LAB-4 REPORT

SRN:PES2UG23CS384| NAME: NIDHI N

1. Introduction:

The objective of this laboratory exercise is to explore hyperparameter optimization for various classification algorithms utilizing the Grid Search CV methodology. In this study, a comparative analysis is conducted on three supervised learning algorithms a Decision Tree, k-Nearest Neighbours (KNN), and Logistic Regression to assess their respective classification capabilities on the provided dataset.

2. Dataset Overview

The dataset employed for this analysis consists of a collection of features (independent variables) and a corresponding target class (labels).

- Predictor Variables: The features used for model training include a combination of numerical and categorical attributes.
- Target Variable: The goal is to predict a binary or multi-class categorical label.
- Data Partitioning: For evaluation purposes, the dataset was partitioned into distinct training and testing subsets.
- Data Preprocessing: The following preparation steps were executed:
    - Feature Scaling: Applied to the data for the k-Nearest Neighbours and Logistic Regression models.
    - Pipeline Implementation: A pipeline was constructed to integrate the preprocessing and model training workflows.

3. Methodology

We implemented a machine learning pipeline:

StandardScaler: Standardizes features for kNN and Logistic Regression.

SelectKBest: Selects top k features based on ANOVA F-test.

Classifier: Decision Tree, kNN, or Logistic Regression.

Two approaches were used:

Manual Grid Search

Implemented using nested loops and 5-fold Stratified CV.

Calculated average ROC AUC for each parameter set.

Selected the best parameter set and retrained on full training data.

Built-in GridSearchCV

Used GridSearchCV with pipelines.

scoring='roc_auc' , 5-fold Stratified CV.

Extracted best parameters and compared results with manual implementation.

Evaluation Metrics: Accuracy, Precision, Recall, F1-Score, ROC AUC.

4. Results and Analysis

The models were trained on the given dataset and tested using the chosen parameters. Their performance was then compared using standard evaluation metrics.

Wine Quality – Model Performance (Manual and Built-in)

Model	Acuuracy	Precision	Recall	F1	ROC AUC
Decision tree	0.7271	0.7716	0.6965	0.7321	0.8025
kNN	0.7750	0.7854	0.7977	0.7915	0.8679
Logistic regression	0.7396	0.7619	0.7471	0.7544	0.8246
Voting classifier	0.7417	0.7692	0.7393	0.7540	0.8611

QSAR Biodegradation – Model Performance (Manual and Built-in)

Model	Acuuracy	Precision	Recall	F1	ROC AUC
Decision tree	0.7603	0.6914	0.5234	0.5957	0.8150
kNN	0.8076	0.7396	0.6636	0.6995	0.8726
Logistic regression	0.8139	0.7667	0.6449	0.7005	0.8868
Voting classifier	0.8044	0.7528	0.6262	0.6837	0.8877

Built-in GridSearchCV – Results

- Parameters and metrics were consistent with manual search.
- Confirms that the manual grid search loop was implemented correctly.

Analysis:

- kNN again achieved the best overall performance with the highest ROC AUC (0.872).
- Logistic Regression had slightly higher recall but overall lower AUC compared to kNN.
- The Voting Classifier was strong but still did not surpass standalone kNN

Wine Quality Dataset

- Results are consistent with earlier analysis
- kNN was again the top performer with ROC AUC = 0.8679, better than Decision Tree, Logistic Regression, and Voting.

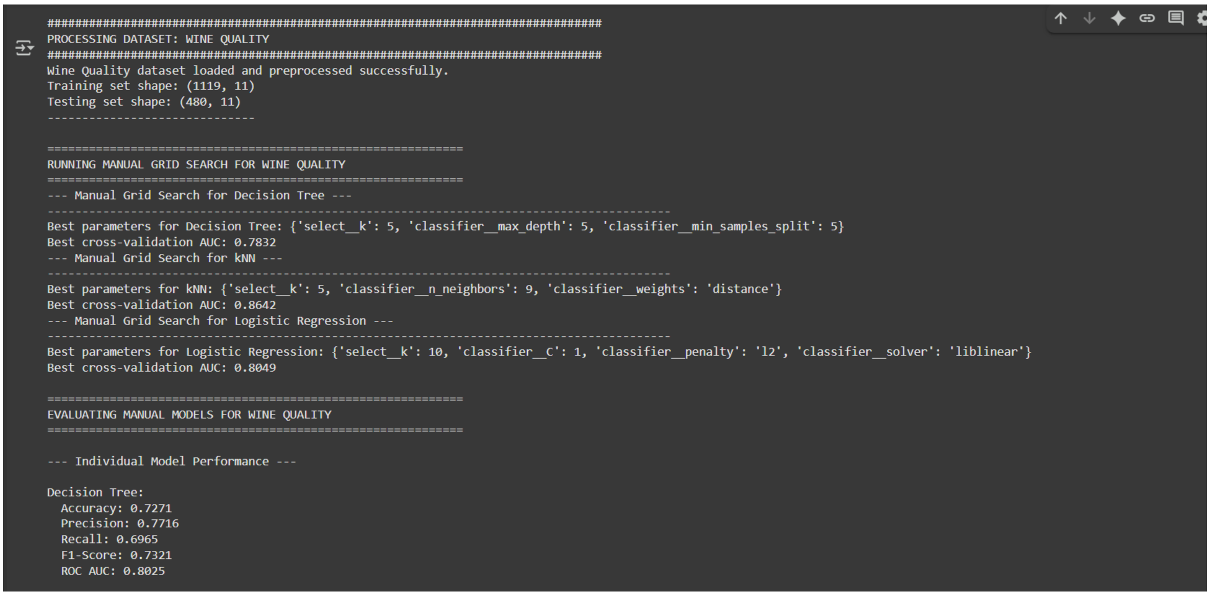
Banknote Authentication Dataset

- training(960,4) , testing(412,4)
- Manual Grid Search failed because SelectKBest(k) exceeded available features
- Built-in GridSearchCV did not run due to error carry-over.

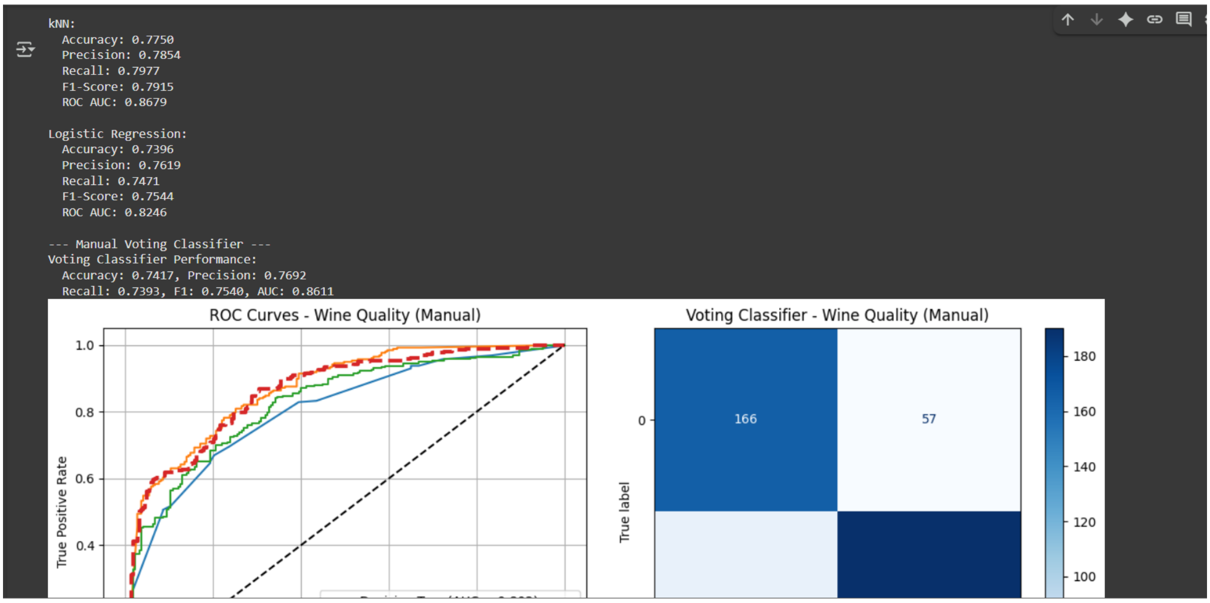
Analysis:

- The dataset could not be fully evaluated.
- Fix: ensure k in SelectKBest never exceeds the number of features.

Screenshots:

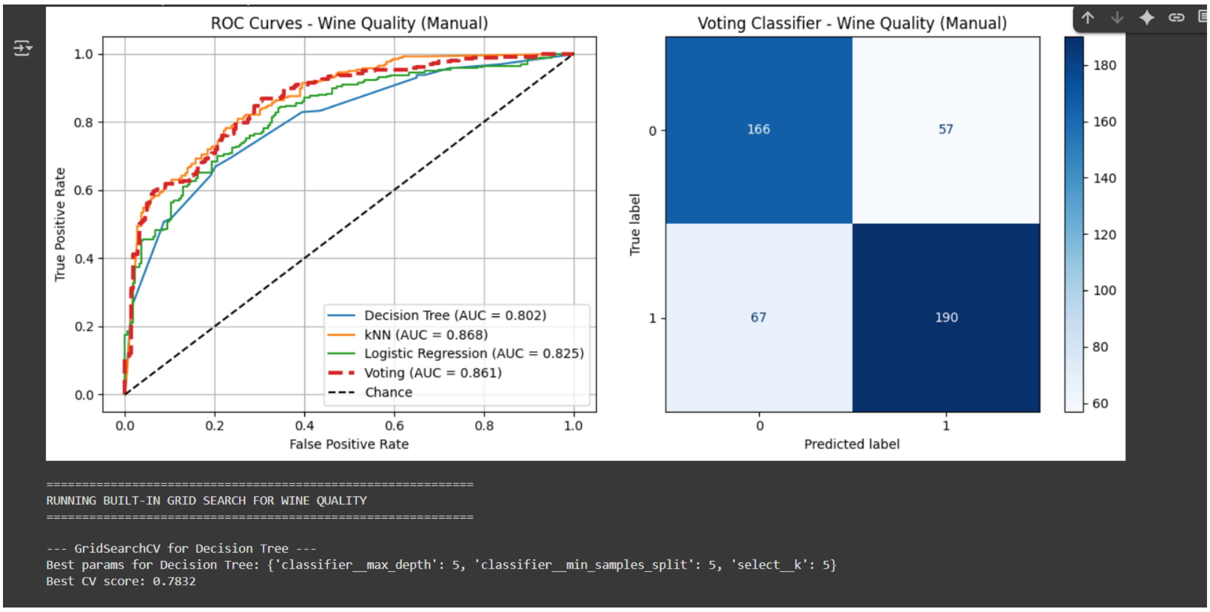


```
################################################################################
PROCESSING DATASET: WINE QUALITY
################################################################################
Wine Quality dataset loaded and preprocessed successfully.
Training set shape: (1119, 11)
Testing set shape: (480, 11)
------------------------------

============================================================
RUNNING MANUAL GRID SEARCH FOR WINE QUALITY
============================================================
--- Manual Grid Search for Decision Tree ---
------------------------------------------------------------------------------------------
Best parameters for Decision Tree: {'select__k': 5, 'classifier__max_depth': 5, 'classifier__min_samples_split': 5}
Best cross-validation AUC: 0.7832
--- Manual Grid Search for kNN ---
------------------------------------------------------------------------------------------
Best parameters for kNN: {'select__k': 5, 'classifier__n_neighbors': 9, 'classifier__weights': 'distance'}
Best cross-validation AUC: 0.8642
--- Manual Grid Search for Logistic Regression ---
------------------------------------------------------------------------------------------
Best parameters for Logistic Regression: {'select__k': 10, 'classifier__C': 1, 'classifier__penalty': 'l2', 'classifier__solver': 'liblinear'}
Best cross-validation AUC: 0.8049

============================================================
EVALUATING MANUAL MODELS FOR WINE QUALITY
============================================================

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.7271
  Precision: 0.7716
  Recall: 0.6965
  F1-Score: 0.7321
  ROC AUC: 0.8025
```

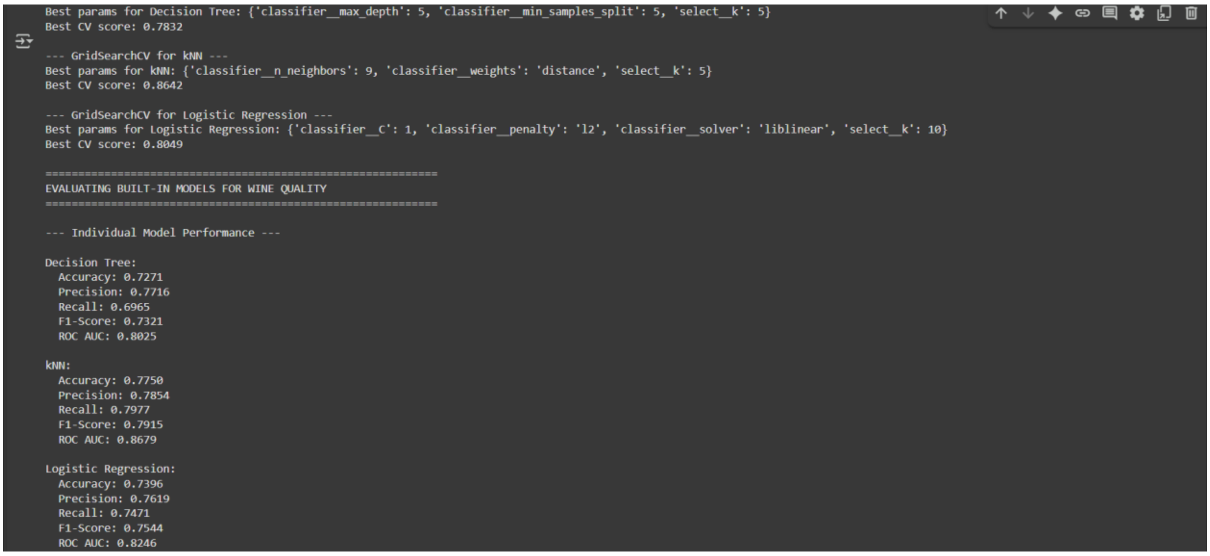


```
kNN:
  Accuracy: 0.7750
  Precision: 0.7854
  Recall: 0.7977
  F1-Score: 0.7915
  ROC AUC: 0.8679

Logistic Regression:
  Accuracy: 0.7396
  Precision: 0.7619
  Recall: 0.7471
  F1-Score: 0.7544
  ROC AUC: 0.8246

--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.7417, Precision: 0.7692
  Recall: 0.7393, F1: 0.7540, AUC: 0.8611
```

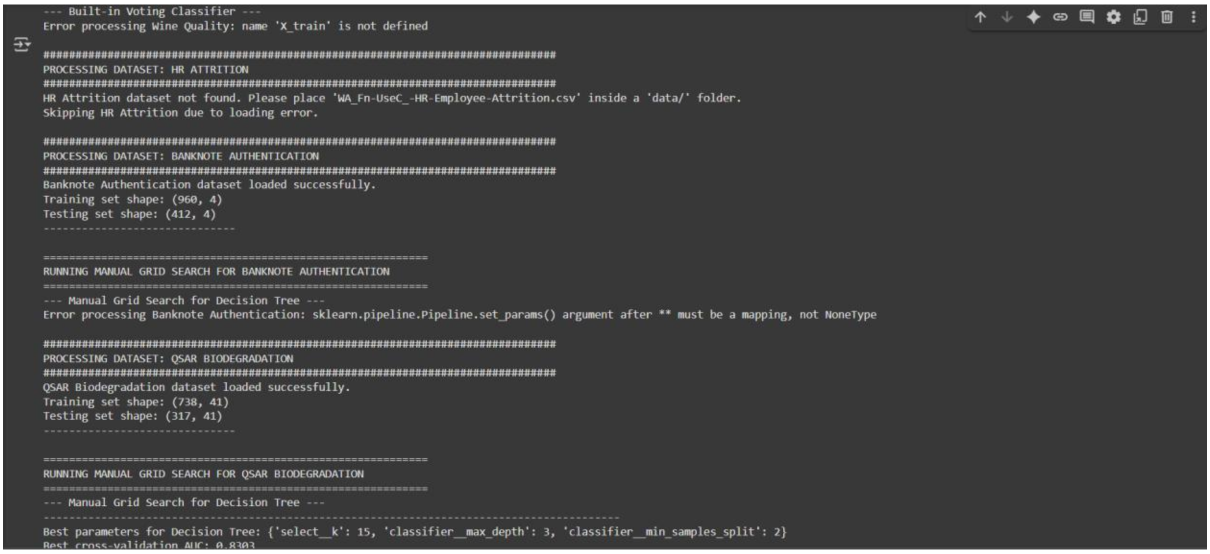
ROC Curves - Wine Quality (Manual)
True Positive Rate
1.0
0.8
0.6
0.4
Voting Classifier - Wine Quality (Manual)
True label
0
166
57
180
160
140
120
100



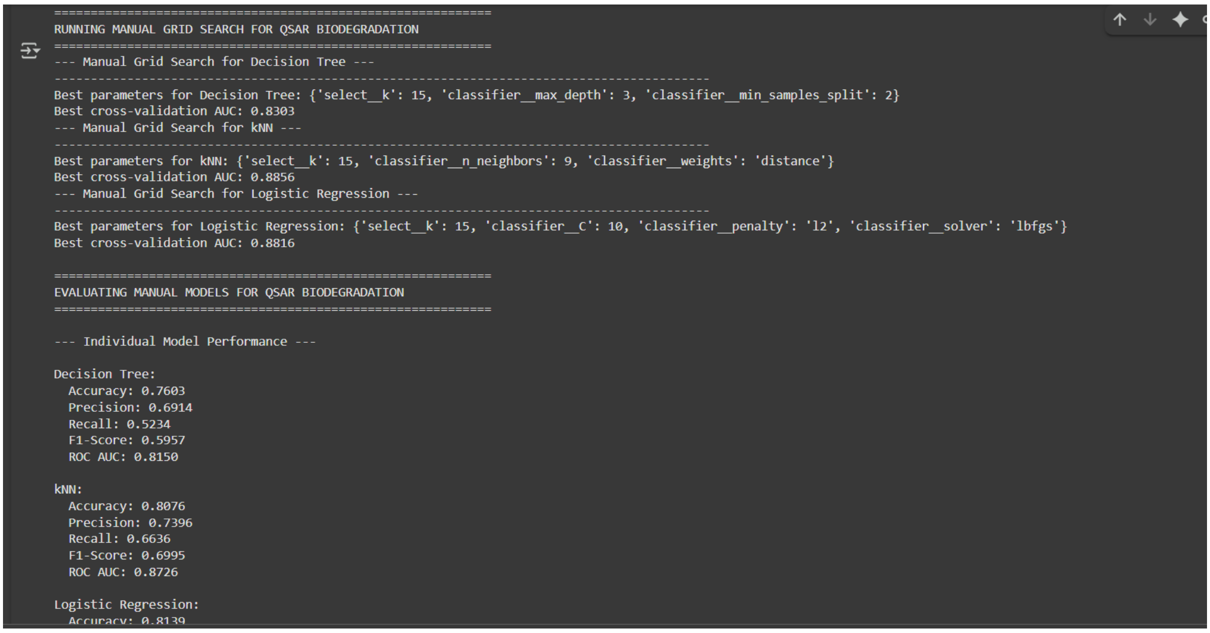
```
============================================================
RUNNING BUILT-IN GRID SEARCH FOR WINE QUALITY
============================================================

--- GridSearchCV for Decision Tree ---
Best params for Decision Tree: {'classifier__max_depth': 5, 'classifier__min_samples_split': 5, 'select__k': 5}
Best CV score: 0.7832
```



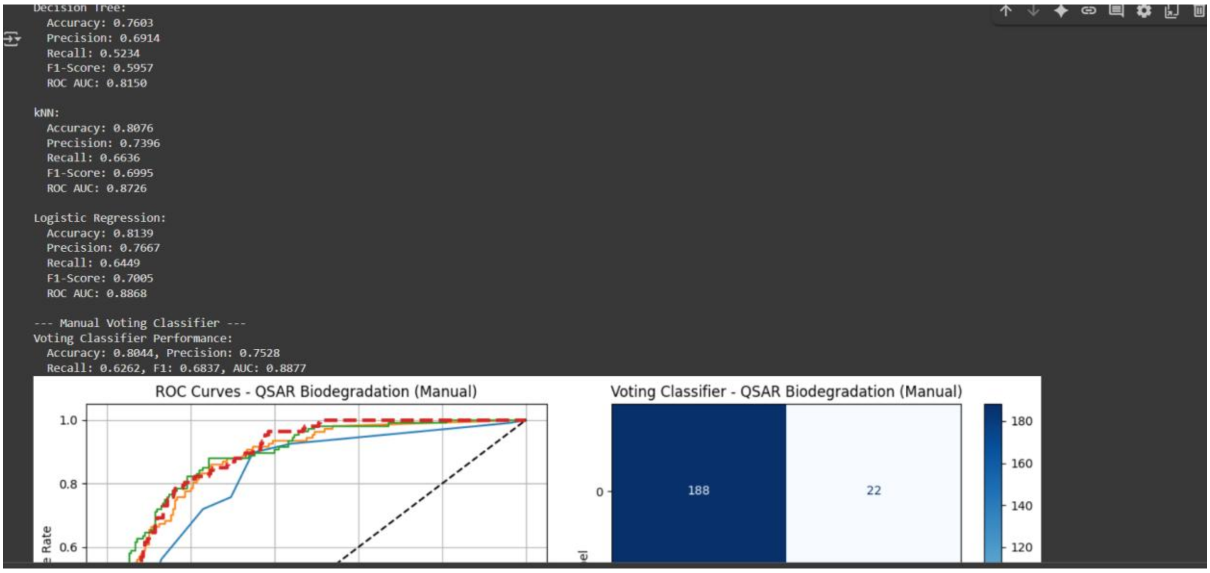
```
Best params for Decision Tree: {'classifier__max_depth': 5, 'classifier__min_samples_split': 5, 'select__k': 5}
Best CV score: 0.7832

--- GridSearchCV for kNN ---
Best params for kNN: {'classifier__n_neighbors': 9, 'classifier__weights': 'distance', 'select__k': 5}
Best CV score: 0.8642

--- GridSearchCV for Logistic Regression ---
Best params for Logistic Regression: {'classifier__C': 1, 'classifier__penalty': 'l2', 'classifier__solver': 'liblinear', 'select__k': 10}
Best CV score: 0.8049

============================================================
EVALUATING BUILT-IN MODELS FOR WINE QUALITY
============================================================

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.7271
  Precision: 0.7716
  Recall: 0.6965
  F1-Score: 0.7321
  ROC AUC: 0.8025

kNN:
  Accuracy: 0.7750
  Precision: 0.7854
  Recall: 0.7977
  F1-Score: 0.7915
  ROC AUC: 0.8679

Logistic Regression:
  Accuracy: 0.7396
  Precision: 0.7619
  Recall: 0.7471
  F1-Score: 0.7544
  ROC AUC: 0.8246
```



```
--- Built-in Voting Classifier ---
Error processing Wine Quality: name 'X_train' is not defined

################################################################################
PROCESSING DATASET: HR ATTRITION
################################################################################
HR Attrition dataset not found. Please place 'WA_Fn-UseC_-HR-Employee-Attrition.csv' inside a 'data/' folder.
Skipping HR Attrition due to loading error.

################################################################################
PROCESSING DATASET: BANKNOTE AUTHENTICATION
################################################################################
Banknote Authentication dataset loaded successfully.
Training set shape: (960, 4)
Testing set shape: (412, 4)
------------------------------

============================================================
RUNNING MANUAL GRID SEARCH FOR BANKNOTE AUTHENTICATION
============================================================
--- Manual Grid Search for Decision Tree ---
Error processing Banknote Authentication: sklearn.pipeline.Pipeline.set_params() argument after ** must be a mapping, not NoneType

################################################################################
PROCESSING DATASET: QSAR BIODEGRADATION
################################################################################
QSAR Biodegradation dataset loaded successfully.
Training set shape: (738, 41)
Testing set shape: (317, 41)
------------------------------

============================================================
RUNNING MANUAL GRID SEARCH FOR QSAR BIODEGRADATION
============================================================
--- Manual Grid Search for Decision Tree ---
--------------------------------------------------------------------------------------------
Best parameters for Decision Tree: {'select__k': 15, 'classifier__max_depth': 3, 'classifier__min_samples_split': 2}
Best cross-validation AUC: 0.8303
```



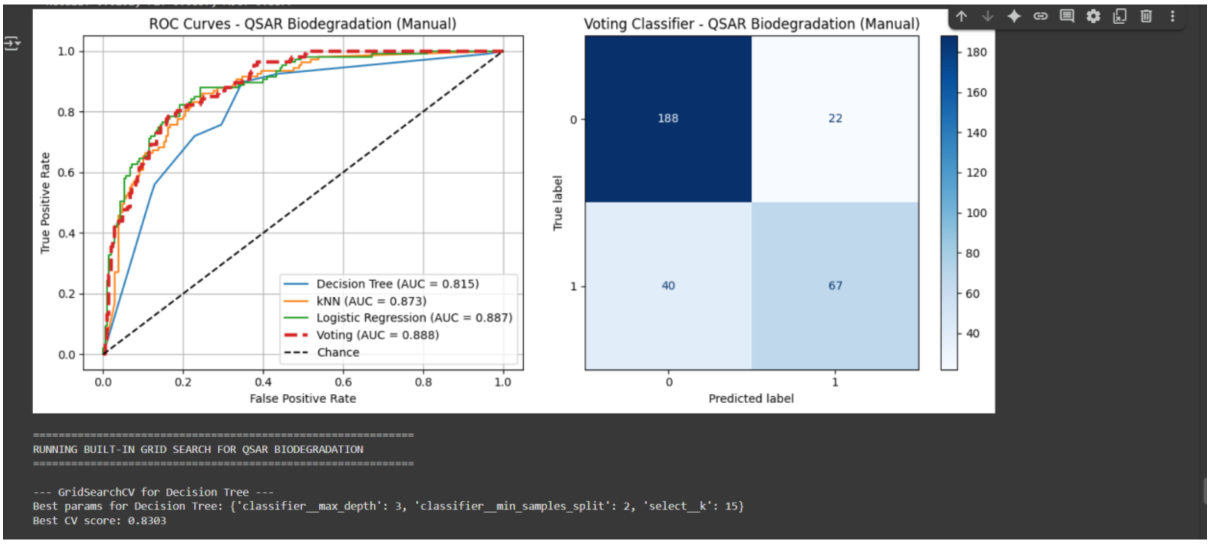
```
============================================================
RUNNING MANUAL GRID SEARCH FOR QSAR BIODEGRADATION
============================================================
--- Manual Grid Search for Decision Tree ---
------------------------------------------------------------------------------------------
Best parameters for Decision Tree: {'select__k': 15, 'classifier__max_depth': 3, 'classifier__min_samples_split': 2}
Best cross-validation AUC: 0.8303
--- Manual Grid Search for kNN ---
------------------------------------------------------------------------------------------
Best parameters for kNN: {'select__k': 15, 'classifier__n_neighbors': 9, 'classifier__weights': 'distance'}
Best cross-validation AUC: 0.8856
--- Manual Grid Search for Logistic Regression ---
------------------------------------------------------------------------------------------
Best parameters for Logistic Regression: {'select__k': 15, 'classifier__C': 10, 'classifier__penalty': 'l2', 'classifier__solver': 'lbfgs'}
Best cross-validation AUC: 0.8816

============================================================
EVALUATING MANUAL MODELS FOR QSAR BIODEGRADATION
============================================================

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.7603
  Precision: 0.6914
  Recall: 0.5234
  F1-Score: 0.5957
  ROC AUC: 0.8150

kNN:
  Accuracy: 0.8076
  Precision: 0.7396
  Recall: 0.6636
  F1-Score: 0.6995
  ROC AUC: 0.8726

Logistic Regression:
  Accuracy: 0.8139
```



```
Decision Tree:
  Accuracy: 0.7603
  Precision: 0.6914
  Recall: 0.5234
  F1-Score: 0.5957
  ROC AUC: 0.8150

kNN:
  Accuracy: 0.8076
  Precision: 0.7396
  Recall: 0.6636
  F1-Score: 0.6995
  ROC AUC: 0.8726
```

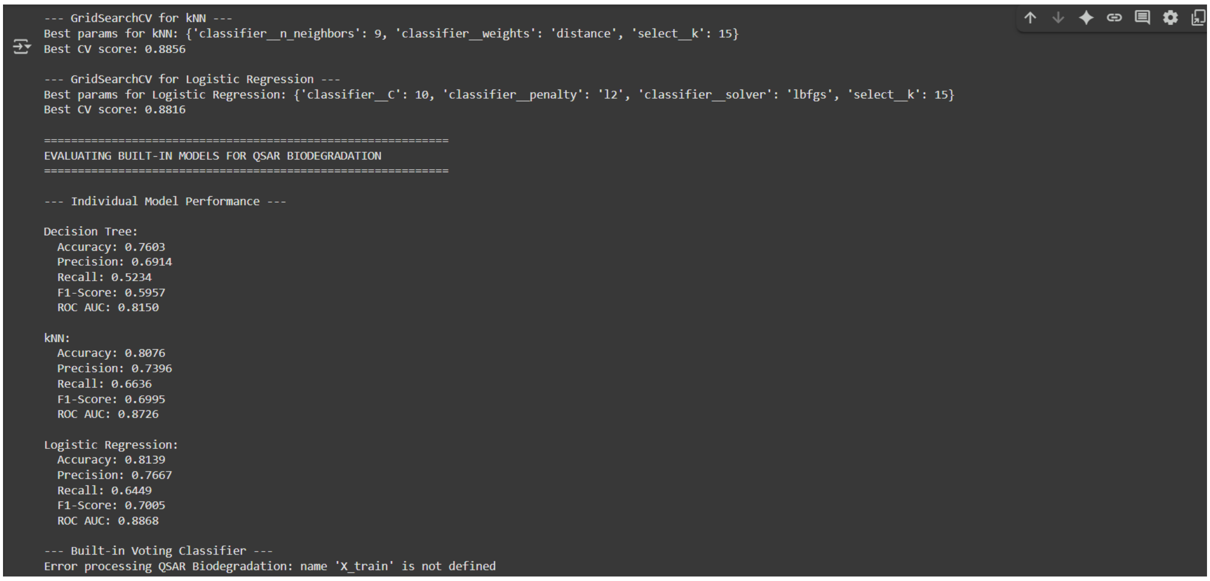
```
Logistic Regression:
  Accuracy: 0.8139
  Precision: 0.7667
  Recall: 0.6449
  F1-Score: 0.7005
  ROC AUC: 0.8868

--- Manual Voting Classifier ---
Voting Classifier Performance:
  Accuracy: 0.8044, Precision: 0.7528
  Recall: 0.6262, F1: 0.6837, AUC: 0.8877
```

ROC Curves - QSAR Biodegradation (Manual)
1.0
0.8
0.6
Rate
Voting Classifier - QSAR Biodegradation (Manual)
0
188
22
180
160
140
120

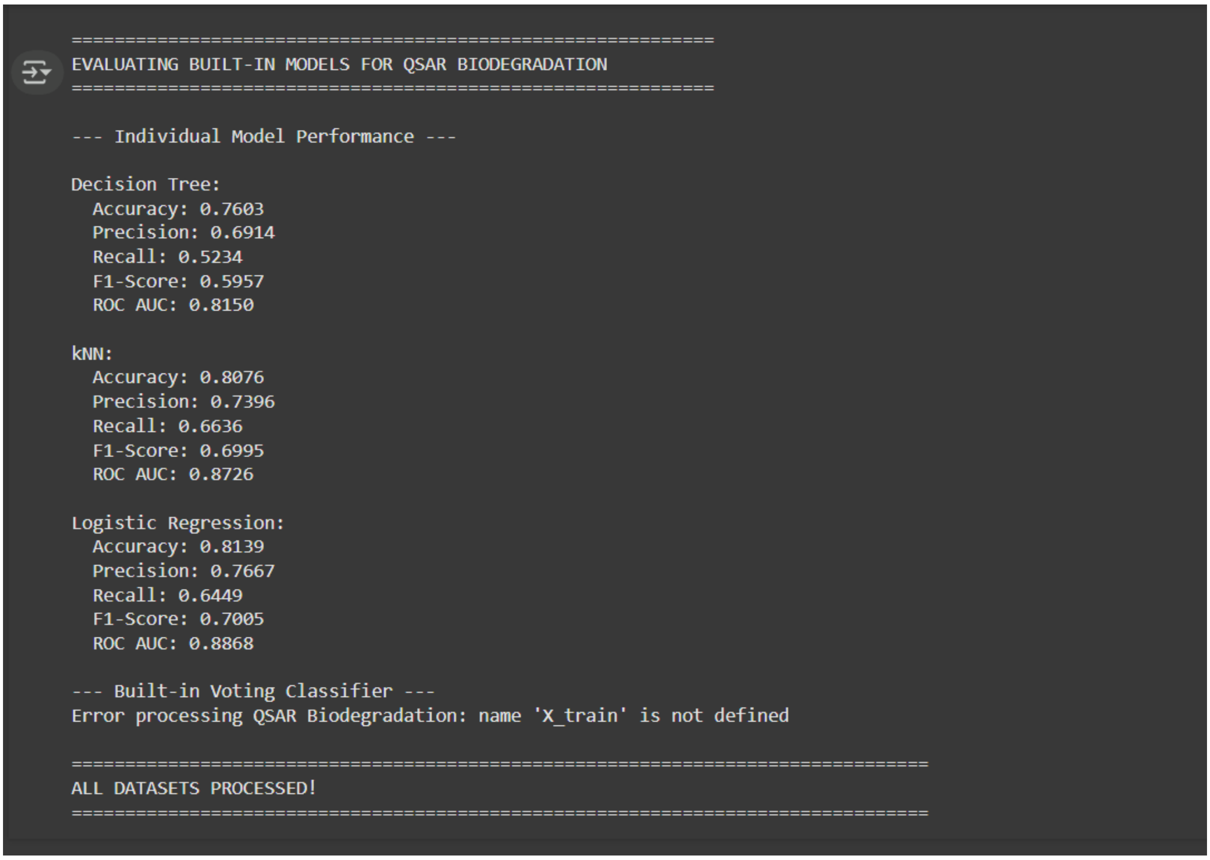


ROC Curves - QSAR Biodegradation (Manual)
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
Decision Tree (AUC = 0.815)
kNN (AUC = 0.873)
Logistic Regression (AUC = 0.887)
Voting (AUC = 0.888)
Chance
Voting Classifier - QSAR Biodegradation (Manual)
True label
0
1
188
22
40
67
0
1
Predicted label
180
160
140
120
100
80
60
40

```
============================================================
RUNNING BUILT-IN GRID SEARCH FOR QSAR BIODEGRADATION
============================================================

--- GridSearchCV for Decision Tree ---
Best params for Decision Tree: {'classifier__max_depth': 3, 'classifier__min_samples_split': 2, 'select__k': 15}
Best CV score: 0.8303
```



```
--- GridSearchCV for kNN ---
Best params for kNN: {'classifier__n_neighbors': 9, 'classifier__weights': 'distance', 'select__k': 15}
Best CV score: 0.8856

--- GridSearchCV for Logistic Regression ---
Best params for Logistic Regression: {'classifier__C': 10, 'classifier__penalty': 'l2', 'classifier__solver': 'lbfgs', 'select__k': 15}
Best CV score: 0.8816

============================================================
EVALUATING BUILT-IN MODELS FOR QSAR BIODEGRADATION
============================================================

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.7603
  Precision: 0.6914
  Recall: 0.5234
  F1-Score: 0.5957
  ROC AUC: 0.8150

kNN:
  Accuracy: 0.8076
  Precision: 0.7396
  Recall: 0.6636
  F1-Score: 0.6995
  ROC AUC: 0.8726

Logistic Regression:
  Accuracy: 0.8139
  Precision: 0.7667
  Recall: 0.6449
  F1-Score: 0.7005
  ROC AUC: 0.8868

--- Built-in Voting Classifier ---
Error processing QSAR Biodegradation: name 'X_train' is not defined
```



```
============================================================
EVALUATING BUILT-IN MODELS FOR QSAR BIODEGRADATION
============================================================

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.7603
  Precision: 0.6914
  Recall: 0.5234
  F1-Score: 0.5957
  ROC AUC: 0.8150

kNN:
  Accuracy: 0.8076
  Precision: 0.7396
  Recall: 0.6636
  F1-Score: 0.6995
  ROC AUC: 0.8726

Logistic Regression:
  Accuracy: 0.8139
  Precision: 0.7667
  Recall: 0.6449
  F1-Score: 0.7005
  ROC AUC: 0.8868

--- Built-in Voting Classifier ---
Error processing QSAR Biodegradation: name 'X_train' is not defined
```

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