**Classification Problems**

### **Dataset 1: Fruit Classification**

*This dataset contains measurements of various fruits, including apples, oranges, and lemons. It includes quantitative features such as weight, width, length, and a color score representing an unspecified measure of fruit color.*

**Dataset description**

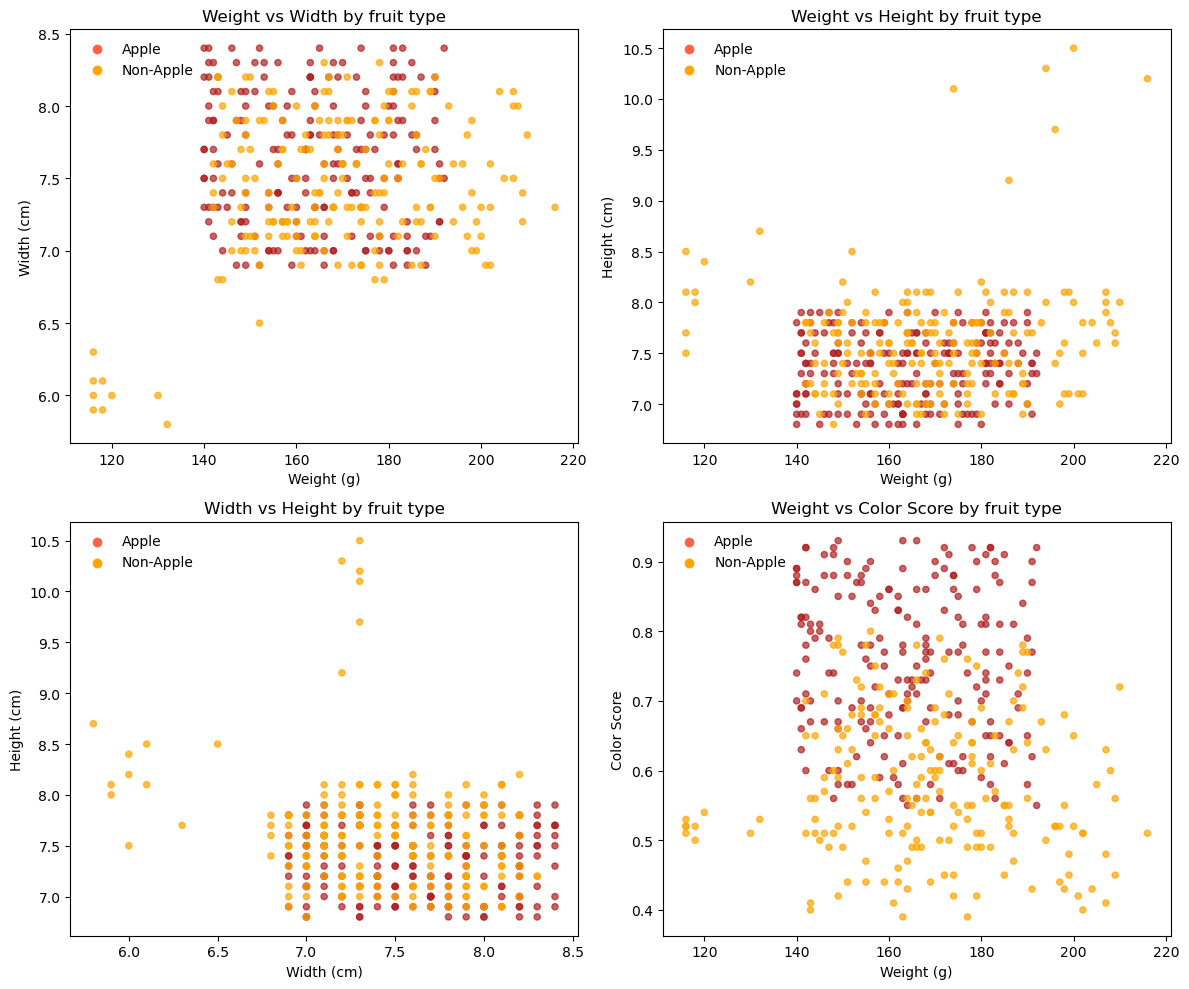
The dataset consists of **399** observations, with no missing values or duplicate entries. Table 1 below describes the features by fruit type. According to the descriptive statistics, there are some distinct differences between the fruits. Lemons are heavier and taller compared to apples and oranges. Apples and oranges have somewhat similar dimensions and weight. The color score ranges from 0.5 to 0.8[[1]](#footnote-0), with lemons being the lightest and apples being the darkest fruits.

***Table 1: Fruit features by type***

| **Fruit** | **Mean weight** | **Mean width** | **Mean height** | **Mean color score** | **Count** |
| --- | --- | --- | --- | --- | --- |
| apple | 164.2 | 7.6 | 7.3 | 0.8 | 199 |
| non-apple | 168.8 | 7.4 | 7.6 | 0.6 | 200 |
| *lemon* | *171.9* | *7.3* | *7.8* | *0.5* | *100* |
| *orange* | *165.6* | *7.5* | *7.4* | *0.6* | *100* |
| Total | 166.5 | 7.5 | 7.5 | 0.7 | 399 |

The scatterplots below illustrate the relationships between different features divided into apples and non-apples. Apples exhibit a more compact distribution in size-related characteristics, with fewer outliers in weight, width, and height, forming a relatively tight cluster[[2]](#footnote-1). In contrast, non-apples display greater variability, suggesting less consistency in their size and weight. While apples tend to have higher color scores, the plots do not indicate a clear relationship between fruit size[[3]](#footnote-2) and color. The scatter plots are in line with descriptive statistics and also demonstrate differences between the fruits (apples versus non-apples) in size, weight and color, assuming that these features can be used to classify fruits into categories.

**Figure 1. Fruit features by type**



**Fitting a binary classifier model**

Next, we fit a binary classifier to predict whether a fruit is an apple or not, using features such as weight, width, height, and color score, and evaluate the model's performance by quantifying its accuracy on a randomized training and testing dataset split. The data was split into 75% (299 entries) for training and 25% (100 entries) for testing, to ensure there are sufficient observations (sample) in the testing category. This split ensures that the model is trained on enough data (at least 100 entries) while also being tested on a separate set to avoid overfitting and ensure reliable performance. The following methods have been tested to determine the best model: Decision Tree, K-Neighbours, SVC, Random Forest, Logistic Regression. According to the results, the Logistic Regression demonstrated the best and well-balanced classification metrics (all being above 0.8) as well as best testing accuracy (0.84). The training accuracy was relatively high (0.78). The high testing score relative to the training score suggests the model is not overfitting and is able to generalize well to new data.

### The **Logistic Regression** model performance metrics are as follows:

### **Training Accuracy: 78%** (65/299 mislabeled points in the training dataset) The model has learned the training data well, with a relatively high accuracy on the training set.

### **Testing Accuracy: 84%** (16/100 mislabeled points in the testing dataset) The model performs excellently on the test data, showcasing strong generalization ability.

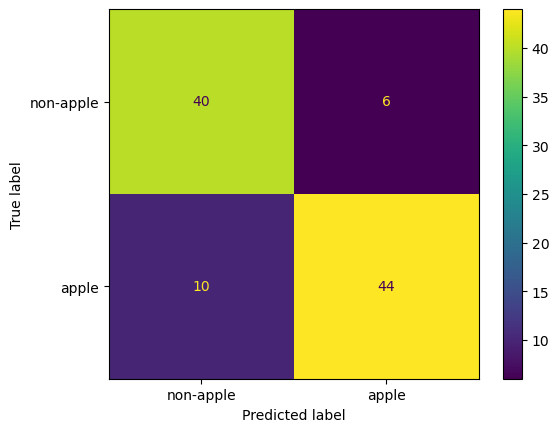
### **Precision: 88%** (True Positives / (True Positives + False Positives))When the model classifies the fruit as an apple, it is correct 88% of the time, indicating a low rate of false positives.

### **Recall: 81%** (True Positives / (True Positives + False Negatives))

### The model correctly identifies 81% of all the apples in the dataset, meaning it captures most of the apples, but misses a few.

### **F1 Score: 85%** (2 \* (Precision \* Recall) / (Precision + Recall)) The F1 score is a balance between how often the model correctly identifies apples (precision) and how many apples it correctly catches (recall). An F1 score of 85% shows the model is doing very well at both.

**Figure 2. Confusion Matrix: Testing**



**Confusion Matrix Breakdown:**

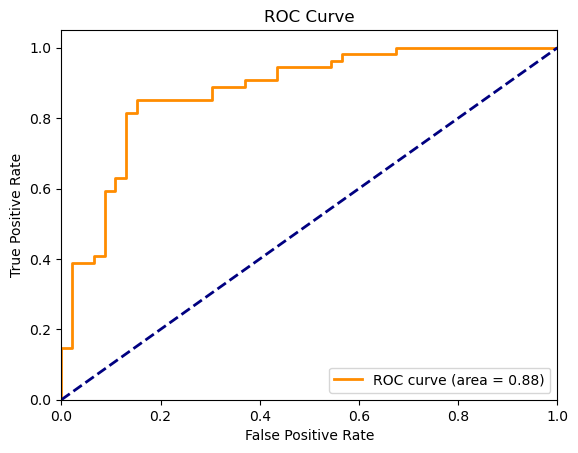
True Positives (TP): 44 correctly classified apples

True Negatives (TN): 40 correctly classified non-apples

False Positives (FP): 6 misclassified apples

False Negatives (FN): 10 misclassified non-apples

**Figure 3. ROC (Receiver Operating Characteristic) curve**



The area under the ROC curve is **0.88**, indicating high discrimination between the two classes (compared to the baseline - blue line), correctly classifying positive and negative cases and having a strong predictive power and performing much better than random guessing.

**Fitting a 3-class classifier model**

We then extend the model to perform multiclass classification, aiming to identify apples, oranges, and lemons. The same data split was used as in the previous model. Various classification methods were tested to determine the best-performing model, including Decision Tree, K-Neighbors, SVC, and Random Forest, all adapted for multiclass classification. Among the methods tested, the **Random Forest** classifier performed best for this three-class classification task. However, while it achieved perfect accuracy on the training set (1.00), its test accuracy was 0.72, suggesting potential overfitting. The high training accuracy indicates the model has memorized patterns in the training data rather than generalizing well to unseen data.

### The **Random Forest classification** model performance metrics are as follows:

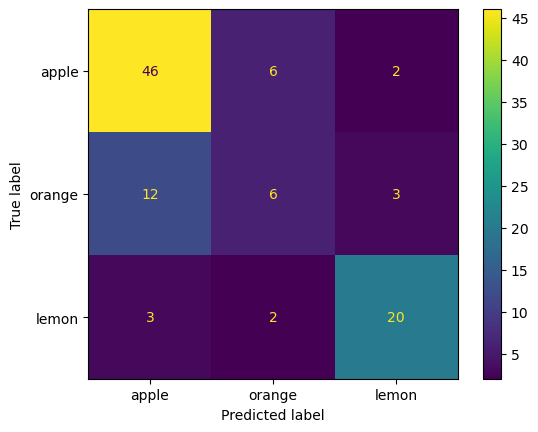
* **Training Accuracy:** 100% (0 mislabeled points in the training dataset)  
   The model classifies all training instances correctly, indicating overfitting.
* **Testing Accuracy:** 72% (28 mislabeled points in the testing dataset)  
   The model correctly classifies 72% of test instances, showing moderate generalization.
* **Precision:** 70%  
   On average, when the model predicts a specific fruit, it is correct 70% of the time.
* **Recall:** 72%  
   The model correctly identifies 72% of fruits, missing 28%.
* **F1 Score:** 70%  
   This balanced measure of precision and recall suggests a reasonable, though improvable, classification performance.

The table below shows the classification metrics for each class.

***Table 2: 3-class classification metrics***

| **Fruit** | **Precision** | **Recall** | **F1 score** |
| --- | --- | --- | --- |
| apple | 75% | 85% | 80% |
| lemon | 80% | 80% | 80% |
| orange | 43% | 29% | 34% |

According to the results demonstrated by the confusion matrix below and the classification metrics, the Random Forest model performs well with apples, achieving high precision (75%) and recall (85%), resulting in an F1 score of 80%. However, the model struggles with oranges, where precision (43%) and recall (29%) are notably low, leading to a low F1 score (34%). The model performs reasonably well with lemons, achieving balanced precision and recall of 80%, resulting in an F1 score of 80%. Hence, while the model is effective at classifying apples and lemons, it needs improvement in orange classification, where it tends to misclassify more frequently. A potential strategy to improve performance could involve refining the model for better generalization, particularly for the orange class.

**Figure 4. Confusion Matrix: Testing**

**Apple:** True Positives (TP): 46 correctly classified apples  
 False Positives (FP): 15 non-apples misclassified as apples  
 False Negatives (FN): 8 apples misclassified as other fruits

**Orange:** True Positives (TP): 6 correctly classified oranges  
 False Positives (FP): 8 non-oranges misclassified as oranges  
 False Negatives (FN): 15 oranges misclassified as other fruits

**Lemon:** True Positives (TP): 20 correctly classified lemons  
 False Positives (FP): 5 non- lemons misclassified as lemons  
 False Negatives (FN): 5 lemons misclassified as other fruits

### **Dataset 2: Dyslexia Classification**

*This dataset contains cognitive response data from an experiment. It includes measures of participants' responses to stimuli under orthographic and semantic conditions, capturing patterns in recognizing ‘same,’ ‘related,’ or ‘new’ words across different contexts[[4]](#footnote-3).*

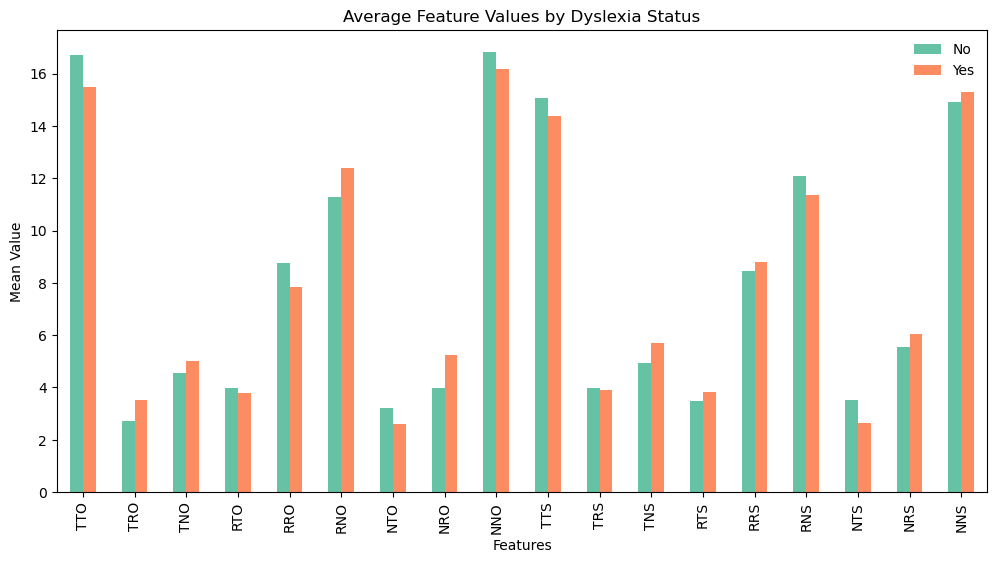
**Dataset description**

The dataset includes **71** entries (39 non-dyslexic and 32 dyslexic individuals), with no missing values or duplicate entries[[5]](#footnote-4). The complete list of measures tested, considering whether the participant has dyslexia, is as follows:

| **Orthographical Conditions** | **Semantical Conditions** |
| --- | --- |
| TTO - Number of answers ‘the same’ on target | TTS - Number of answers ‘the same’ on target |
| TRO - Number of answers ‘related’ on target | TRS - Number of answers ‘related’ on target |
| TNO - Number of answers ‘new’ on target | TNS - Number of answers ‘new’ on target |
| RTO - Number of answers ‘the same’ on related stimuli | RTS - Number of answers ‘the same’ on related stimuli |
| RRO - Number of answers ‘related’ on related stimuli | RRS - Number of answers ‘related’ on related stimuli |
| RNO - Number of answers ‘new’ on related stimuli | RNS - Number of answers ‘new’ on related stimuli |
| NTO - Number of answers ‘the same’ on new stimuli | NTS - Number of answers ‘the same’ on new stimuli |
| NRO - Number of answers ‘related’ on new stimuli | NRS - Number of answers ‘related’ on new stimuli |
| NNO - Number of answers ‘new’ on new stimuli | NNS - Number of answers ‘new’ on new stimuli |

The below figure demonstrates the mean score of each measure (18 in total) by whether the participant has dyslexia. The graph demonstrates that some measures have higher means for those who have dyslexia (i.e TTO, RTO, RRO, etc.) while for others it’s the opposite (i.e. TRO, RNO, NRO, etc) with no apparent trends between the categories (i.e. same, related, new target/stimuli).

**Figure 6. Dyslexia features mean scores**

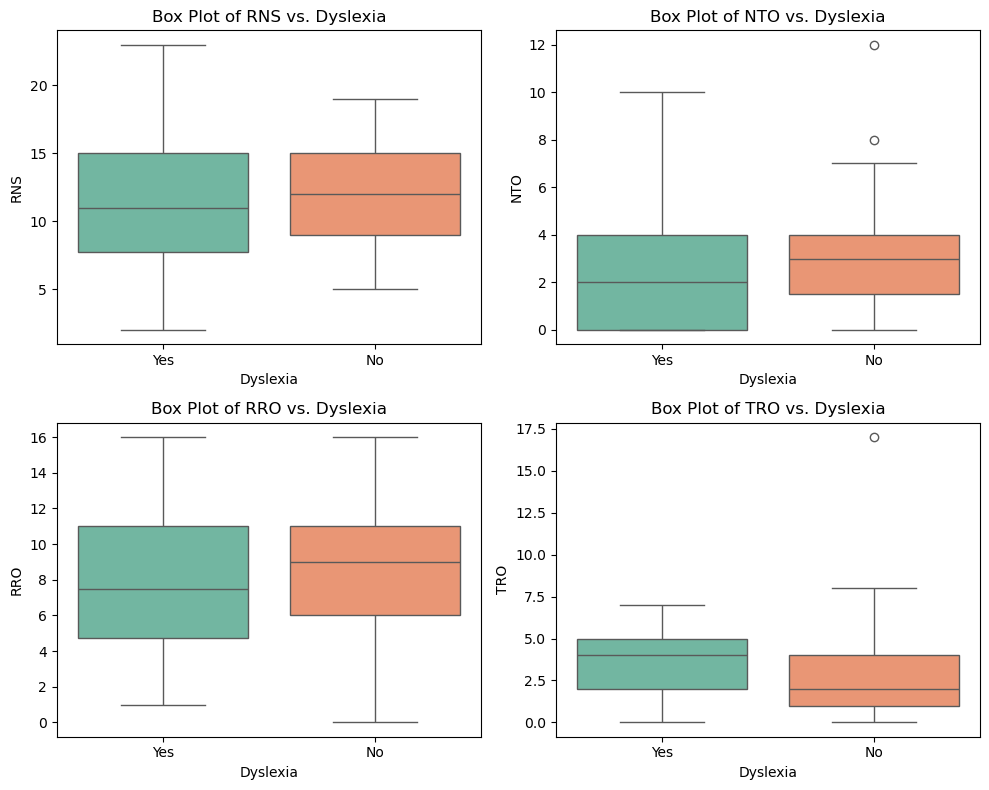
****

Given the large number of measures and the small dataset (71 cases) that may hinder achieving a high-accuracy model[[6]](#footnote-5), we’ll focus on the most relevant variables based on prior research. According to Obidzinski’s[[7]](#footnote-6) decision tree model, the following variables are key in distinguishing between dyslexia and non-dyslexia groups based on memory performance:

* **RNS (Number of answers ‘new’ on related stimuli in the semantical condition)**: The most important predictor, a critical aspect of memory processing in dyslexia.
* **NTO (Number of answers ‘target’ on new stimuli in orthographical condition)**: The second most important predictor, a key sign of dyslexia-related memory deficits.
* **RRO (Number of answers ‘related’ on related stimuli in the orthographical condition)**: A significant predictor, assessing the ability to process related stimuli in the orthographical condition.
* **TRO (Number of answers ‘related’ on target stimuli in the orthographical condition)**: Although the least important, it still provides valuable insight into memory performance related to target stimuli.

Hence, we will focus on these four variables for the classification model. The figure below shows the mean scores for the selected variables by dyslexia status.

**Figure 7. Main predictor features by dyslexia status**



The boxplots illustrate differences in selected variables based on dyslexia status. Individuals with dyslexia exhibit lower and more variable scores for the mean number of answers ‘new’ on related stimuli in the semantical condition, ‘target’ on new stimuli in the orthographical condition, and ‘related’ on related stimuli in the orthographical condition. In contrast, the mean number of answers ‘related’ on target stimuli in the orthographical condition is significantly higher for those with dyslexia than for those without.

**Fitting a binary classifier model**

Next, we fit a binary classifier to predict whether the individual has dyslexia based on the selected measures and evaluate the model's performance by quantifying its accuracy on a randomized training and testing dataset split. The data was split into 75% (53 entries) for training and 25% (18 entries) for testing. The following methods have been tested to determine the best model: Decision Tree, K-Neighbours, SVC, Random Forest, Logistic Regression. According to the results, the Decision Tree demonstrated the best and well-balanced classification metrics (all being close or above 0.7), which, given the small sample, is a good result.

The **Decision Tree** model performance metrics are as follows:

### **Training Accuracy: 100%** (0 mislabeled points in the training dataset) The model has learned the training data well, with a relatively high accuracy on the training set.

### **Testing Accuracy: 67%** (6/18 mislabeled points in the testing dataset) The model works relatively well on new data, showing good generalization.

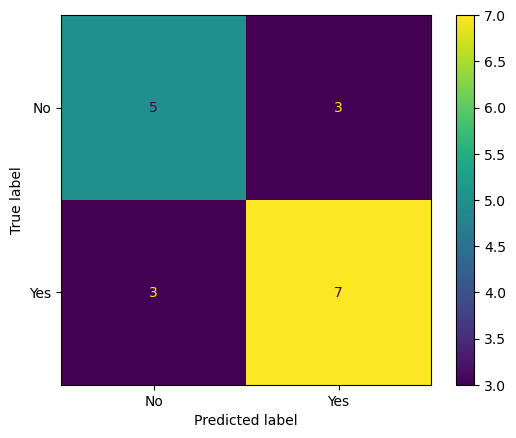
### **Precision: 70%** (True Positives / (True Positives + False Positives))When the model identifies dyslexia, it is correct 70% of the time, meaning false positives are low.

### **Recall: 70%** (True Positives / (True Positives + False Negatives))

### The model finds 70% of the dyslexic cases, but misses some.

### **F1 Score: 70%** (2 \* (Precision \* Recall) / (Precision + Recall)) The F1 score shows the balance between how well the model identifies dyslexia (precision) and how many dyslexic cases it catches (recall). A score of 70% means the model performs fair at both.

**Figure 8. Confusion Matrix: Testing**



**Confusion Matrix Breakdown:**

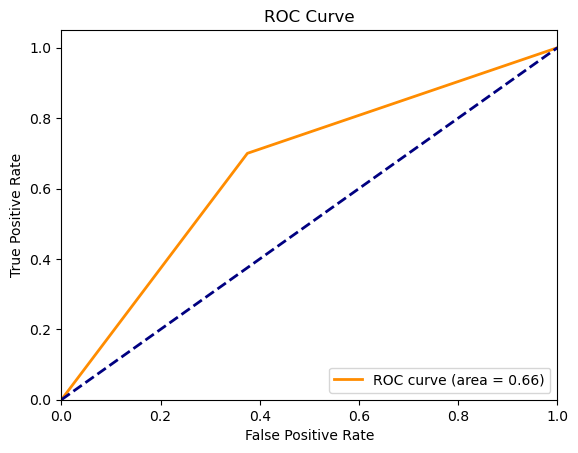
True Positives (TP): 7 correctly classified dyslexia cases

True Negatives (TN): 5 correctly classified non-dyslexia cases

False Positives (FP): 3 misclassified dyslexia cases

False Negatives (FN): 3 misclassified non-dyslexia cases

**Figure 9. ROC (Receiver Operating Characteristic) curve**



The area under the ROC curve (AUC) is 0.66, which indicates relatively low discrimination between the two classes. It suggests that the model has some ability to distinguish between positive and negative cases compared to the baseline (blue line), performing better than random guessing but could still be improved for stronger predictive performance.

1. The color scheme is an undefined variable, however, we conclude it goes from light to dark based on the color of lemons and oranges. [↑](#footnote-ref-0)
2. The outliers are not extreme and will not be removed for this classification. [↑](#footnote-ref-1)
3. Weight and Height have the same relationship with color scheme, so only weight is scattered across color as one of the size measures. [↑](#footnote-ref-2)
4. Source: <https://osf.io/n752f/> [↑](#footnote-ref-3)
5. The dataset contains several outliers on 8 measures, however, given the overall size of the dataset, no cases will be removed. [↑](#footnote-ref-4)
6. The models with all the variables have very low accuracy (i.e, a Logistic Regression Classifier Model using all the features resulted in a model with Training score: of 0.79 and a Testing score of 0.44. [↑](#footnote-ref-5)
7. Obidziński, M. (2020). Response frequencies in the conjoint recognition memory task as predictors of developmental dyslexia diagnosis: A decision-trees approach. *Dyslexia, 26*(2), 157-167. https://doi.org/10.1002/dys.1655 [↑](#footnote-ref-6)