Random Forest in Classification of SLI/ASD/TD

Moving towards accurate and early prediction of language delay with network science and machine learning approaches

https://www.nature.com/articles/s41598-021-85982-0

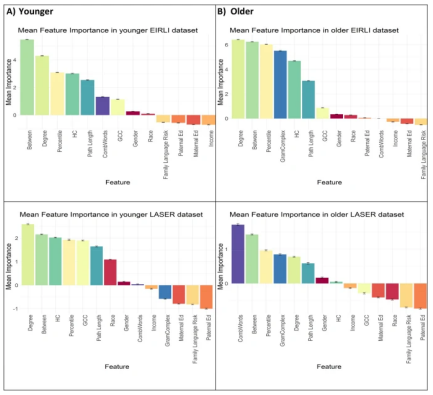
What is Random Forest:
First, to understand how it works, we need to understand how random forest is made of. Random Forest (RF) models are constructed by assembling a collection of individual decision trees, each of which is considered a "weak learner." A "weak learner" is a simple model that, on its own, may not be very accurate and is susceptible to overfitting (fitting the training data too closely) and overgeneralization (not performing well on new, unseen data). In the context of RF, these individual decision trees are intentionally kept simple and prone to these issues. In the RF approach, the inherent weakness of each individual decision tree is counteracted by combining many of them into a forest. This ensemble approach leverages the collective strength of multiple weak learners to create a more robust and accurate model. The randomness introduced during the tree-building process in RF helps to diversify the individual trees, making the overall model less prone to the shortcomings associated with overfitting and overgeneralization. Thus, overall, Random Forest (RF) models are made by creating a group of individual decision trees, each of which is a "weak learner" prone to overfitting and overgeneralization.

Why Random Forest?:
In the filed of the research about classification of language imparament, Random Forest have been the most efficeinet as it was “outperforming over 100 other classifier algorithms in a variety of datasets and testing conditions.” The advantage shared about Random Forest was that it is very accurate classifications without needing a lot of assumptions about how the data is organized. Specifically, with the nautre of Random Forest, RF approaches help us understand which variables are most important for accurate classification. By systematically adding or removing single variables (features) and comparing model accuracy, we can figure out how much each feature contributes to the overall accuracy.

How does Random Forest Work?
Random Forest (RF) models are made by creating a group of individual decision trees, each of which is a "weak learner" prone to overfitting and overgeneralization. These trees are built on random orderings and subsets of variables within the dataset, creating "split points" at each tree based on binary predictor thresholds. Each random subset of predictors is called a "node," and predictor thresholds are chosen to optimize model accuracy. This process is repeated to produce trees of varying "depths" with multiple nodes, forming a single decision tree. This entire process is repeated multiple times (e.g., 500 times in our study), creating a "forest" of trees that, when combined, forms a robust "stronger learner" model less prone to overfitting. Thus, in all, the important features would be

- **Random Subset of Variables**: Making the splits to be differnet for every tree that will be under the forest
- **Binary Predictor Threshold**: How the data will be split in different branches
- **Nodes and Dephts**

Research findings:
The research using Random Forest was finding as such approach yielded robust and reliable predictions of later LL outcome with classification accuracies in single datasets exceeding 90%. Furthermore, the research also found the ranking of features as they conducted the Random Forest research as it shows beside:



Comparison of Machine Learning Algorithms for Autism Spectrum Disorder Classification

https://www.atlantis-press.com/proceedings/ijcse-20/125946383

Reserach Finding:

To detect Autism Spectrum Disorder, the classification process in this research involves applying specific parameters to each algorithm, and their performance in classifying the ASD dataset was assessed. After comparing the specificity and sensitivity values, it was determined that the **Random Forest algorithm**, when utilizing all features, outperforms other algorithms in effectively classifying ASD in children and adolescents.

Radom Forest:

Random Forest is designed to be highly efficient and accurate for the classification process. By employing the concept of bootstraps, it generates diverse versions of the training data through random sampling with replacement from the original set. This process results in the creation of distinct trees that collectively form a forest. In the classification phase, each tree predicts a class based on its specific training subset. Ultimately, the final prediction from the Random Forest is derived by consolidating the individual predictions made by all the trees.

To determine the overall result, the research utilized the bagging method, which involves combining predictions from multiple trees, each trained on a different bootstrap subset. This aggregation enhances the model's overall performance and robustness, aligning with the Random Forest's voting approach: determining the representative class based on the majority votes within the forest.