Homework 2 Analysis

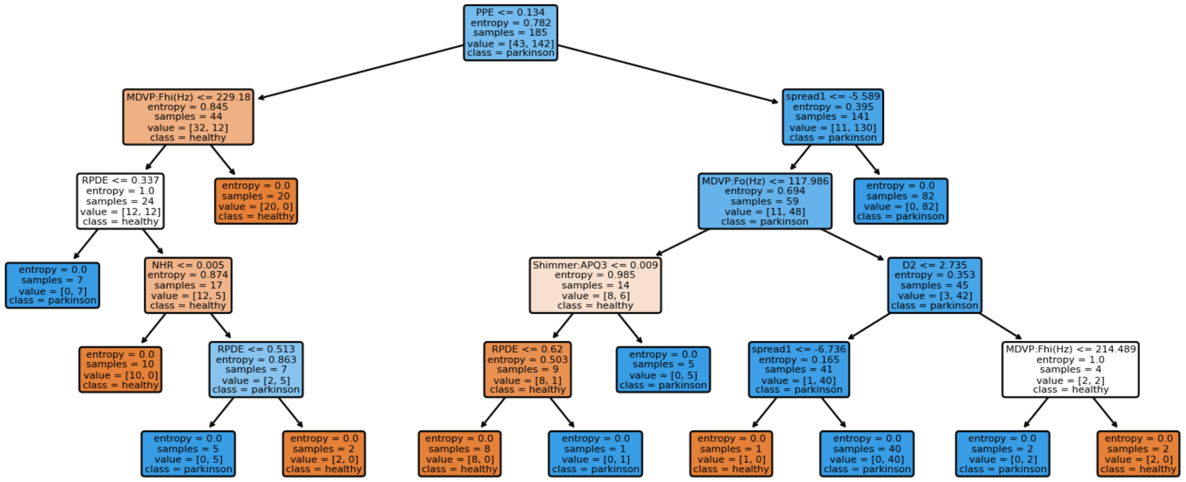
Based on the decision tree analysis for predicting Parkinson's disease, the tree depth reached five nodes. In pursuit of higher accuracy, we tried changing the max_depth function and we found out the tree will never go over 6 nodes even if the max depth would be set over 6. In contrast, when the depth was set to lower than 5, the training accuracy decreased.

The tree created 13 leaf nodes indicating the tree's decision outcomes. Only 7 out of 22 attributes were used, highlighting their higher relevance in predicting the disease. Most likely, the other attributes had a lower information gain for their current data, therefore not being that relevant.

The confusion matrix revealed a tendency towards false positives over false negatives, which, while not ideal, is preferable in medical diagnostics to avoid missing potential Parkinson's cases. Our tree is more likely to misclassify a healthy person as a sick person rather than the other way around. It is preferred because our main focus is predicting the sick cases rather than preventing fake positive diagnoses. A sick person should start treatment as soon as possible, especially considering the progressive nature of Parkinsons.

The substantial accuracy gap between training (100%) and testing data (60%) would usually suggest overfitting, but in our case we can’t assume that, taking into consideration our really small testing sample. If the gap consists with bigger testing samples, overfitting would be a fair assumption.

Improving classifier performance could involve refining the tree's complexity, possibly by adjusting the max depth or the minimum number of samples required for a split, and expanding the testing dataset for a more robust evaluation. Additionally, pruning the tree would also help remove branches that do not help the efficiency. Changing the random state of the tree would result in more constant accuracies and a cross validation would create a better validation system before using the testing data.


PPE <= 0.134
entropy = 0.782
samples = 185
value = [43, 142]
class = parkinson
MDVP:Fhi(Hz) <= 229.18
entropy = 0.845
samples = 44
value = [32, 12]
class = healthy
spread1 <= -5.589
entropy = 0.395
samples = 141
value = [11, 130]
class = parkinson
RPDE <= 0.337
entropy = 1.0
samples = 24
value = [12, 12]
class = healthy
entropy = 0.0
samples = 20
value = [20, 0]
class = healthy
MDVP:Fo(Hz) <= 117.986
entropy = 0.694
samples = 59
value = [11, 48]
class = parkinson
entropy = 0.0
samples = 82
value = [0, 82]
class = parkinson
entropy = 0.0
samples = 7
value = [0, 7]
class = parkinson
NHR <= 0.005
entropy = 0.874
samples = 17
value = [12, 5]
class = healthy
Shimmer:APQ3 <= 0.009
entropy = 0.985
samples = 14
value = [8, 6]
class = healthy
D2 <= 2.735
entropy = 0.353
samples = 45
value = [3, 42]
class = parkinson
entropy = 0.0
samples = 10
value = [10, 0]
class = healthy
RPDE <= 0.513
entropy = 0.863
samples = 7
value = [2, 5]
class = parkinson
RPDE <= 0.62
entropy = 0.503
samples = 9
value = [8, 1]
class = healthy
entropy = 0.0
samples = 5
value = [0, 5]
class = parkinson
spread1 <= -6.736
entropy = 0.165
samples = 41
value = [1, 40]
class = parkinson
MDVP:Fhi(Hz) <= 214.489
entropy = 1.0
samples = 4
value = [2, 2]
class = healthy
entropy = 0.0
samples = 5
value = [0, 5]
class = parkinson
entropy = 0.0
samples = 2
value = [2, 0]
class = healthy
entropy = 0.0
samples = 8
value = [8, 0]
class = healthy
entropy = 0.0
samples = 1
value = [0, 1]
class = parkinson
entropy = 0.0
samples = 1
value = [1, 0]
class = healthy
entropy = 0.0
samples = 40
value = [0, 40]
class = parkinson
entropy = 0.0
samples = 2
value = [0, 2]
class = parkinson
entropy = 0.0
samples = 2
value = [2, 0]
class = healthy