Assignment 4 – Machine Learning

Questions 1 to 3

Group 15

Introduction

The objective of this assignment is to develop and evaluate tree-based models for predicting whether a subject with Mild Cognitive Impairment (MCI) will convert to Alzheimer's Disease (progressive MCI or pMCI) or remain stable (stable MCI or sMCI) based on glucose metabolism features from 14 cortical brain regions.

Question 1 – Decision Tree Classifier

We trained a **Decision Tree Classifier** using a grid search to optimize the `criterion` hyperparameter among `gini`, `entropy`, and `log_loss`. The dataset included 202 sMCI and 202 pMCI individuals for training, and 361 sMCI and 194 pMCI individuals for testing.

Best Criterion

The best performing criterion found was `gini`.

Performance on Test Data

- Accuracy: 59.1%
- Precision: 44.5%
- Recall (Sensitivity): 69.1%
- Specificity: 53.7%
- Balanced Accuracy: **61.4%**

Discussion

The decision tree shows reasonable recall for the pMCI class, but suffers from relatively low specificity and precision. The model likely overfits due to its complexity and limited generalization capacity, as reflected by the moderate balanced accuracy.

Tree Visualization

Figure 1 displays the structure of the trained Decision Tree.

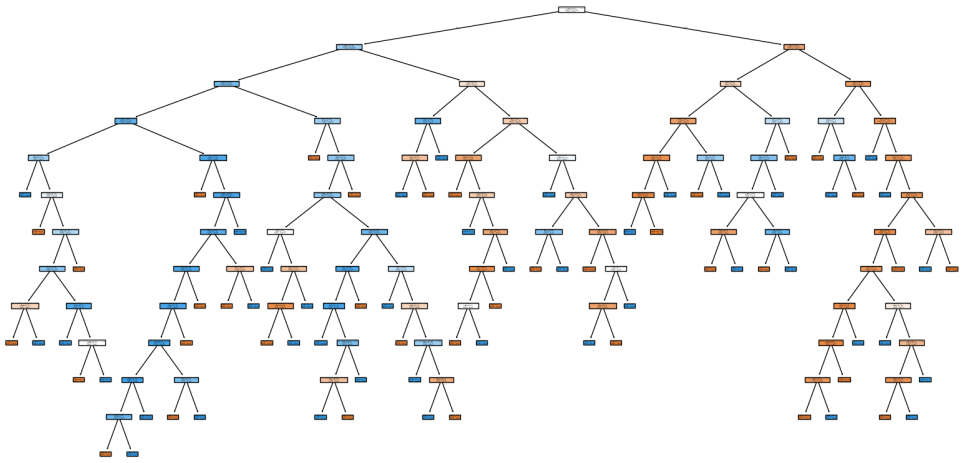

Figure 1: Trained Decision Tree Classifier

Question 2 – Analysis of the Tree Structure

The Decision Tree uses multiple features repeatedly, particularly in early splits, indicating higher importance. From visual inspection of the tree, the most frequently used features near the root include:

- `ctx-rh-precuneus`
- `ctx-lh-inferiorparietal`
- `ctx-lh-middletemporal`

This suggests these regions may carry stronger predictive signals for distinguishing sMCI and pMCI classes in this dataset.

Question 3 – Random Forest Classifier

We trained a **Random Forest Classifier** using GridSearchCV to tune the number of estimators (`n_estimators` = 100 or 200) and the splitting criterion (`gini`, `entropy`, or `log_loss`).

Best Parameters

- `criterion`: `entropy`
- `n_estimators`: 100

Performance on Test Data

- Accuracy: 64.5%
- Precision: 49.5%
- Recall (Sensitivity): 78.4%
- Specificity: 57.1%
- Balanced Accuracy: **67.7%**

Discussion

The Random Forest model significantly outperforms the single Decision Tree in terms of recall and balanced accuracy. This reflects the advantage of ensembling in reducing overfitting and increasing generalization.

Additionally, feature importance analysis highlights several regions consistently contributing to the classification, though the predictive signal remains weak overall, likely due to the subtle nature of pMCI conversion patterns.

1 Question 4

After extensive hyperparameter tuning using grid search with cross-validation on a Random Forest classifier, the best model was selected based on balanced accuracy. The final chosen model uses the `entropy` criterion, `class_weight="balanced"`, `n_estimators=200`, `max_depth=20`, and `min_samples_split=5`. This configuration achieved the highest validation balanced accuracy of 66.98%.

However, this performance remains considerably lower than what was observed in previous assignments using SVM-based approaches. Despite optimization efforts, tree-based classifiers do not appear to generalize well on this dataset. This suggests that the underlying decision boundaries may not be well captured by tree-based models, and alternative approaches such as kernel methods might be better suited for this classification task.