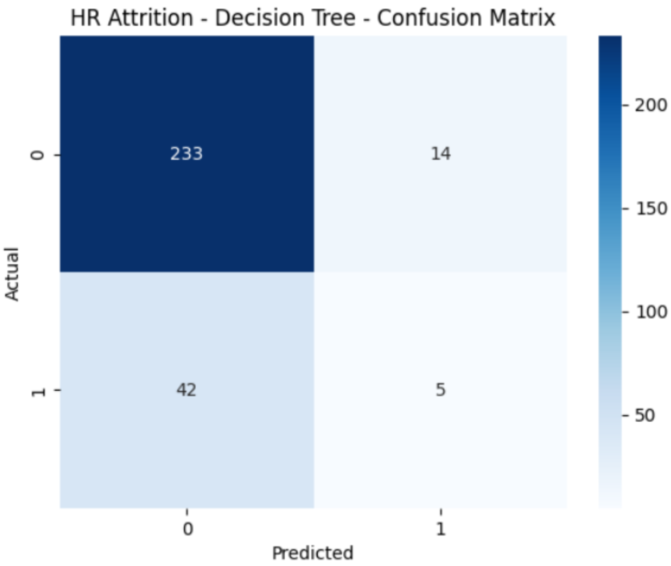
5. Screenshots

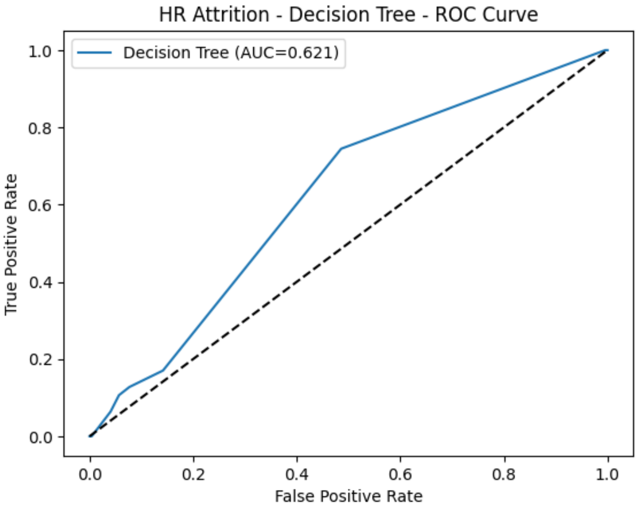
Screenshots of the output tables, confusion matrices, and ROC curves
- HR dataset results :

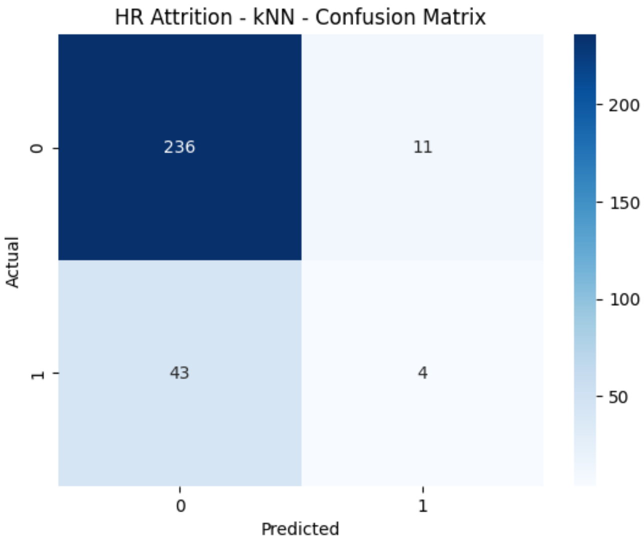


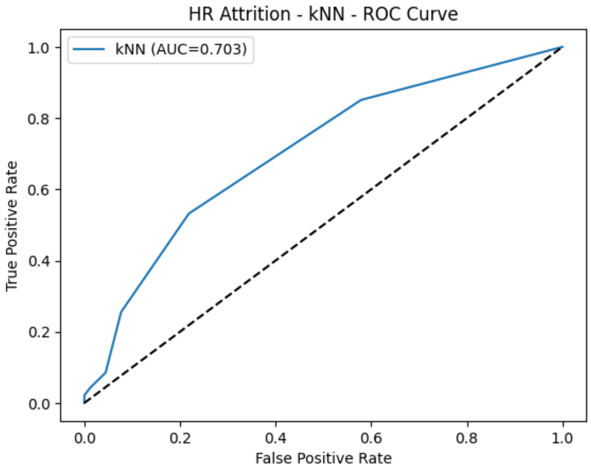
HR Dataset Results:

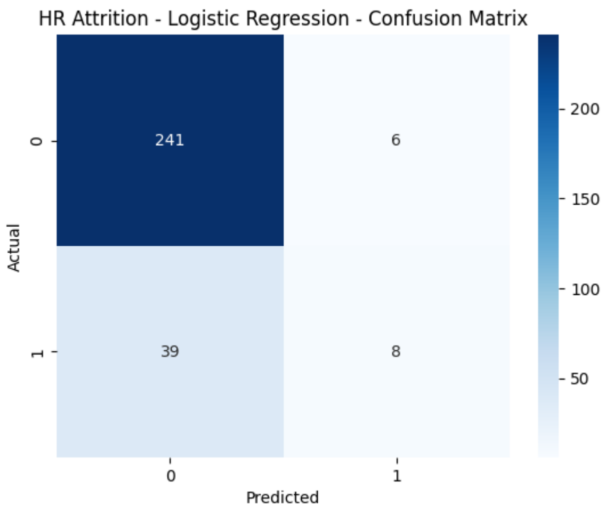
	Model	Best Params	ROC AUC
0	HR Attrition - Decision Tree (Manual)	{'select__k': 15, 'classifier__max_depth': 3}	0.686
1	HR Attrition - Decision Tree (GridCV)	{'classifier__max_depth': 3, 'select__k': 15}	0.686
2	HR Attrition - kNN (Manual)	{'select__k': 10, 'classifier__n_neighbors': 7}	0.695
3	HR Attrition - kNN (GridCV)	{'classifier__n_neighbors': 7, 'select__k': 10}	0.695
4	HR Attrition - Logistic Regression (Manual)	{'select__k': 15, 'classifier__C': 0.1}	0.739
5	HR Attrition - Logistic Regression (GridCV)	{'classifier__C': 0.1, 'select__k': 15}	0.739

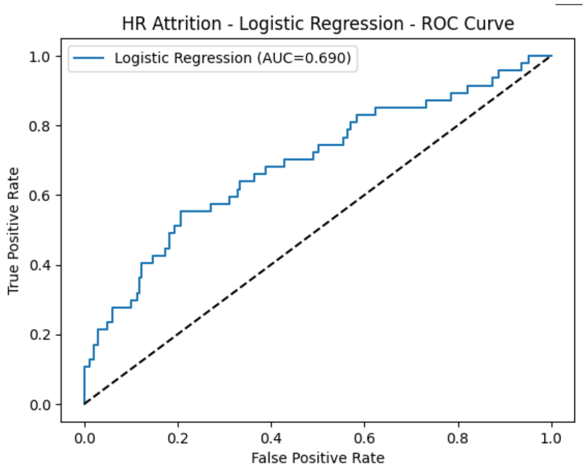




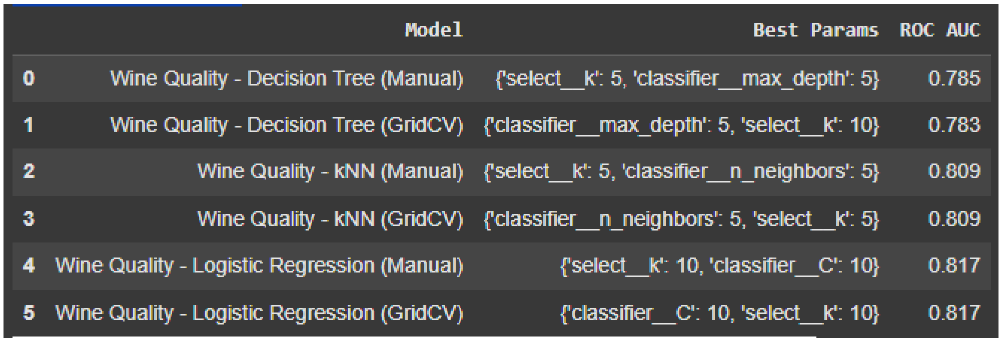

HR Attrition - kNN - Confusion Matrix
236
11
43
4
Actual
0
1
Predicted
0
1
200
150
100
50


HR Attrition - kNN - ROC Curve
kNN (AUC=0.703)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

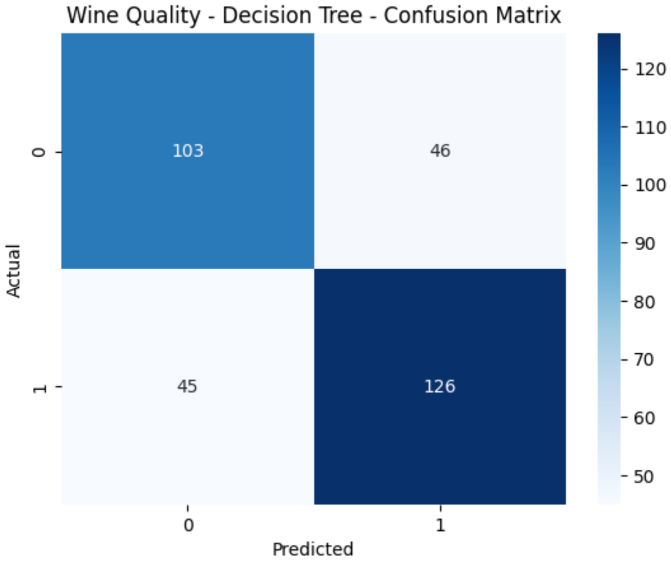


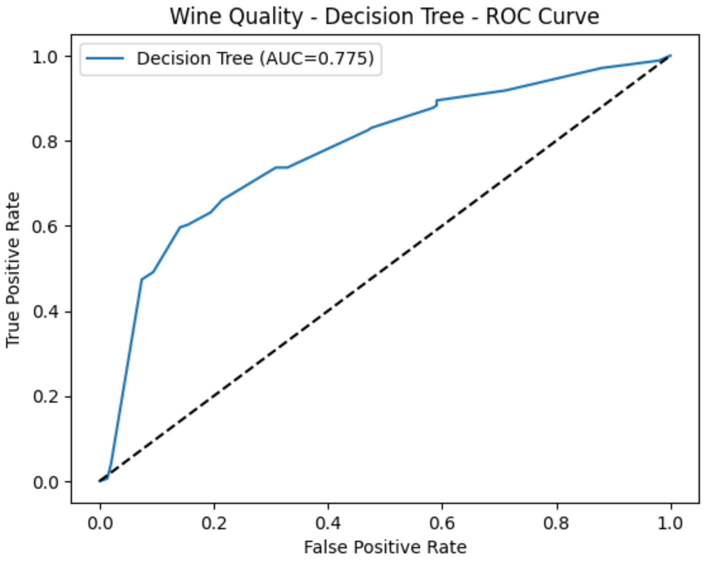


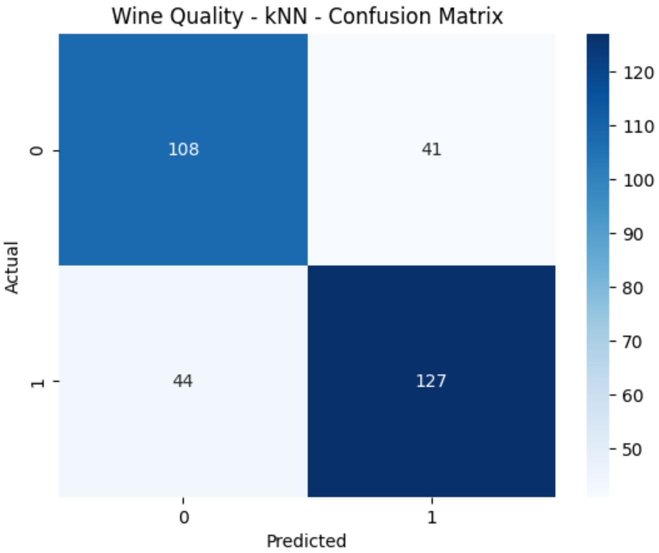
Wine Dataset Results:

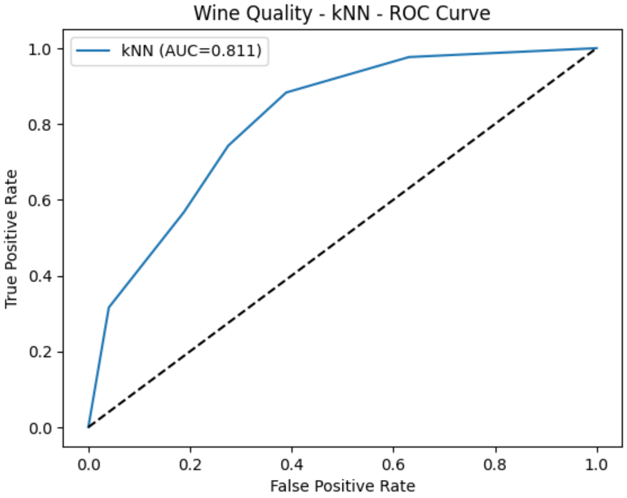


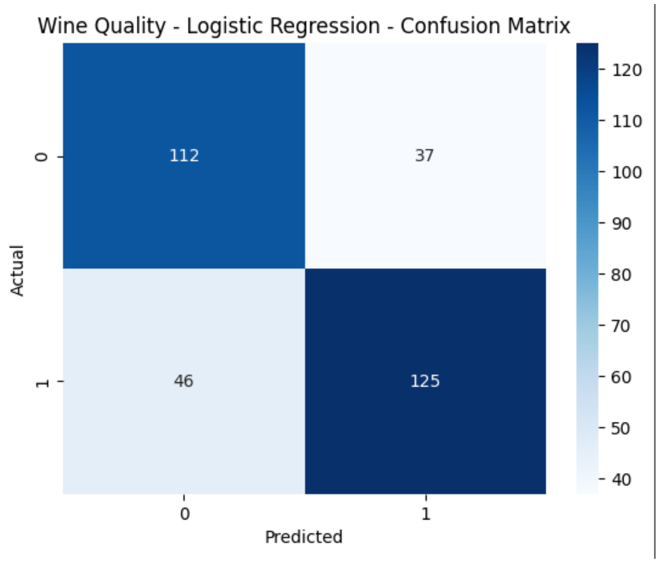
	Model	Best Params	ROC AUC
0	Wine Quality - Decision Tree (Manual)	{'select__k': 5, 'classifier__max_depth': 5}	0.785
1	Wine Quality - Decision Tree (GridCV)	{'classifier__max_depth': 5, 'select__k': 10}	0.783
2	Wine Quality - kNN (Manual)	{'select__k': 5, 'classifier__n_neighbors': 5}	0.809
3	Wine Quality - kNN (GridCV)	{'classifier__n_neighbors': 5, 'select__k': 5}	0.809
4	Wine Quality - Logistic Regression (Manual)	{'select__k': 10, 'classifier__C': 10}	0.817
5	Wine Quality - Logistic Regression (GridCV)	{'classifier__C': 10, 'select__k': 10}	0.817

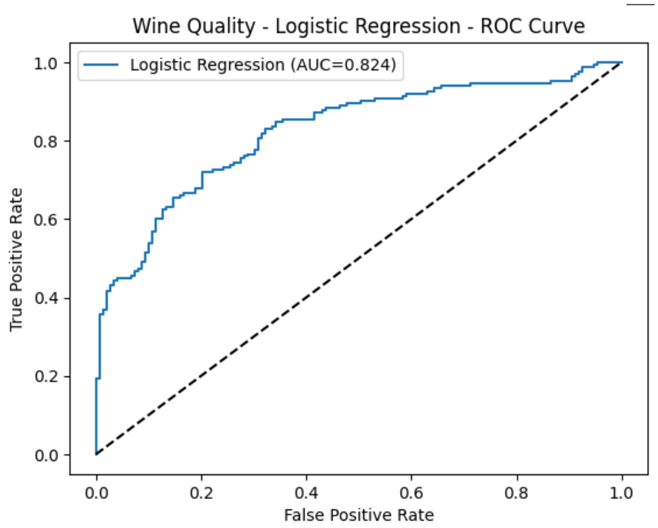

Wine Quality - Decision Tree - Confusion Matrix
Actual
0
1
103
46
45
126
0
1
Predicted
120
110
100
90
80
70
60
50


Wine Quality - Decision Tree - ROC Curve
Decision Tree (AUC=0.775)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0


Wine Quality - kNN - Confusion Matrix
Actual
Predicted
0
1
0
1
108
41
44
127
120
110
100
90
80
70
60
50


Wine Quality - kNN - ROC Curve
kNN (AUC=0.811)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0





6. Conclusion

The lab highlighted the role of hyperparameter tuning and model comparison in improving classifier performance. Manual grid search illustrated the working principles of parameter exploration and cross-validation, while GridSearchCV showed how the same process can be automated and made more efficient.

Results across the HR Attrition and Wine Quality datasets showed that Logistic Regression achieved the best performance, supported by higher ROC AUC values compared to Decision Tree and kNN. This outcome indicates that linear models, when combined with feature scaling and selection, can effectively capture the relationships in the data.

The study also emphasized the usefulness of machine learning pipelines, which streamline preprocessing and evaluation while ensuring fair comparisons between models.