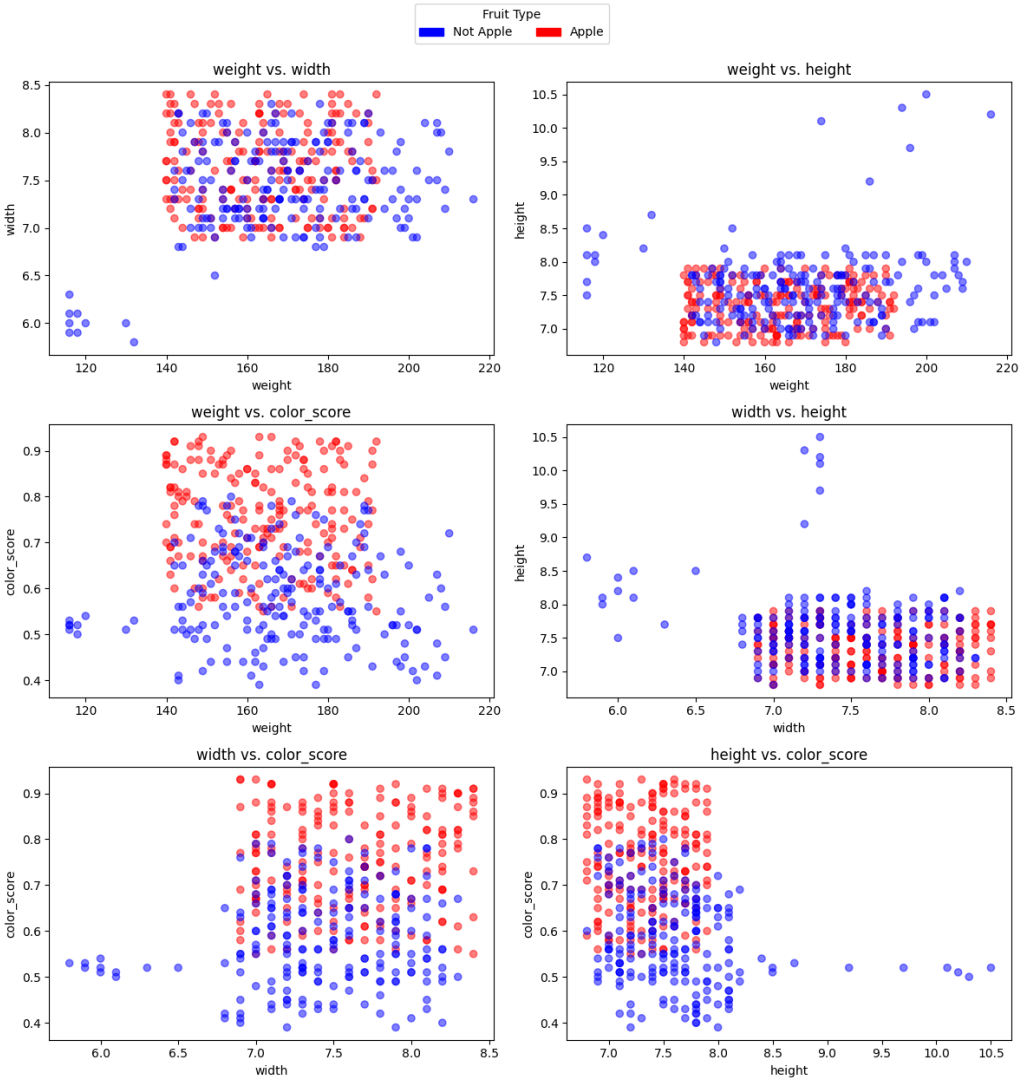
Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Fruit Classification and Dyslexia Detection: A Preliminary Machine Learning Analysis



PART I

To the left is the first dataset, this contains 3 different fruits – apples, oranges and lemons. There are 399 rows with 6 columns: ***fruit_name***, ***weight***, ***width***, ***height*** and ***color_score***. The dataset shows a significant class imbalance, with 199 apples compared to 100 oranges and 100 lemons. However, the data is relatively balanced for the binary classification of apples versus non-apples. Visualising the variables reveals distinct patterns that differentiate these classes, with varying degrees of clarity. Notably, plots of ***weight*** versus ***color_score*** and ***width*** versus ***color_score*** show some separation, suggesting colour score's potential importance. The dataset also has 12 outliers, which were removed as they fell outside the range of $\mu \pm 3\sigma$.

PART II

To determine the optimal binary classifier for identifying apples, models were trained to differentiate apples from non-apples. The features used for the model were ***weight***, ***width***, ***height***, and ***color_score*** (X data), while the target variable was the binary classification of ***fruit_name*** being equal to ***apple*** (y data). After splitting the data into test and training data (weighted 20% - 80% respectively to allow adequate training), I fitted the models to the training data. I evaluated three different classifiers: Decision Tree, K-Nearest Neighbours and Support Vector Machines (SVM), as these are 3 models that use different approaches and should excel in differing areas. Next the hyperparameters in the classifiers were then tweaked until the rate of improvement diminished, for example, the optimum 'max_depth' for the Decision Tree was 300 and the best Cost Parameter (c) value was 100 with gamma of 0.001.

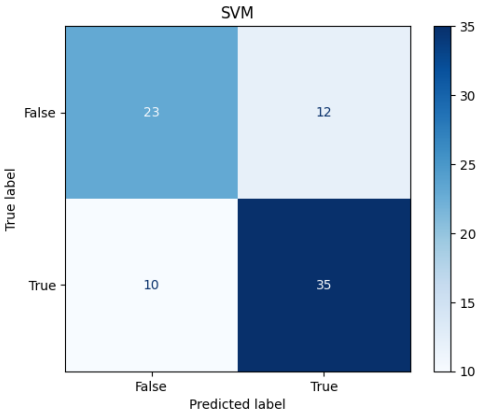
Model	Train Score	Test Score
Decision Tree	1	0.7
K-Nearest Neighbours	0.815	0.575
SVM	0.743	0.725

The results reveal a clear performance disparity among the classifiers. The Decision Tree exhibited the best performance on the training data, although a score of 1 is typical of overfitting despite the positive score of 0.7 on the test data. While K-Nearest Neighbors scored well on the training data it exhibits clear signs of overfitting from its substantially lower test score of ≈ 0.575, suggesting it struggles to generalise beyond the training data's specific patterns. Finally, SVM, which despite having the lowest score on the training data, managed to achieve the best score on the test data – higher than its training score.

This is good evidence of the model generalising the data, indicating that it has learned underlying patterns rather than simply memorising the training data. However, it is important to analyse these results further, particularly looking at the F1 score and Confusion matricies; these are able to provide more quantative understanding of the results balancing precision and recall.

Looking at the F1 scores we see our results confirmed. SVM maintaining the highest score further confirms its ability to balance recall and precision. While K-Nearest Neighbours lower score reveals its inability to generalise the data; the F1 Score of 0.6 indicates a poor balance between precision and recall, suggesting the model is missing some positives while avoiding false positives.

Model	F1 Score
Decision Tree	0.73
K-Nearest Neighbours	0.6
SVM	0.76



Furthermore, the confusion matrix confirms the expected performance of the SVM, reflecting a balanced distribution of true positives and true negatives.

PART III

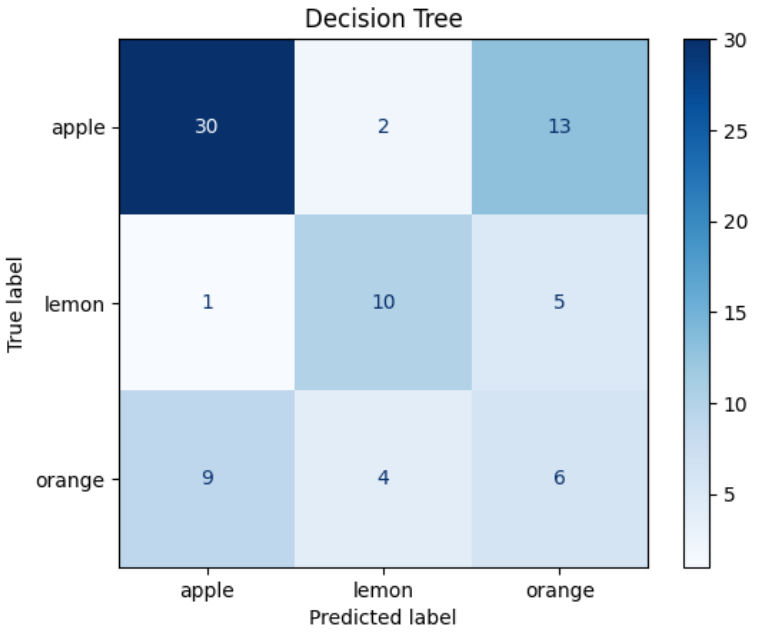
Next, I extended the analysis to a 3-class classification problem involving apples, oranges, and lemons. Employing the same three models and feature set as before, I aimed to identify the optimal classifier. While the features used remain the same as the previous model: ***weight***, ***width***, ***height***, and ***color_score*** (X data), the target variable has changed. Now I have implemented a 3-class classifier in the form of just the ***fruit_name*** variable (y data), this produced 3 different outcomes: ***apple, orange,*** and ***lemon***. Again, this data was then split into testing and training data using the same ratios as prior, before training the models on the training data. Afterwards we can evaluate them giving the following results:

Model	Train Score	Test Score
Decision Tree	1	0.575
K-Nearest Neighbours	0.746	0.512
SVM	0.605	0.637

The shift to multi-class classification yields significantly different results compared to the binary scenario. Now all 3 models are performing much worse overall. K-Nearest Neighbours still exhibits the poorest performance, indicating probable overfitting. Notably, the Support Vector Machine (SVM) and Decision Tree models present an intriguing contrast. While the Decision Tree achieves a higher test score, its perfect training score of 1 strongly suggests overfitting. Conversely, the SVM, despite a much lower test score, actually improves its test scores compared to its training score, demonstrating good signs of generalisation. This consistency implies that the SVM has effectively learned the underlying patterns of the data without memorising the training set, possibly making it the superior model. However, this hypothesis was later challenged when gaining deeper insights into model performance.

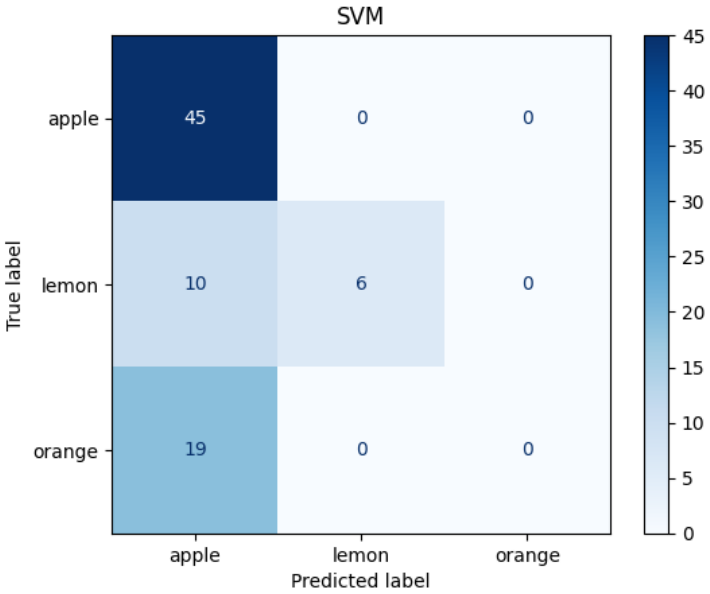
Model	F1 Score
Decision Tree	0.87
K-Nearest Neighbours	0.38
SVM	0.55

To do this F1 scores and confusion matrices were examined. Perhaps as expected K-Nearest Neighbours had an extremely low F1 score, due to its low precision score. Although looking at SVM and Decision Tree we see a surprise in the form of a large score disparity. This suggests that the previous hypothesis of SVM being able to generalise the data was false and perhaps there is an imbalance in the data. While Decision Tree, despite its lower F1 Score, may be the more suitable choice.



The Decision Tree model demonstrates strong performance, particularly in predicting ***apple*** instances, as evidenced by its confusion matrix. However, it exhibits slightly lower accuracy for ***lemon*** and ***orange*** with ***orange*** being the least accurately classified. The skewed distribution of the test data, heavily favouring ***apple*** is also apparent in the confusion matrix.

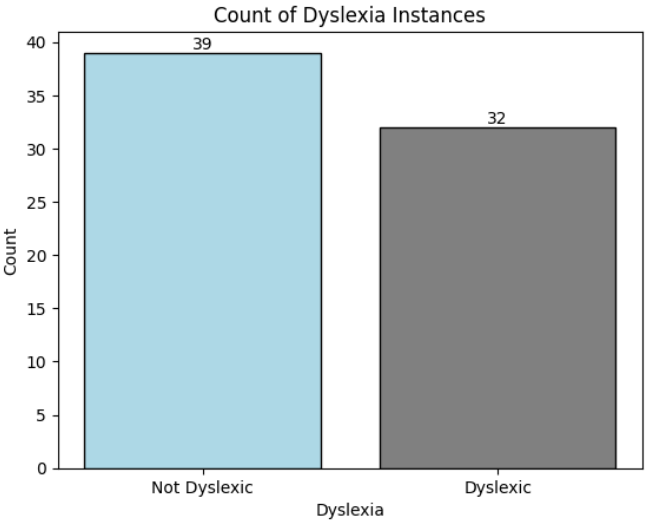
The SVM model further confirms this, revealing a significant performance disparity. Its confusion matrix highlights a critical flaw: it almost always predicts ***apple*** capitalising on the dataset's imbalance. This explains its seemingly high accuracy in the context of the skewed test data, but it fails to generalise to other classes.



In conclusion, the Decision Tree outperformed both SVM and K-Nearest Neighbours in the 3-class fruit classification task, demonstrating the most balanced performance. While SVM initially appeared promising, it was ultimately hindered by overfitting and poor generalisation, likely due to class imbalance. K-Nearest Neighbours exhibited significant overfitting and low overall performance, making it unsuitable for this dataset.

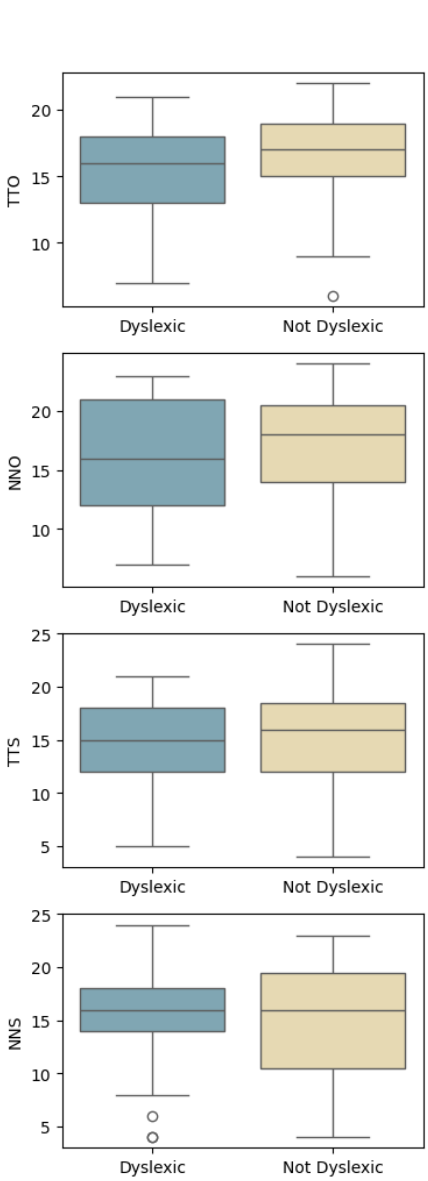
The Decision Tree's success highlights its ability to handle the complexities of multi-class fruit classification, despite its potential for overfitting in highly complex datasets. This analysis underscores the importance of evaluating models using a combination of metrics, including F1 scores and confusion matrices, to accurately assess performance and select the most appropriate classifier.

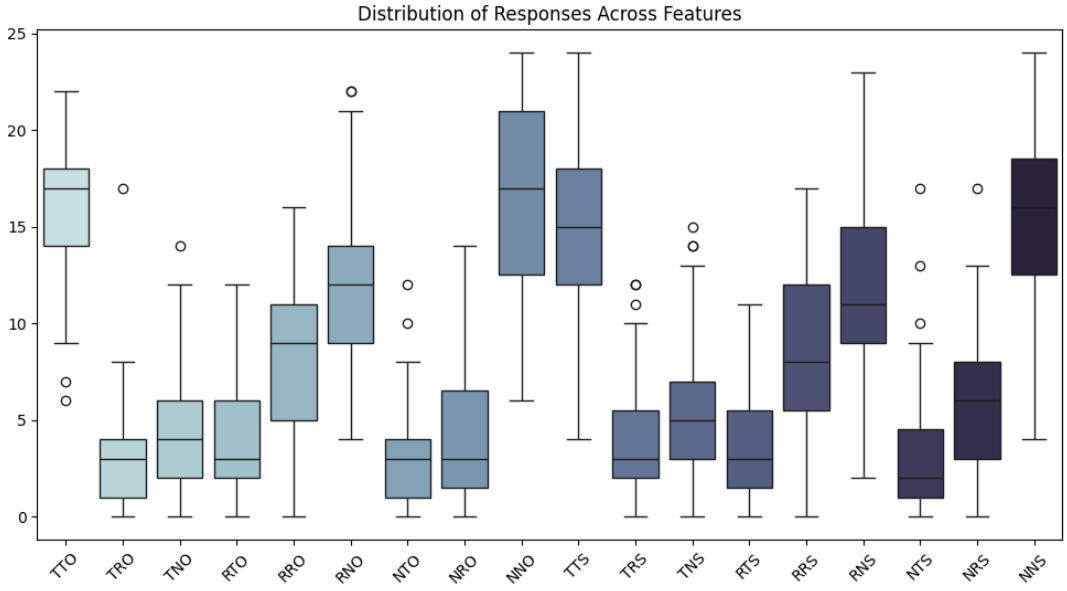
PART IV



Moving to the 'Response frequencies in the conjoint recognition memory task as predictors of developmental dyslexia diagnosis: A decision-trees approach' dataset from OSF Home. We see that the data is slightly weighted towards participants without dyslexia so that is something we must bear in mind throughout the analysis. Looking at the box plots of response frequencies from the memory task dataset, we can observe that TTO, NNO, TTS, and NNS appear to be among the more frequent responses. These variables indicate a tendency for participants to frequently identify items as either 'exactly the same' or 'completely new' across both visual and semantic-based judgments.

- TTO indicates correct identification of items as “exactly the same” based on visual similarity.
- NNO indicates correct identification of items as “completely new” based on visual differences.
- TTS indicates correct identification of new items as “exactly the same” based on semantics.
- NNS indicates correct identification of new items as “new” based on semantics.





It's possible that these higher frequencies suggest that participants were generally more consistent in their judgments of exact matches and complete mismatches compared to judgments of relatedness. However, further analysis would be needed to confirm this interpretation. We can further investigate the relationship between these variables by splitting up the box plots between dyslexic and not dyslexic. When taking this approach, we are able to see slight disparities in the mean and definitely in the standard deviation.

PART V

To classify developmental dyslexia based on the conjoint recognition memory task data, binary classifiers were trained. Obidziński' s original paper used a decision tree, citing the dataset's small size. To validate this choice, K-Nearest Neighbours and Support Vector Machine (SVM) models were also evaluated.

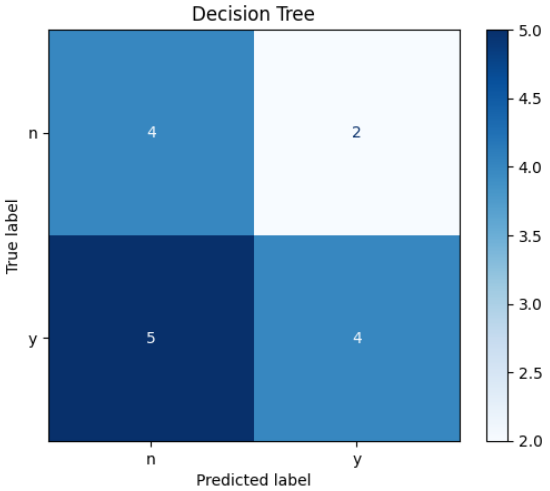
Model	Train Score	Test Score
Decision Tree	0.857	0.867
K-Nearest Neighbours	0.768	0.8
SVM	0.911	0.933

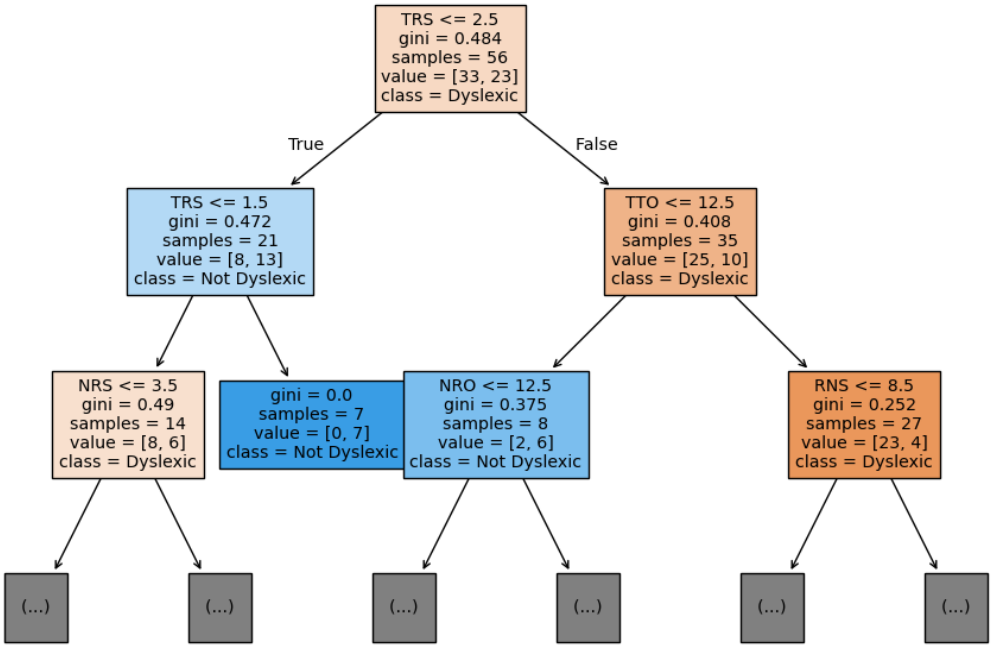
Initially, the scores appear promising, surpassing previous results in this report. All three models exhibit strong performance on both training and test sets, with SVM achieving the highest test score of 0.933. While this might suggest effective generalisation, further analysis of F1 scores and confusion matrices reveals a more nuanced picture.

Contrary to initial impressions, the F1 scores indicate that the Decision Tree, as suggested by Obidziński, is the most robust model for this dataset. This contradicts the first assumption that SVM and K-Nearest Neighbours were superior based on test scores alone.

Model	F1 Score
Decision Tree	0.53
K-Nearest Neighbours	0.43
SVM	0.47

The limited size of the dataset becomes strikingly apparent when examining the confusion matrix, which reveals a mere 20 total test samples. This small sample size significantly increases the potential for models to achieve high accuracy through random chance, rather than by accurately capturing the underlying relationships within the data.





To identify the key predictors contributing to the Decision Tree model's predictive power, feature importance analysis was conducted. The following outlines the model's decision-making process, illustrating how it quantified the association between each predictor and the presence of developmental dyslexia. The tree has been limited to a max depth of 3 as this led to optimal performance.

The Decision Tree's feature importance analysis reveals the following ranking of predictors, with their respective importance scores:

Feature	Importance
TRS	0.283
RNS	0.230
TTO	0.165
NRO	0.114
NRS	0.076
RRS	0.061
TNO	0.037
RRO	0.034

As shown, TRS, RNS, and TTO emerge as the most influential variables in the model's predictive process. Notably, TTO, which was previously observed to be a high frequency response, ranks third in importance, suggesting its significant role in the classification of developmental dyslexia.
It is also crucial to acknowledge that the remaining variables not listed; namely, TRO, TNS, RTS, RTO, RNO, NTO, NTS, and NNO were assigned an importance score of 0. This indicates that the Decision Tree model deemed these variables to have no discernible predictive value in distinguishing between individuals with and without developmental dyslexia.

The Decision Tree's focus on a select set of response frequencies indicates its ability to identify key cognitive markers of developmental dyslexia. While effective with this limited data, further research with larger datasets is needed to validate these findings and assess the model's ability to generalise the data efficiently.