Ensemble approaches

Introduction

This report presents the classification performance of several ensemble learning approaches on two distinct datasets: the digits dataset from scikit-learn and the Mammographic Mass Data Set from the UCI repository. Ensemble methods, which combine the predictions of multiple base learners, are explored to assess their effectiveness in improving predictive accuracy and robustness compared to individual classifiers. The study includes Random Forest, Bagging, AdaBoost, Majority Vote Classifier, and Gradient Boosting. Furthermore, the impact of key hyperparameters on the performance of AdaBoost and Gradient Boosting is investigated.

Methodology

Datasets:

- The Digits dataset contains 1,797 samples of 8x8 images (64 features) representing digits from 0 to 9.
- Mammographic Mass Data Set , a dataset from the UCI repository containing 961 instances with 5 features (BI-RADS assessment, Age, Shape, Margin, Density) and a binary target variable (Severity: benign or malignant).
- Both datasets are split into 80% training and 20% testing sets, with stratification to maintain class distribution.
- The Mammographic Mass dataset contained missing values, which were preprocessed using median imputation.

Classifiers:

The following classifiers were implemented and evaluated:

Base Classifiers

- Decision Tree (Entropy criterion)
- Logistic Regression (L2 regularization, C=0.001)
- K-Nearest Neighbors (k=1, Minkowski metric)

Ensemble Methods

- Random Forest (n_estimators=100, criterion=gini)
- Bagging (base: Decision Tree, n_estimators=100)
- AdaBoost (base: Decision Tree, n_estimators=100, learning_rate=0.1)

- Gradient Boosting (n_estimators=100, max_depth=3)
- Voting Classifier (hard voting: DT, LR, KNN)

Evaluation Metrics

- Accuracy Score
- Training Time (seconds)
- Classification Report (Precision, Recall, F1-score)
- ROC Curve and AUC (binary classification only)

Results and Analysis

Task 1: AdaBoost Calculation

To calculate the updated weights for the next boosting round (i+1) I used python, the following steps were performed: (please see index.py)

1. The weighted error rate (ε) was computed by summing the weights of the misclassified instances from the previous round. $\boldsymbol{\epsilon} = \boldsymbol{w}(\hat{\boldsymbol{y}} \neq \boldsymbol{y})$
2. The coefficient (α_j), representing the weight of the classifier's vote, was calculated using the formula $\alpha_j = 0.5\, ln\frac{1-\epsilon}{\epsilon}$.
3. The weights of each instance were updated $\boldsymbol{w} = \boldsymbol{w} * exp\left(-\alpha_j * \hat{\boldsymbol{y}} * \boldsymbol{y}\right)$: for correctly classified instances, the weights were decreased by multiplying by exp(-α_j), while for incorrectly classified instances, the weights were increased by multiplying by exp(α_j).
4. Finally, the updated weights were normalized to ensure they sum to 1, providing the weight distribution for the next boosting iteration. $\boldsymbol{w} = \frac{\boldsymbol{w}}{\Sigma_i w_i}$

Final Results: Updated Weights

index	x	y	weights	$(\hat{\boldsymbol{y}})$	updated weights
1	1.0	1	0.072	1	0.054
2	2.0	1	0.072	1	0.054
3	3.0	1	0.072	1	0.054
4	4.0	-1	0.072	-1	0.054
5	5.0	-1	0.072	-1	0.054
6	6.0	-1	0.072	-1	0.054
7	7.0	1	0.167	1	0.125
8	8.0	1	0.167	-1	0.249

9	9.0	1	0.167	-1	0.249
10	10.0	-1	0.072	-1	0.054

Task 2: ensemble learning approaches

ROC curve analysis was performed but due to the multi-class nature of the Digits dataset, ROC curves were not generated for its evaluation. For the Mammographic dataset, the plots visualize the trade-off between the true positive rate and the false positive rate for each classifier. The Area Under the Curve (AUC) values are also displayed in the legend of each ROC plot.

Digits Dataset:

Classifier	Accuracy	Training Time (seconds)
Decision Tree	0.8750	0.0334
Logistic Regression	0.9194	0.0733
KNN	0.9806	0.0015
Random Forest	0.9694	0.3054
Bagging	0.9667	4.9057
AdaBoost	0.8778	0.0321
Majority Vote Classifier	0.9667	0.0934
Gradient Boosting	0.9667	6.5463
AdaBoost (n_estimators=50)	0.8778	0.0286
AdaBoost (n_estimators=100)	0.8778	0.0323
AdaBoost (n_estimators=150)	0.8778	0.0345
Gradient Boosting (max_depth=1)	0.9500	2.8308
Gradient Boosting (max_depth=3)	0.9667	7.7517
Gradient Boosting (max_depth=5)	0.9583	13.6859

For the Digits dataset, KNN achieved the highest accuracy among all classifiers and with a very low training time, suggesting that the digit features are well-separated in the feature space. The Ensemble methods Random Forest, Bagging, Majority Vote Classifier, and Gradient Boosting also performed very well, achieving accuracies above 96%. Bagging and Gradient Boosting had significantly higher training times compared to other methods. Hyperparameter tuning for AdaBoost (n_estimators) did not show any meaningful change in accuracy for the tested values. For Gradient Boosting, hyperparameter tuning of ‘max_depth’ showed that a ‘max_depth’ of 3 yielded the best accuracy among the tested values, with ‘max_depth =5’

increasing training time without a substantial improvement in performance. A lower 'max_depth' of 1 resulted in a slightly lower accuracy but a much faster training time.

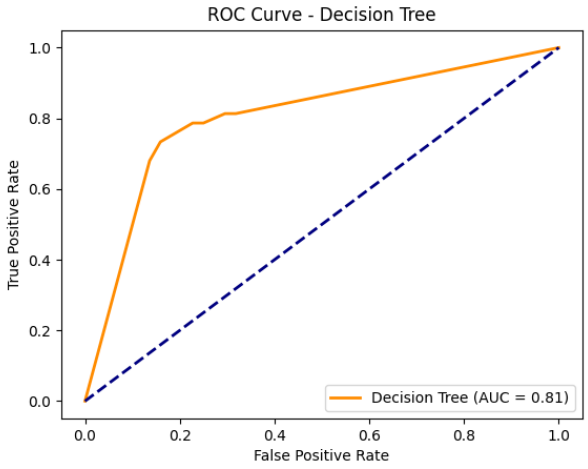
Mammographic Dataset:

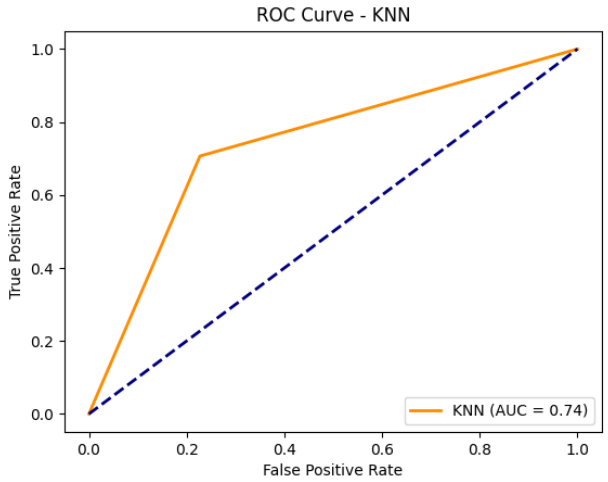
Classifier	Accuracy	Training Time (seconds)
Decision Tree	0.7513	0.0042
Logistic Regression	0.8083	0.0224
KNN	0.7358	0.0104
Random Forest	0.7565	0.3655
Bagging	0.7565	0.2297
AdaBoost	0.7461	0.5772
Majority Vote Classifier	0.7979	0.0501
Gradient Boosting	0.7979	0.2158
AdaBoost (n_estimators=50)	0.7565	0.3475
AdaBoost (n_estimators=100)	0.7461	0.5274
AdaBoost (n_estimators=150)	0.7358	0.7191
Gradient Boosting (max_depth=1)	0.7979	0.2452
Gradient Boosting (max_depth=3)	0.7979	0.2942
Gradient Boosting (max_depth=5)	0.7668	0.4017

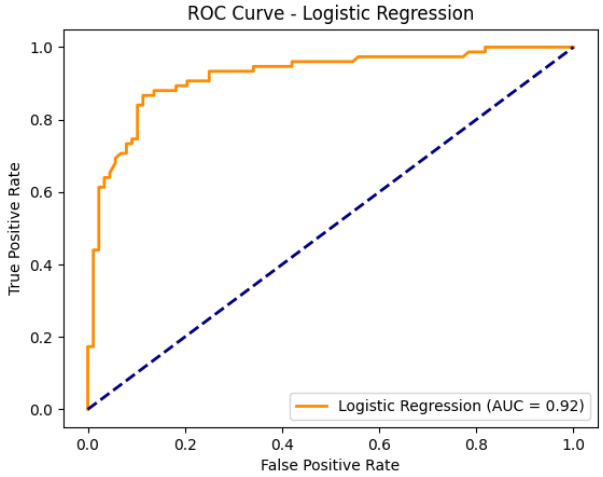
For the Mammographic dataset, Logistic Regression achieved the highest accuracy score and the only accuracy over 80%. The Majority Vote Classifier, despite using hard voting, also showed competitive accuracy. AdaBoost's overall performance was lower compared to other ensembles on this dataset. For Gradient Boosting, increasing the parameter 'max_depth' to 5 resulted in a decrease in accuracy and an increase in training time. The training times for ensemble methods on the Mammographic dataset were lower than on the Digits dataset, likely due to the smaller size of the Mammographic dataset.

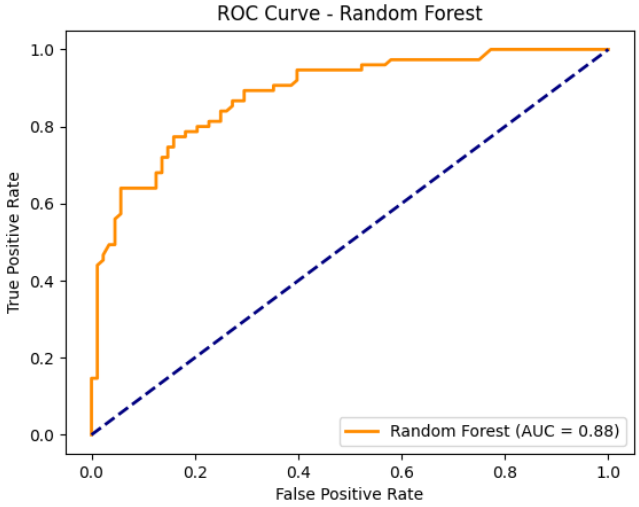
Plots

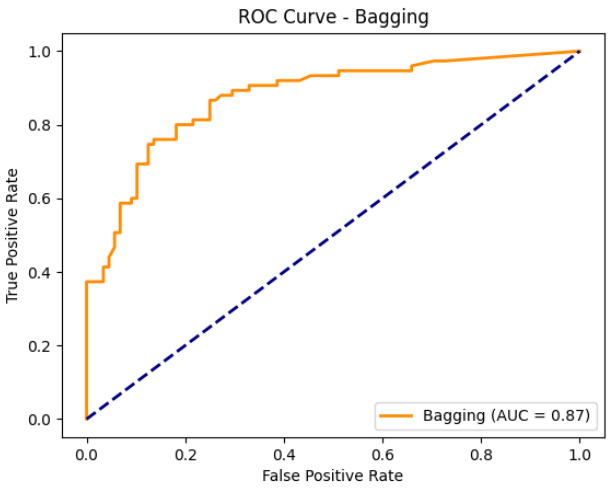
Below are a few of the plots that will be generated by the program.

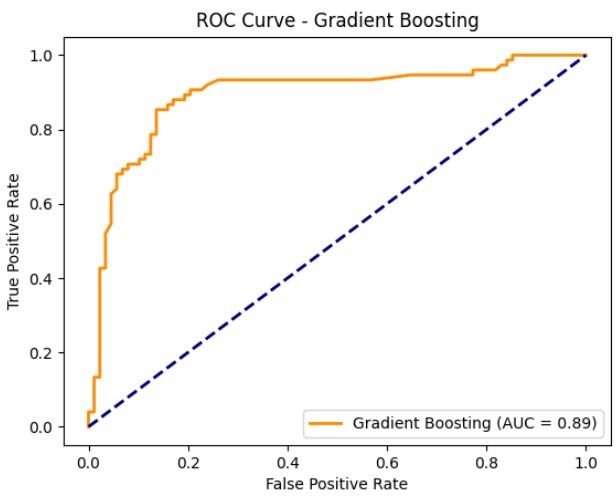

ROC Curve - Decision Tree
True Positive Rate
False Positive Rate
Decision Tree (AUC = 0.81)
0.0
0.2
0.4
0.6
0.8
1.0

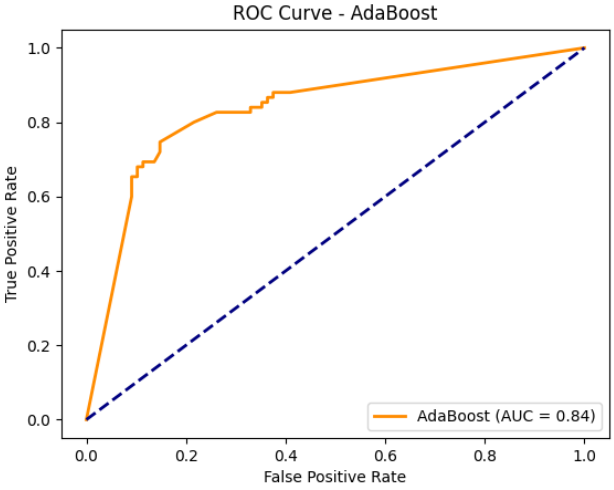

ROC Curve - KNN
True Positive Rate
False Positive Rate
KNN (AUC = 0.74)
0.0
0.2
0.4
0.6
0.8
1.0

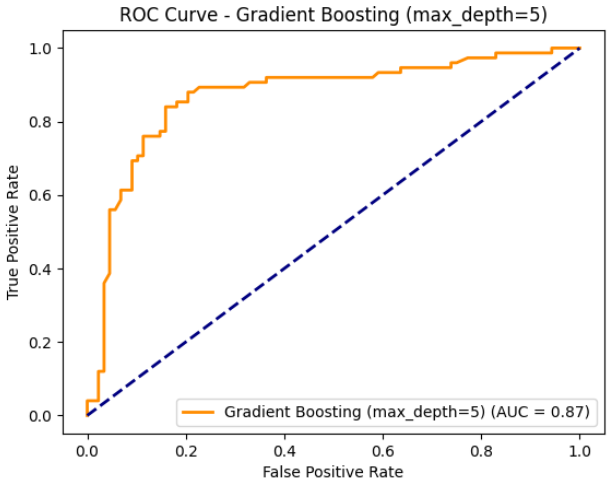

ROC Curve - Logistic Regression
True Positive Rate
False Positive Rate
Logistic Regression (AUC = 0.92)
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve - Random Forest
True Positive Rate
False Positive Rate
Random Forest (AUC = 0.88)
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve - Bagging
True Positive Rate
False Positive Rate
Bagging (AUC = 0.87)
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve - Gradient Boosting
True Positive Rate
False Positive Rate
Gradient Boosting (AUC = 0.89)
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve - AdaBoost
True Positive Rate
False Positive Rate
AdaBoost (AUC = 0.84)
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve - Gradient Boosting (max_depth=5)
True Positive Rate
False Positive Rate
Gradient Boosting (max_depth=5) (AUC = 0.87)
0.0
0.2
0.4
0.6
0.8
1.0