Kyle Tranfaglia

COSC 411

Project 03

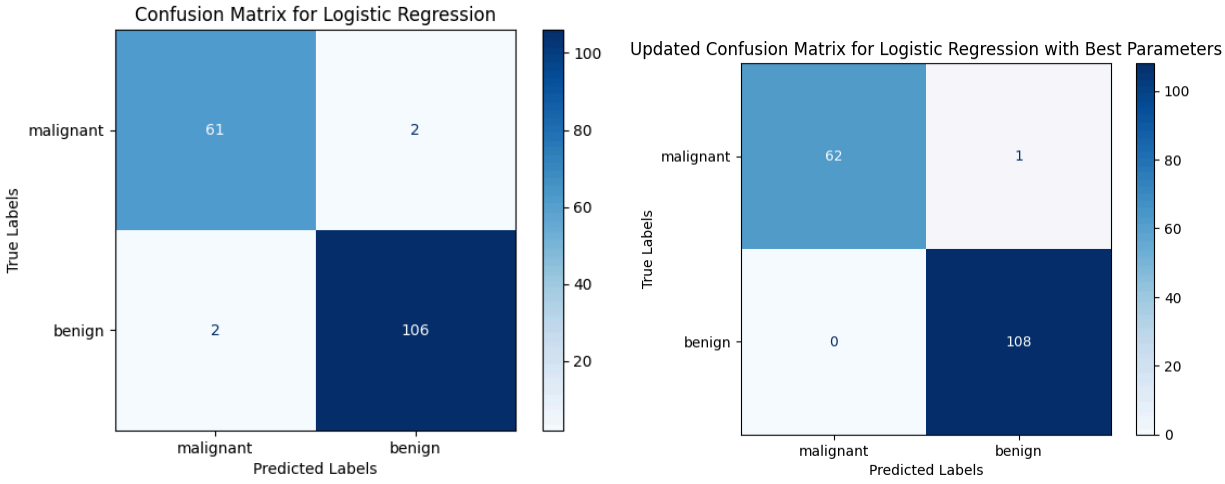
06 Dec. 2024

Baseline and Ensemble Model Building Report

Part 1: Classification on the Breast Cancer Dataset

Task 1: Baseline Models

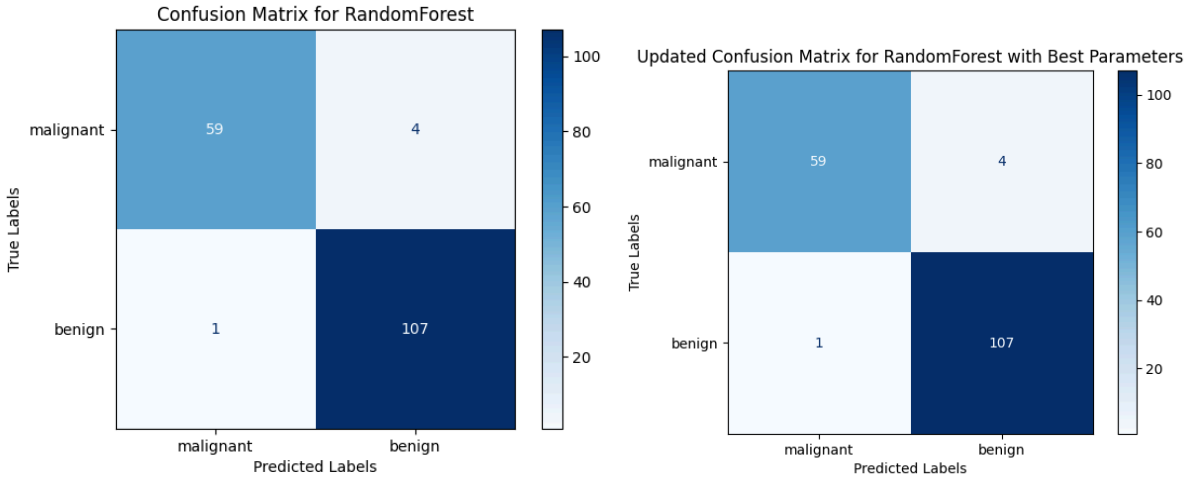
- The best Classifier is Logistic Regression, with an accuracy of 97.66%
- After a Grid search with cross-validation, the following was found:
    - Best Parameters: {'C': 0.1, 'penalty': 'l2', 'solver': 'liblinear'}
    - Best Cross-Validation Accuracy: 97.74%
    - Updated Accuracy Score for Logistic Regression: 99.42%



Task 2: Ensemble Models

- The Best Classifier is RandomForest, with an accuracy of 97.08%
- After a Grid search with cross-validation, the following was found:
    - Best Parameters: {'max_depth': 10, 'min_samples_split': 2, 'n_estimators': 400}

- Best Cross-Validation Accuracy: 95.72%
- Updated Accuracy Score for RandomForest: 97.08%



Discussion

The model building concluded that a baseline model best fits this dataset, with the hyper-tuned Logistic Regression model achieving an accuracy score of 99.42%. The best ensemble model, Random Forest, achieved a solid accuracy of 97.08%, though it was lower than Logistic Regression. Referring to the confusion matrices, the Logistic Regression model misclassified only one case, which was particularly concerning as it labeled a malignant case as benign. However, this is still better than the hyper-tuned Random Forest model, which misclassified four malignant and one benign case. The misclassification of malignant cases as benign is especially problematic, as it could lead to missed diagnoses and delays in treatment. Nevertheless, Logistic Regression's single misclassification suggests it is more reliable in identifying malignant cases than the Random Forest model. This misclassification is likely due to an atypical malignant case close to benign.

Another model was built using AdaBoost to attempt to tune hyperparameters enough to outperform Random Forest, but it only matched its results. All models show lower

cross-validation accuracy compared to their test set performance. This discrepancy arises because cross-validation trains the model on smaller subsets of data, leading to slightly more conservative performance estimates. In contrast, the model in the test set evaluation is trained on the entire dataset. Overall, the models perform very well.

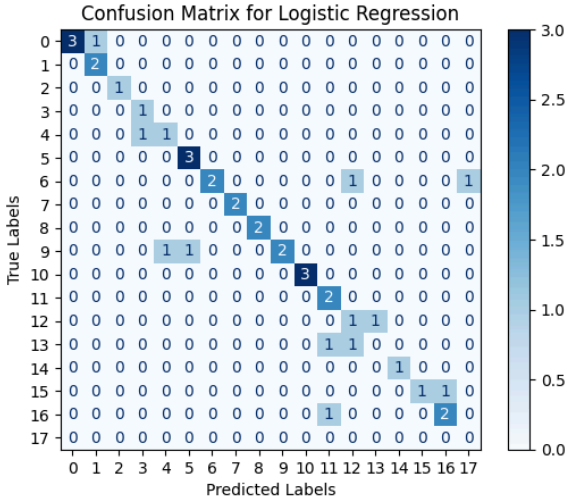
Part 2: Classification on the Project3 Dataset

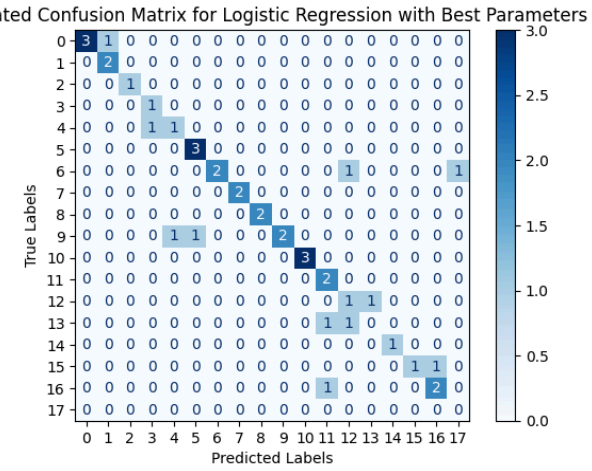
Task 1: Feature Normalization

- Features (columns) 1-65 were normalized using the Scikit-learn StandradScaler function
- I found that scaling this data was very significant for model performance, as the average accuracy amongst the tested classifiers was remarkably better when using the scaled data

Task 2: Baseline Models

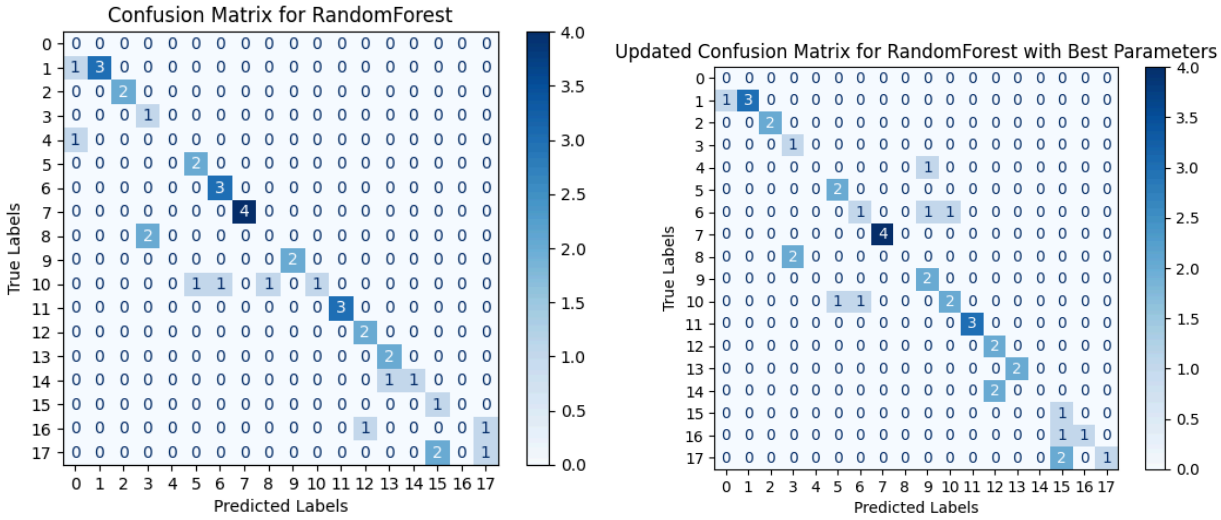
- The best Classifier is Logistic Regression, with an accuracy of 72.50%
- After a Grid search with cross-validation, the following was found:
    - Best Parameters: {'C': 1, 'penalty': 'l2', 'solver': 'saga'}
    - Best Cross-Validation Accuracy: 64.72%
    - Updated Accuracy Score for Logistic Regression: 72.50%





Task 3: Ensemble Models

- The Best Classifier is RandomForest, with an accuracy of 70.00%
- After a Grid search with cross-validation, the following was found:
    - Best Parameters: {'max_depth': 20, 'min_samples_split': 2, 'n_estimators': 500}
    - Best Cross-Validation Accuracy: 63.45%
    - Updated Accuracy Score for RandomForest: 67.50%



Discussion

The analysis of the Project3 dataset highlighted the importance of feature normalization, with scaling significantly improving model performance. Logistic Regression emerged as the best baseline classifier with an accuracy of 72.50%. However, after hyperparameter tuning, the model's cross-validation accuracy decreased to 64.72%, with the updated test set accuracy of 72.50%. This discrepancy suggests that while the model performs well on the test set, it may have overfitted during training, as evidenced by the lower performance on the cross-validation folds. This indicates that the model is likely more sensitive to the small dataset size (196 entries) and struggles with generalization. After hyperparameter tuning, Random Forest achieved a test accuracy of 70.00%, with a slightly lower cross-validation accuracy of 63.45%. This suggests

that Random Forest may generalize better than Logistic Regression, though both models still face challenges with the small dataset. Despite its higher performance, the Random Forest model's modest improvement after hyperparameter tuning indicates that further adjustments may be necessary for substantial gains.

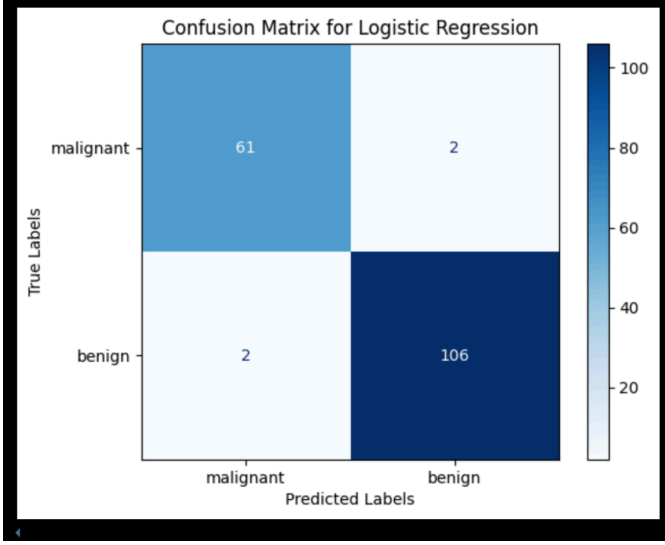
The modest accuracy scores reflect the challenges of working with a small dataset of 196 entries and 65 features. High dimensionality and limited data points make it difficult for both models to distinguish meaningful patterns and avoid overfitting. Acquiring more data, reducing feature dimensionality, and considering alternative evaluation metrics (like precision or recall) would help improve model performance. Overall, while both models performed reasonably well, further optimization is required to handle the dataset's limitations more appropriately.

Outputs

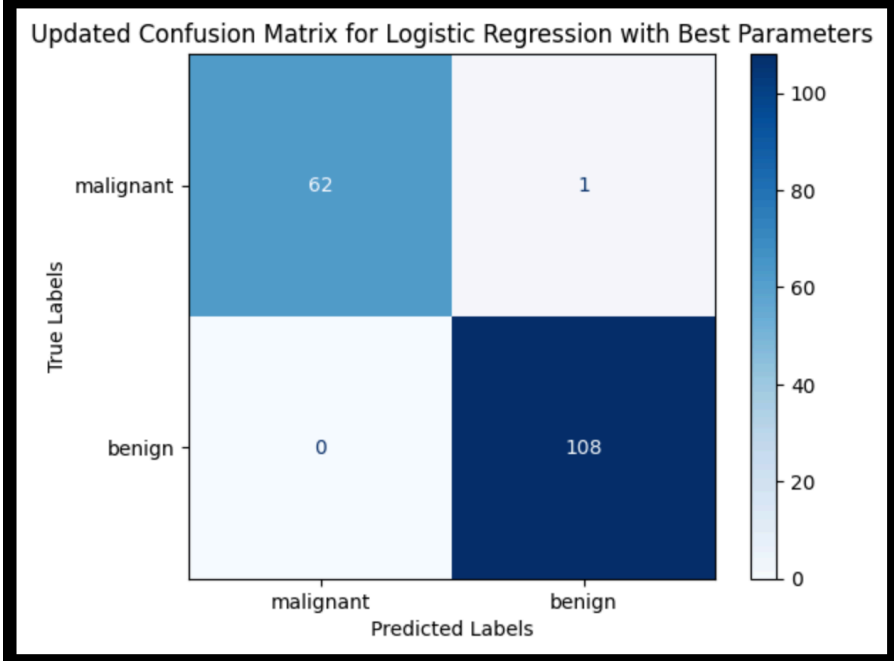
Part 1: Classification on the Breast Cancer Dataset

Task 1: Baseline Models

```
Baseline Classifier Accuracies:
Logistic Regression: Test Accuracy = 97.66%, Cross-Validation Accuracy = 94.72%
KNN: Test Accuracy = 95.91%, Cross-Validation Accuracy = 91.45%
ANN: Test Accuracy = 95.91%, Cross-Validation Accuracy = 92.96%
Decision Tree: Test Accuracy = 94.15%, Cross-Validation Accuracy = 90.96%
Naive Bayes: Test Accuracy = 94.15%, Cross-Validation Accuracy = 93.72%
SVM: Test Accuracy = 93.57%, Cross-Validation Accuracy = 89.43%

Best Classifier is Logistic Regression with an accuracy of 97.66%

Classification Report for Logistic Regression:

              precision    recall  f1-score   support

   malignant       0.97      0.97      0.97        63
      benign       0.98      0.98      0.98       108

    accuracy                           0.98       171
   macro avg       0.97      0.97      0.97       171
weighted avg       0.98      0.98      0.98       171
```

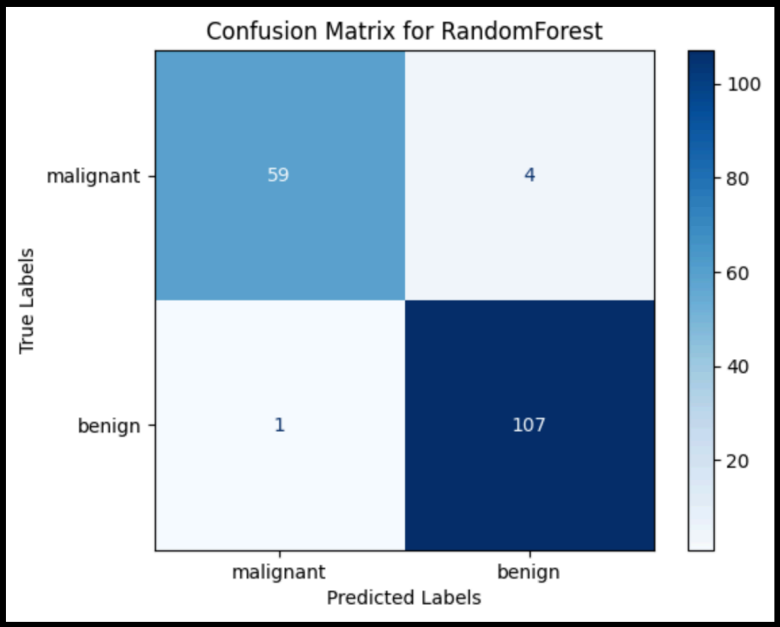


```
Best Parameters: {'C': 0.1, 'penalty': 'l2', 'solver': 'liblinear'}
Best Cross-Validation Accuracy: 97.74%

Updated Classification Report for Hypertuned Logistic Regression:

              precision    recall  f1-score   support

   malignant       1.00      0.98      0.99        63
      benign       0.99      1.00      1.00       108

    accuracy                           0.99       171
   macro avg       1.00      0.99      0.99       171
weighted avg       0.99      0.99      0.99       171

Updated accuracy for Logistic Regression: 99.42%
```

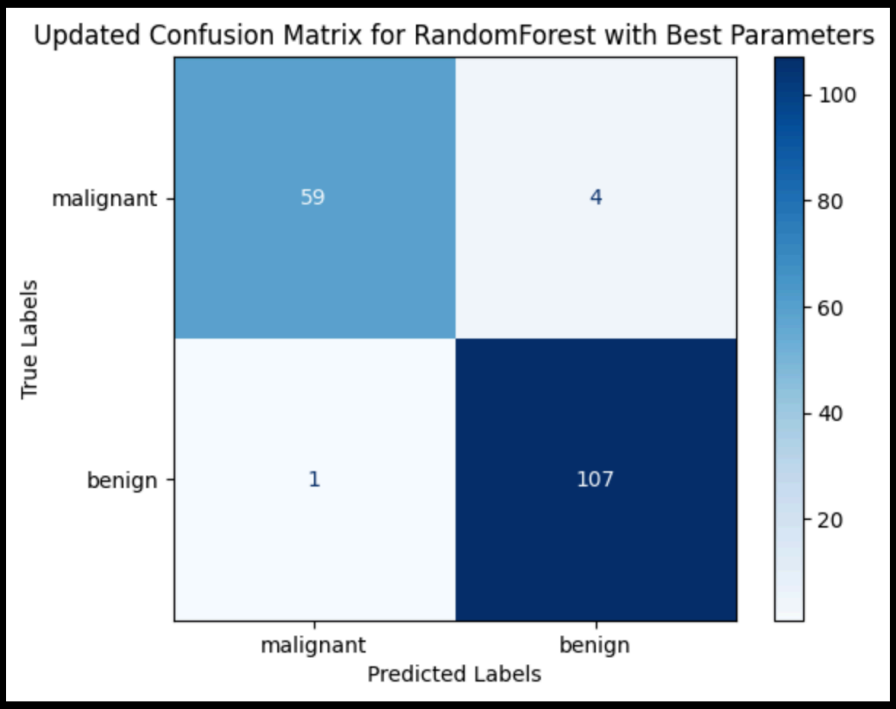


Task 2: Ensemble Models

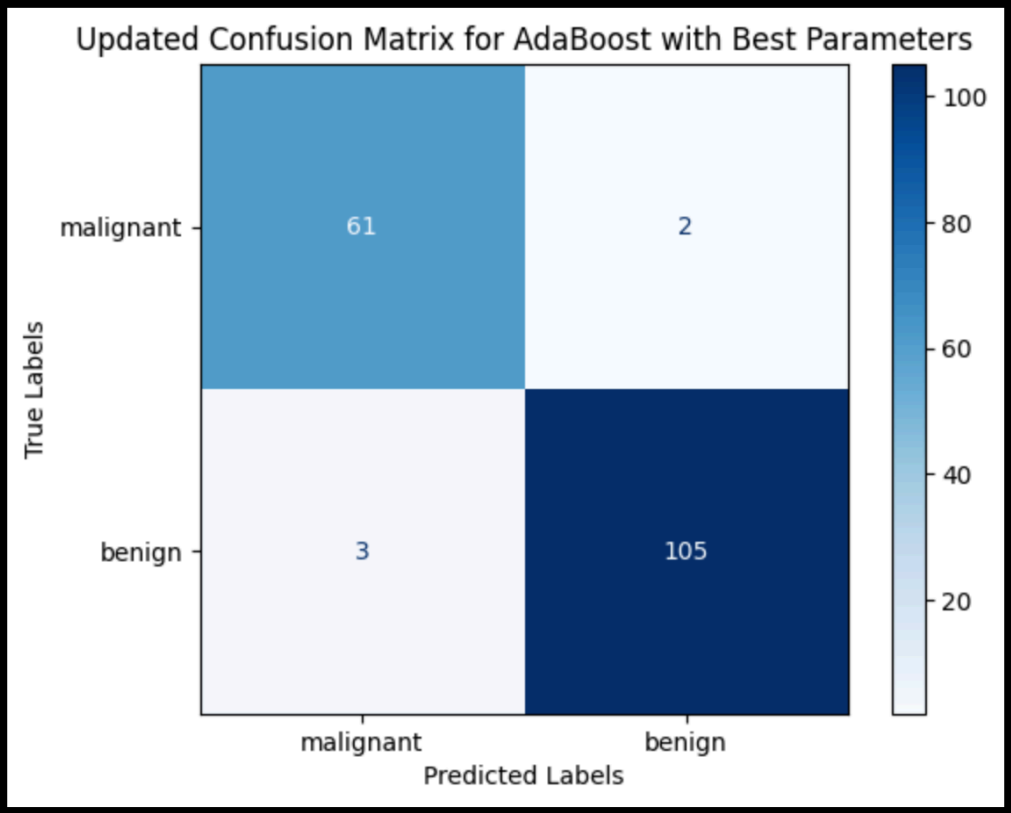
```
Esemble Classifier Accuracies:
RandomForest: Test Accuracy = 97.08%, Cross-Validation Accuracy = 94.97%
AdaBoost: Test Accuracy = 97.08%, Cross-Validation Accuracy = 97.49%
GradientBoosting: Test Accuracy = 95.91%, Cross-Validation Accuracy = 95.48%
Bagging: Test Accuracy = 94.74%, Cross-Validation Accuracy = 93.46%

Best Classifier is RandomForest with an accuracy of 97.08%

Classification Report for RandomForest:

              precision    recall  f1-score   support

   malignant       0.98      0.94      0.96        63
      benign       0.96      0.99      0.98       108

    accuracy                           0.97       171
   macro avg       0.97      0.96      0.97       171
weighted avg       0.97      0.97      0.97       171
```



```
Best Parameters: {'max_depth': 10, 'min_samples_split': 2, 'n_estimators': 400}
Best Cross-Validation Accuracy: 95.72%

Updated Classification Report for Hypertuned RandomForest:

              precision    recall  f1-score   support

   malignant       0.98      0.94      0.96        63
      benign       0.96      0.99      0.98       108

    accuracy                           0.97       171
   macro avg       0.97      0.96      0.97       171
weighted avg       0.97      0.97      0.97       171

Updated accuracy for RandomForest: 97.08%
```



```
Updated Classification Report for Hypertuned AdaBoost:

              precision    recall  f1-score   support

   malignant       0.95      0.97      0.96        63
      benign       0.98      0.97      0.98       108

    accuracy                           0.97       171
   macro avg       0.97      0.97      0.97       171
weighted avg       0.97      0.97      0.97       171

Updated accuracy for AdaBoost: 97.08%
```



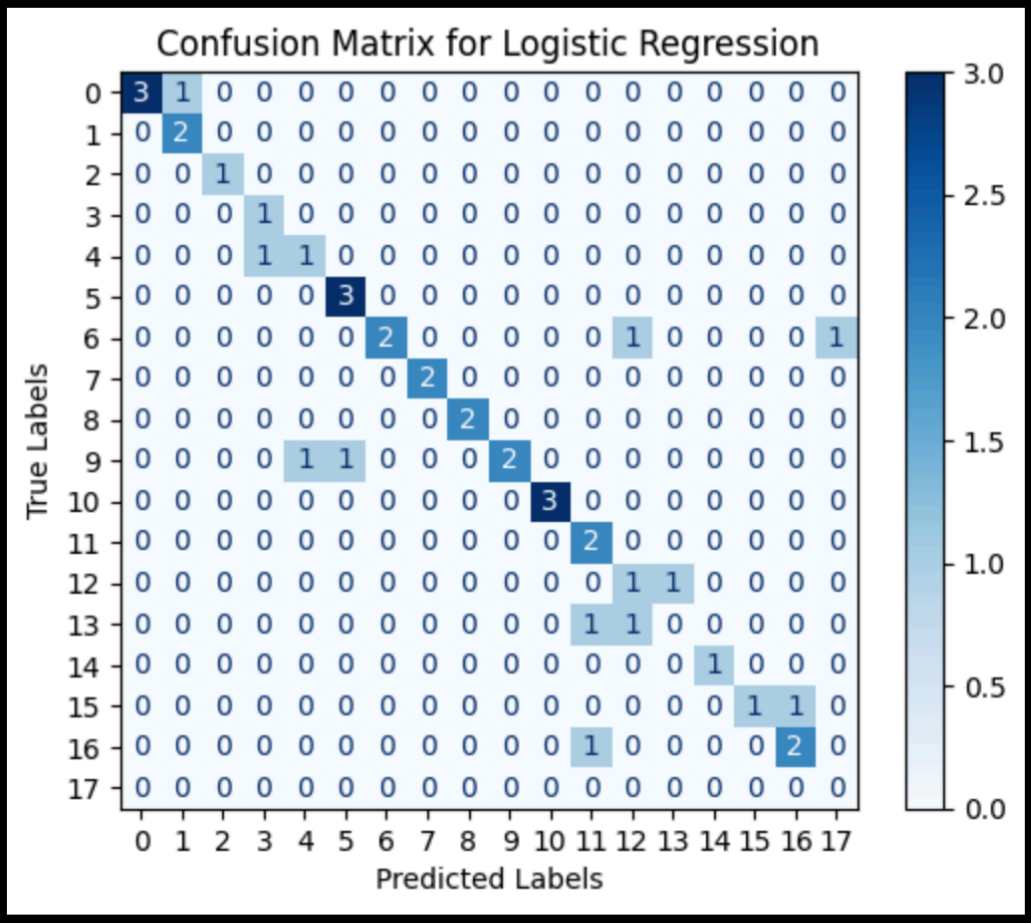
Part 2: Classification on the Project3 Dataset

Task 1: Feature Normalization

No Results (preprocessing)

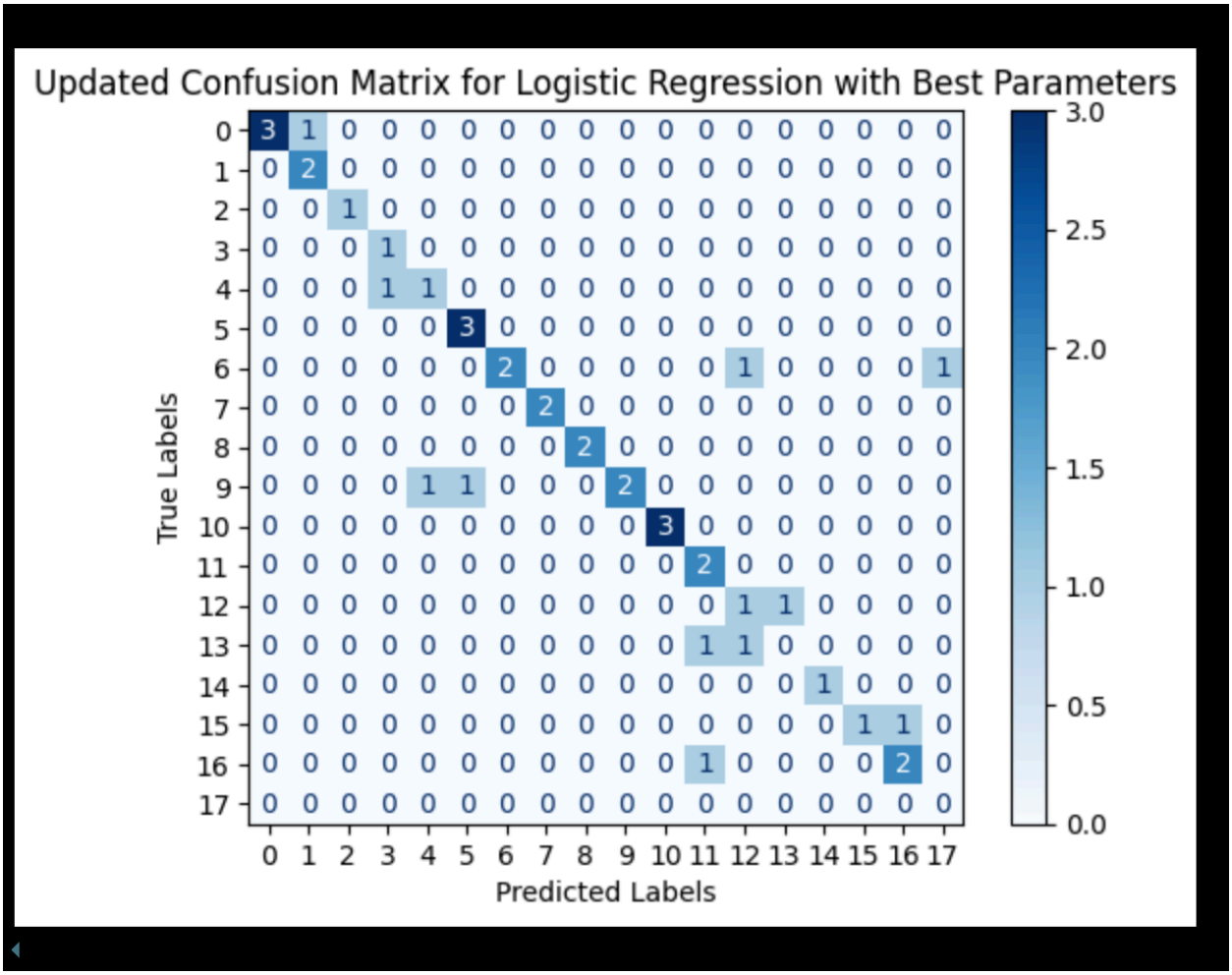
Task 2: Baseline Models

```
Baseline Classifier Accuracies:
Logistic Regression: Test Accuracy = 72.50%, Cross-Validation Accuracy = 64.05%
Naive Bayes: Test Accuracy = 65.00%, Cross-Validation Accuracy = 55.08%
ANN: Test Accuracy = 65.00%, Cross-Validation Accuracy = 57.74%
SVM: Test Accuracy = 55.00%, Cross-Validation Accuracy = 51.27%
KNN: Test Accuracy = 50.00%, Cross-Validation Accuracy = 49.31%
Decision Tree: Test Accuracy = 42.50%, Cross-Validation Accuracy = 46.79%

Best Classifier is Logistic Regression with an accuracy of 72.50%

Classification Report for Logistic Regression:

              precision    recall  f1-score   support

           2       1.00      0.75      0.86         4
           3       0.67      1.00      0.80         2
           4       1.00      1.00      1.00         1
           5       0.50      1.00      0.67         1
           6       0.50      0.50      0.50         2
           7       0.75      1.00      0.86         3
           8       1.00      0.50      0.67         4
           9       1.00      1.00      1.00         2
          10       1.00      1.00      1.00         2
          12       1.00      0.50      0.67         4
          13       1.00      1.00      1.00         3
          14       0.50      1.00      0.67         2
          15       0.33      0.50      0.40         2
          16       0.00      0.00      0.00         2
          17       1.00      1.00      1.00         1
          18       1.00      0.50      0.67         2
          19       0.67      0.67      0.67         3
          20       0.00      0.00      0.00         0

    accuracy                           0.72        40
   macro avg       0.72      0.72      0.69        40
weighted avg       0.79      0.72      0.73        40
```



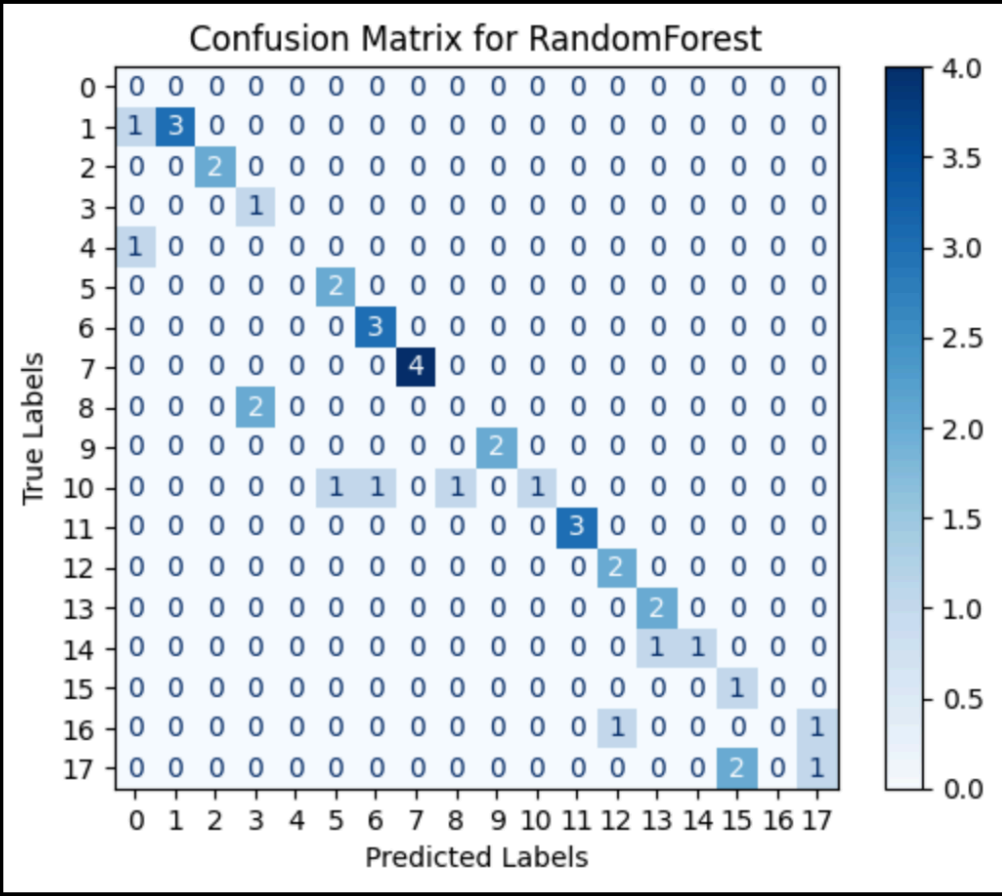

```
Best Parameters: {'C': 1, 'penalty': 'l2', 'solver': 'saga'}
Best Cross-Validation Accuracy: 64.72%
```

```
Updated Classification Report for Hypertuned Logistic Regression:

              precision    recall  f1-score   support

           2       1.00      0.75      0.86         4
           3       0.67      1.00      0.80         2
           4       1.00      1.00      1.00         1
           5       0.50      1.00      0.67         1
           6       0.50      0.50      0.50         2
           7       0.75      1.00      0.86         3
           8       1.00      0.50      0.67         4
           9       1.00      1.00      1.00         2
          10       1.00      1.00      1.00         2
          12       1.00      0.50      0.67         4
          13       1.00      1.00      1.00         3
          14       0.50      1.00      0.67         2
          15       0.33      0.50      0.40         2
          16       0.00      0.00      0.00         2
          17       1.00      1.00      1.00         1
          18       1.00      0.50      0.67         2
          19       0.67      0.67      0.67         3
          20       0.00      0.00      0.00         0

    accuracy                           0.72        40
   macro avg       0.72      0.72      0.69        40
weighted avg       0.79      0.72      0.73        40

Updated accuracy for Logistic Regression: 72.50%
```


Updated Confusion Matrix for Logistic Regression with Best Parameters
True Labels
Predicted Labels
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
3.0
2.5
2.0
1.5
1.0
0.5
0.0

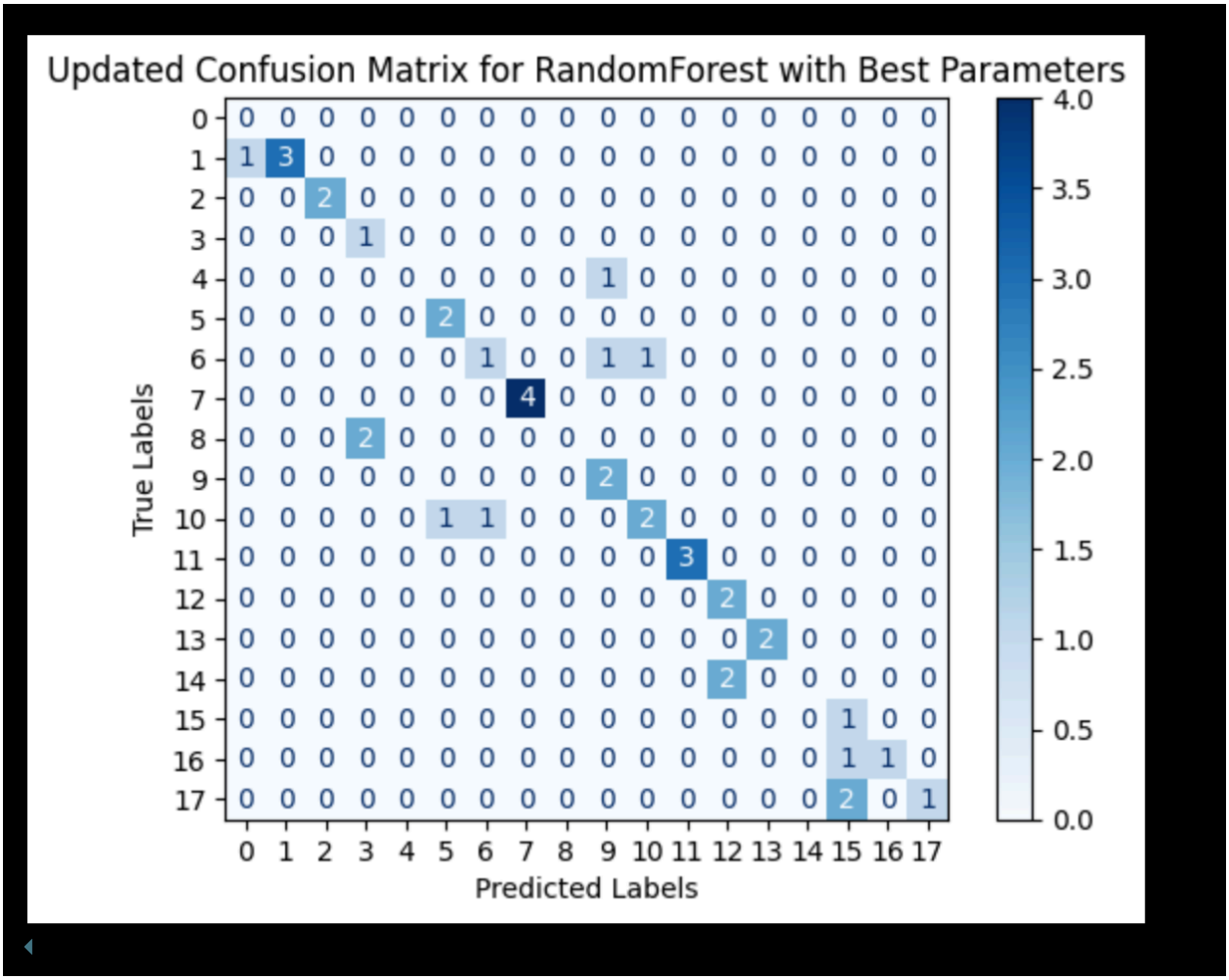
Task 3: Ensemble Models

```
Esemble Classifier Accuracies:
RandomForest: Test Accuracy = 70.00%, Cross-Validation Accuracy = 62.84%
Bagging: Test Accuracy = 57.50%, Cross-Validation Accuracy = 53.87%
GradientBoosting: Test Accuracy = 57.50%, Cross-Validation Accuracy = 37.84%
AdaBoost: Test Accuracy = 17.50%, Cross-Validation Accuracy = 17.94%

Best Classifier is RandomForest with an accuracy of 70.00%

Classification Report for RandomForest:

              precision    recall  f1-score   support

           1       0.00      0.00      0.00         0
           2       1.00      0.75      0.86         4
           3       1.00      1.00      1.00         2
           4       0.33      1.00      0.50         1
           5       0.00      0.00      0.00         1
           6       0.67      1.00      0.80         2
           7       0.75      1.00      0.86         3
           8       1.00      1.00      1.00         4
           9       0.00      0.00      0.00         2
          10       1.00      1.00      1.00         2
          12       1.00      0.25      0.40         4
          13       1.00      1.00      1.00         3
          14       0.67      1.00      0.80         2
          15       0.67      1.00      0.80         2
          16       1.00      0.50      0.67         2
          17       0.33      1.00      0.50         1
          18       0.00      0.00      0.00         2
          19       0.50      0.33      0.40         3

    accuracy                           0.70        40
   macro avg       0.61      0.66      0.59        40
weighted avg       0.74      0.70      0.67        40
```




```
Best Parameters: {'max_depth': 20, 'min_samples_split': 2, 'n_estimators': 500}
Best Cross-Validation Accuracy: 63.45%
```

```
Updated Classification Report for Hypertuned RandomForest:

              precision    recall  f1-score   support

           1       0.00      0.00      0.00         0
           2       1.00      0.75      0.86         4
           3       1.00      1.00      1.00         2
           4       0.33      1.00      0.50         1
           5       0.00      0.00      0.00         1
           6       0.67      1.00      0.80         2
           7       0.50      0.33      0.40         3
           8       1.00      1.00      1.00         4
           9       0.00      0.00      0.00         2
          10       0.50      1.00      0.67         2
          12       0.67      0.50      0.57         4
          13       1.00      1.00      1.00         3
          14       0.50      1.00      0.67         2
          15       1.00      1.00      1.00         2
          16       0.00      0.00      0.00         2
          17       0.25      1.00      0.40         1
          18       1.00      0.50      0.67         2
          19       1.00      0.33      0.50         3

    accuracy                           0.68        40
   macro avg       0.58      0.63      0.56        40
weighted avg       0.70      0.68      0.65        40

Updated accuracy for RandomForest: 67.50%
```


Updated Confusion Matrix for RandomForest with Best Parameters
True Labels
Predicted Labels
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 1 0 0 1 1 0 0 0 0 0 0 0
0 0 0 0 0 0 0 4 0 0 0 0 0 0 0 0 0 0
0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0
0 0 0 0 0 1 1 0 0 0 2 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 3 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0