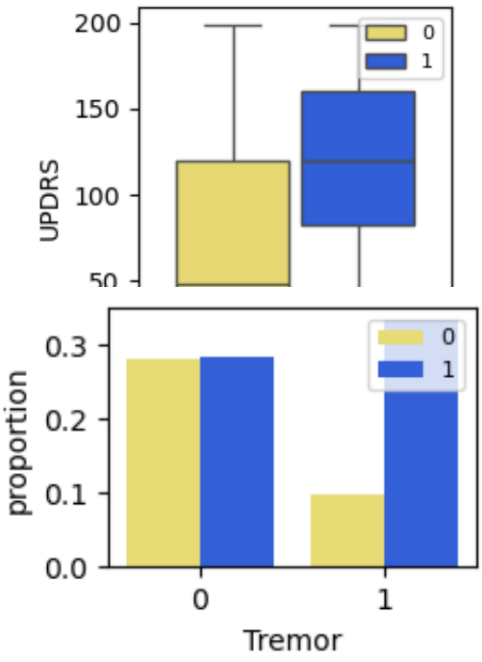
Data 622: Machine Learning and Big Data
Assignment 2
Dirk Hartog

Decision Trees are one of many supervised classification techniques used to predict categorical outcomes. An expansion of the decision tree is a random forest model, which constructs many decision trees, combining their outputs into a single result. Detecting the presence or absence of a disease is one area where decision trees and random forest models can assist doctors in the decision making process. The use of decision trees doesn't come without some drawbacks. The article titled "The GOOD, The BAD & The UGLY of Using Decision Trees" describes advantages, disadvantages and problems with conventional usage of decision trees and how these may relate to the consideration decision tree models for a particular project.

For this project I used the Parkinson's Disease Data set that can be found at https://www.kaggle.com/datasets/rabieelkharoua/parkinsons-disease-dataset-analysis. The Parkinson's data set has a relatively small number of observations (2105) with many features (33) including a diagnostic variable that is labeled with a 1 for the presence of parkinson and a 0 representing the absence of Parkinson's disease. Two different decision trees and a random forest model were built using this data set to evaluate the accuracy and variance between models. Evaluating the performance of a model should always look at trying to optimize the model for better accuracy and usability. Addressing some of the issues described in the article will provide insights to why a decision tree might be appropriate and how to optimize the models to mitigate some of its weaknesses.

The initial exploratory data analysis revealed that for the numeric categories the distributions within each class of the target variable were very similar. The feature that stood out the most was the distribution of scores in the UPDRS (Unified Parkinson's Disease Rating Scale) a questionnaire that evaluates various aspects of the disease, where higher scores relate to greater severity of the disease. In the Parkinson's group the median value was 120 versus in the no parkinsons group the median score was 48.



When looking at the categorical values a similar finding was visualized. The proportions of observations with and without parkinsons within each level of the categorical variable matched the proportions of observations with and without the disease (Fig. 2), except within the Tremor variable. These categories are highlighted as the two decision trees that were created used them as the root node.

The decision trees and random forest models were built using the python library scikit.learn. When building the decision trees, a max depth of the tree was set to 3 and the splitter parameter was set to 'best' which uses the feature that will have the best initial split of the data. In the second decision tree the maximum depth parameter was kept at 3 but the splitter parameter was set to 'random' in an attempt to build the tree with a different feature as a root node. If the root node happened to be the same as the first one, the model was run again until a different feature was chosen. For both decision tree models, a decision tree visual was generated and a confusion matrix was constructed. Cross validation was run on both decision trees to improve accuracy. For the random forest model a random grid search was used to find the best parameters for the number of trees (between 10 and 50) with a maximum depth (between 1 and 5). These ranges were chosen to keep the depth of trees somewhat consistent to the individual decision tree models. The output metrics for all models are reported in the table below.

	Decision Tree_UPDRS	Decision Tree_Tremor	Random Forest
F1_score	0.852071	0.813049	0.854260
Sensitivity	0.918367	0.826531	0.971939
Specificity	0.612500	0.662500	0.662500
Precision	0.794702	0.800000	0.762000
Accuracy	0.802215	0.764241	0.794304
Cross_val Accuracy	0.81	0.750	

As we can see the model with the best accuracy was the decision tree that used UPDRS as the first split. When forcing the second decision tree to use a different feature for the root node the accuracy of the model declined. Small increases in specificity of the subsequent models were noted but not enough to warrant a significant improvement. Surprisingly, the random forest model performed just as well as the first model, but was the best at identifying true positives as seen with almost 97.2% sensitivity. Most of the models had high sensitivity and low specificity. This indicates the model was much better at identifying the positive cases versus the negative ones. This might be related to a class imbalance issue where the trained model was biased against the majority class. Cross validation revealed a slight improvement in the first model's accuracy over the five iterations but on a lower accuracy for the second model's performance.

There are some disadvantages of using a decision tree for this data set that were mentioned in the article "The GOOD, The BAD & The UGLY of Using Decision Trees". When a decision tree becomes very large, the ability to interpret the model can be difficult. Fortunately, the potential to prune branches and select features allows for the flexibility to control the complexity and improve the models interpretability. Another drawback is that decision trees tend to have high variance and over fit the training data. In this specific case of identifying Parkinson's, it may be acceptable that we misclassify negative cases as positive (type I error) as missing a positive might be detrimental in applying early interventions. Decision trees do not require a lot of computational power and as information about factors relating to Parkinson's disease become more apparent a decision tree can be updated easily. The ease at which decision trees can be modified and evolve can prevent it from becoming ineffective over time.

Decision trees and random forest models can be effective ways in assisting the decision making process by objectively providing a logical pathway to an outcome. They can be a good first line model to introduce to an audience that may not have a technical background and could easily understand the decisions made at each step. Parameters of the tree or forest can be modified to address the issue of overfitting and they are easy to update, run and analyze as understanding of topics continues to evolve.