

This is a simplified demonstration of the use of growth trajectory features extraction and use for the prediction of growth trajectories clusters. Different studies might adapt, or extend the algorithm below in order to be tailored to the needs of different research works.

1. Feature extraction algorithm

This process is executed for each anthropometric measure. It requires as input a dataframe with the longitudinal growth measurements of the anthropometric measure for all subjects in the population. For each population, the algorithm returns a data frame with the features from all 5 anthropometric measures (raw height, raw weight, standardized length, weight and bmi). A total of 15 features per measure are returned for each population.

1. For each subject in the dataset (i.e., each row in the input data frame)
 - a. Train a linear model = $\text{measure} \sim \text{time points}$
 - b. Calculate the conditional change as the difference in the residual sum of squares between the population and the local linear model (1 feature)
 - c. From the linear model, save the intercept and slope (2 features)
 - d. Integrate the linear model.
 - e. From the integrand, save the AUC and the tempo (2 features)
 - f. If the subject has more than 2 non-NA time points, train a quadratic model = $\text{measure} \sim (\text{time points})^2$
 - i. Calculate the conditional change as the difference in the residual sum of squares between the population and the local linear model (1 feature)
 - ii. From the quadratic model, save the intercept and the slope (2 features)
 - iii. Integrate the quadratic model.
 - iv. From the integrand, save the AUC and the tempo (2 features)
 - g. From the raw data, save the start and end values of the anthropometric measure (2 features), if there is +2SD (rapid growth) or -2SD (rapid decrease) of change in the anthropometric measure (3 features)

2. Growth Pattern Membership Prediction

This is the main process of prediction. This process is kept generic of the specific classification algorithm (Random Forest or XGBoost in this work), so it can be followed with any specific algorithm. It receives as input a data frame with 75 features for all 5 anthropometric measures for each population and a clustering based on LCMM (or any other preferred method), where the clusters have been labelled by experts (note: this step is optional and is used only to interpret the results).

1. Split the dataset in training-validation and test set (usually 75-25)
2. Set up a grid search 10-fold cross-validation training control to optimally configure the classification algorithm.
3. Select the hyperparameters to tune (e.g., number of variables at each split, number of trees and the maximum number of terminal nodes per tree for Random Forest, or the subsampling rate, the learning rate, the minimum sum of instance weight per child and the max depth of each tree for XGBoost)

4. Train the final model with the tuned hyperparameters
5. Calculate and plot the SHAP values for all or for the most important features for explainability.
6. Calculate the confusion matrix between the reference and predicted values for the test set.
 - a. From the confusion matrix, calculate the sensitivity (or recall), specificity, precision, accuracy and Kappa statistic. Using precision and recall, we can also calculate the F1-score as their harmonic average.
 - b. In case of multiple classes (i.e., more than 2 classes), metrics (except for accuracy and Kappa) are returned per class, instead of the entire classifier. In this case, we can report the average metric across all classes.
7. If more than 2 classes, for each class C
 - a. Create a “true class” vector, where 1 is when the reference class is equal to C , and 0 for any of the other classes.
 - b. Calculate the true positive rate and false positive rate between “true class” and the predicted classes.
 - c. Plot the ROC curve for the class C
 - d. Calculate the AUC of the ROC curve
 - e. (Optional) Calculate the calibrated probabilities of the classifier.
 - f. (Optional) Plot the original and calibrated probabilities.
 - g. (Optional) Compare the original and calibrated AUC.
8. (Optional for explainability) Extract the best tree based on the minimum error rate.
 - a. Compare the split values of the features present in the tree with the SHAP values calculated in step 5.