Human gut microbiome of children with ASD

Joao Fernando Marques

2022-11-19

Introduction

The dataset used in this report is from the research paper by Zhou Dan et al. published on April 21st of 2020 - Altered gut microbial profile is associated with abnormal metabolism activity of Autism Spectrum Disorder downloaded from kaggle.com. Autistic Spectrum Disorder (ASD) is a severe neurodevelopmental disorder that is primarily characterized by abnormal behavioral symptoms, and children with ASD often have gastrointestinal symptoms, such as gaseousness, diarrhea, and constipation. A few studies a small number of subjects have shown that individuals with ASD have different gut bacterial microbiota from typically developing (TD) individuals, and that early life perturbations of the developing gut microbiota can impact neurodevelopment and potentially lead to adverse mental health outcomes later in life through the gut-microbiome-brain axis. Thus, it still needs to be more thoroughly analyzed whether the abundance and diversity of altered bacteria indicates the potential interaction effects between ASD and gastrointestinal symptoms.

Classic methods of studying the bacterial microbiota have shortcomings, the main one being that not all bacteria can be cultured under standard laboratory conditions, but rapid advances in high-throughput sequencing technologies have profoundly changed the study of microbiomes across diverse environments. Next-generation sequencing (NGS) uses the sequence of 16S rRNA, a highly conserved region in bacterial genomes to identify bacteria within samples, but comes with several analytical challenges. The method generates various reads from the different 16S presents in the sample, generating data that is compositional, that is, only reflects the relative abundance of each bacteria in the sample. If a dominant bacteria increases in the sample, the proportion of all the others will decrease, even if their absolute abundance remain constant. Also, the number of reads per sample varies (read depth), so some form of data normalization is necessary, the data are often very sparse, and there are usually more predictors than samples. So these characteristics of microbiome data must be considered in statistical analyses.

Given this, the objective of this report is to classify individuals in ASD or TD based on their microbiome profile, as well to compare different machine-learning methods to random forest when used in microbiome data. Random forest was chosen as a start point because of the various machine-learning algorithm, it is able to identify non-linear relationships, deal with variable interactions, is robust to overfitting and works well with high dimensional data with low signal-to-noise ratio, such as microbiome data.

Methods

Data exploration

The raw file from kaggle has 1322 rows and 256 columns, but the predictors, the bacteria, are the rows. The first column identifies the bacteria by a short ID, called Operational Taxonomic Unit (OTU), the second column gives the taxonomic classification of each bacteria, and the other columns are the samples, with a code identifying if a sample is from ASD or TD individuals. In the paper is said that data were collected from 286 children, but there are only 254 samples in the kaggle database, so some bacteria that were only present in those missing 32 samples have 0 reads and have to be removed.

This file was separated in three data.frames, Data, Target and Taxonomy with the following code:

```
##### Read file #####
Raw <- fread("GSE113690_Autism_16S_rRNA_OTU_assignment_and_abundance.csv",
```

```
            stringsAsFactors = T)

# Create Taxonomy data.frame, with OTU and Taxonomy columns
Taxonomy <- Raw %>%
  select(OTU, taxonomy)

# Remove Taxonomy and OTU columns,
# transpose the results and
# rename the columns with the OTU column to cross-reference the taxonomy table.
# The rownames are then transformed into a column and the tags are
# recoded as ASD or TD (According to the supplementary material of the paper).
# Any bacteria without reads are then removed from the object.
Data <- Raw %>%
  select(-taxonomy, -OTU) %>%
  t() %>%
  data.frame() %>%
  `colnames<-`(Raw$OTU) %>%
  rownames_to_column("target") %>%
  mutate(target = str_remove_all(target, "\\d"),
         target = factor(target),
         target = fct_recode(target, ASD = "B", TD = "A"),
         ) %>%
  select(target, where(~ is.numeric(.x) && sum(.x) != 0))

# A new data.frame is created with the target column and
# this column is removed from the Data object.
Target <- select(Data, target)
Data <- select(Data, -target)

# The bacteria present in the taxonomy object is filtered based on the remaining
# bacteria in the dataset.
# The taxonomy column is separated in 8 columns, with each column representing
# a taxonomy rank.
Taxonomy <- Taxonomy %>%
  filter(OTU %in% colnames(Data)) %>%
  separate(taxonomy, sep = ";_",
           into = c("domain",
                    "kingdom",
                    "phylum",
                    "class",
                    "order",
                    "family",
                    "genus",
                    "species"))
```

After the data cleaning step, we ended with 254 samples with 1315 different bacteria observed in these samples. Figure 1 shows the number of samples in each of the groups, ASD or TD. Since all samples have the same read depth (3.1757×10^4 reads), we can use the abundance table direct in the analysis, otherwise we would need to process it further, rarefying the samples to equal read depth, or at least transforming it to relative abundance.

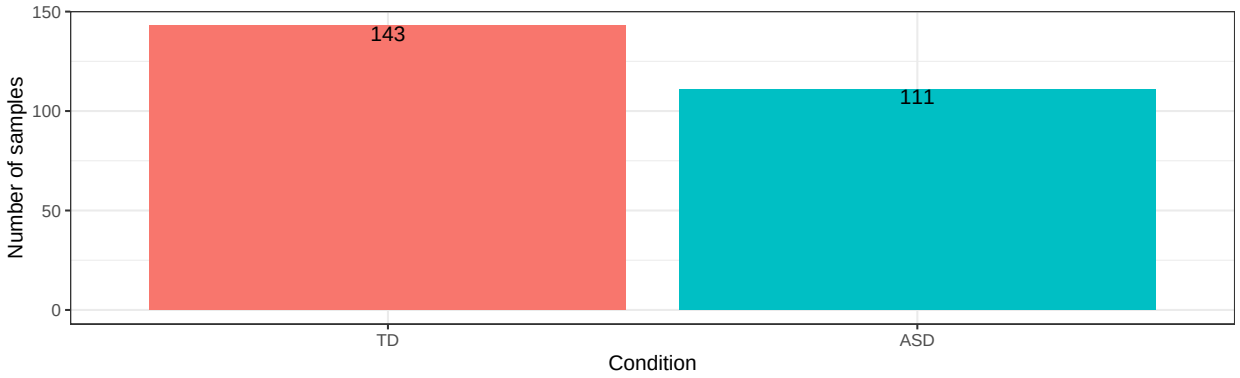

Figure 1: Number of samples in each group

We can explore relations of the bacteria observed with barplots of different taxonomy categories. Figure 2 shows the barplot for Phylum in each sample, but in this category rank it is difficult to see any difference, and is already difficult to see difference in the colors because of the number of phyla present.

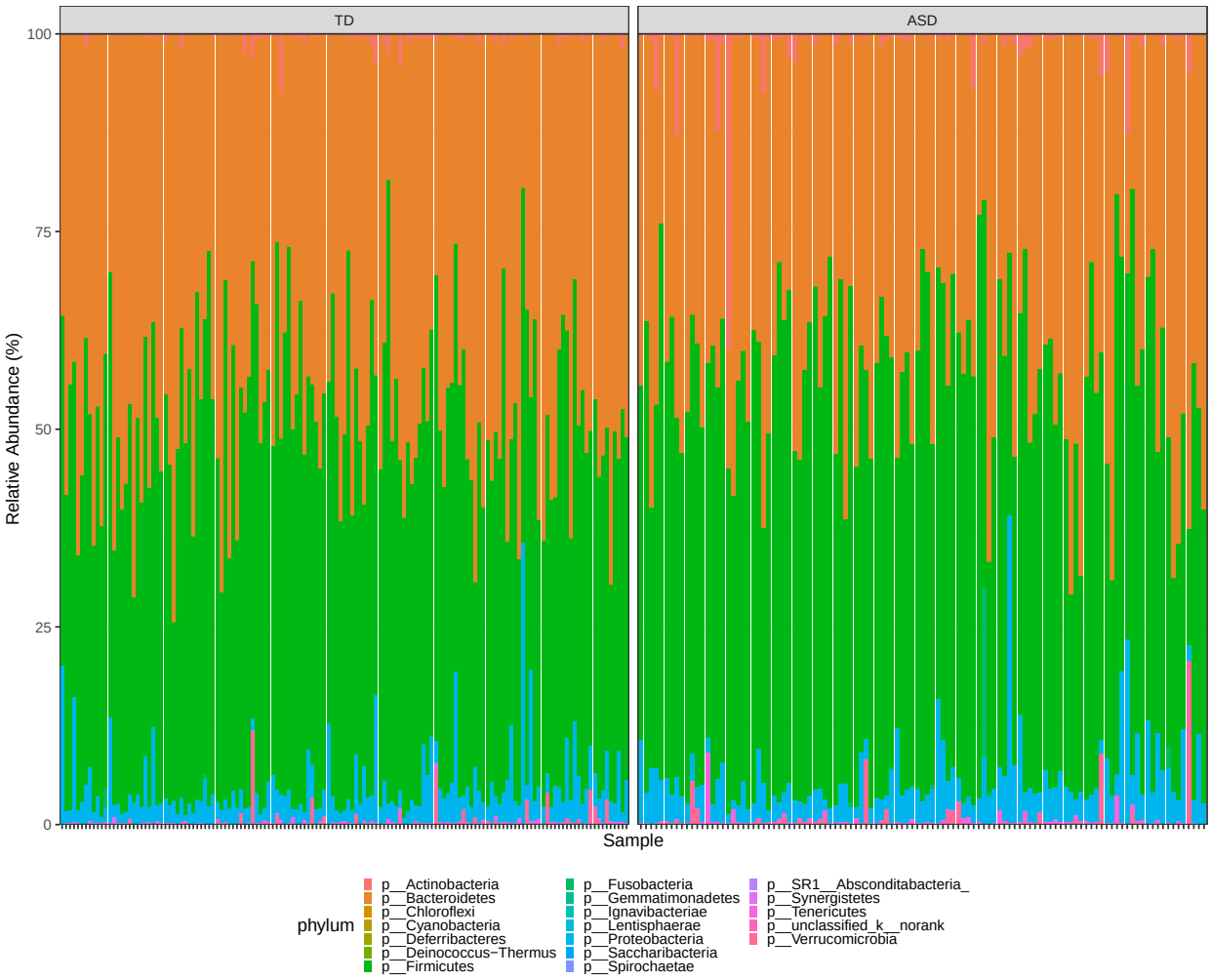

Figure 2: Stacked bar-plot representation of microbiota compositions with taxonomic features collapsed at the level of phyla

To inspect whether lower ranks show differences between groups we can use ordination maps. A widely used method is Principal Component Analysis (PCA), but here we come across another peculiarity of ecological data, not only for microbiome, but for all data that count species. These types of data are not euclidian, so we can't use any statistical method that relies in euclidian distance and have to use other methods. This happens because the value of "0" have a special meaning in ecological data. Think about three samples where you searched for a bird species. In sample "A" you did not see any birds, in sample "B" you saw 5 birds and in sample "C" you saw 150 birds. If we analyse these data with euclidian distance, sample "B" is much more similar to sample "A" than it is to sample "C" (distance from "B" to "A" is 5, and distance from "B" to "C" is 145), but for the real, living species, sample "B" is more similar to "C", because the species manages to survive in it, and does not survive in sample "A". Sample "A" could be from a desert, sample "B" from a forest edge, and sample C" from the forest interior. Figure 3 shows what happens if we apply PCA to the microbiome dataset.

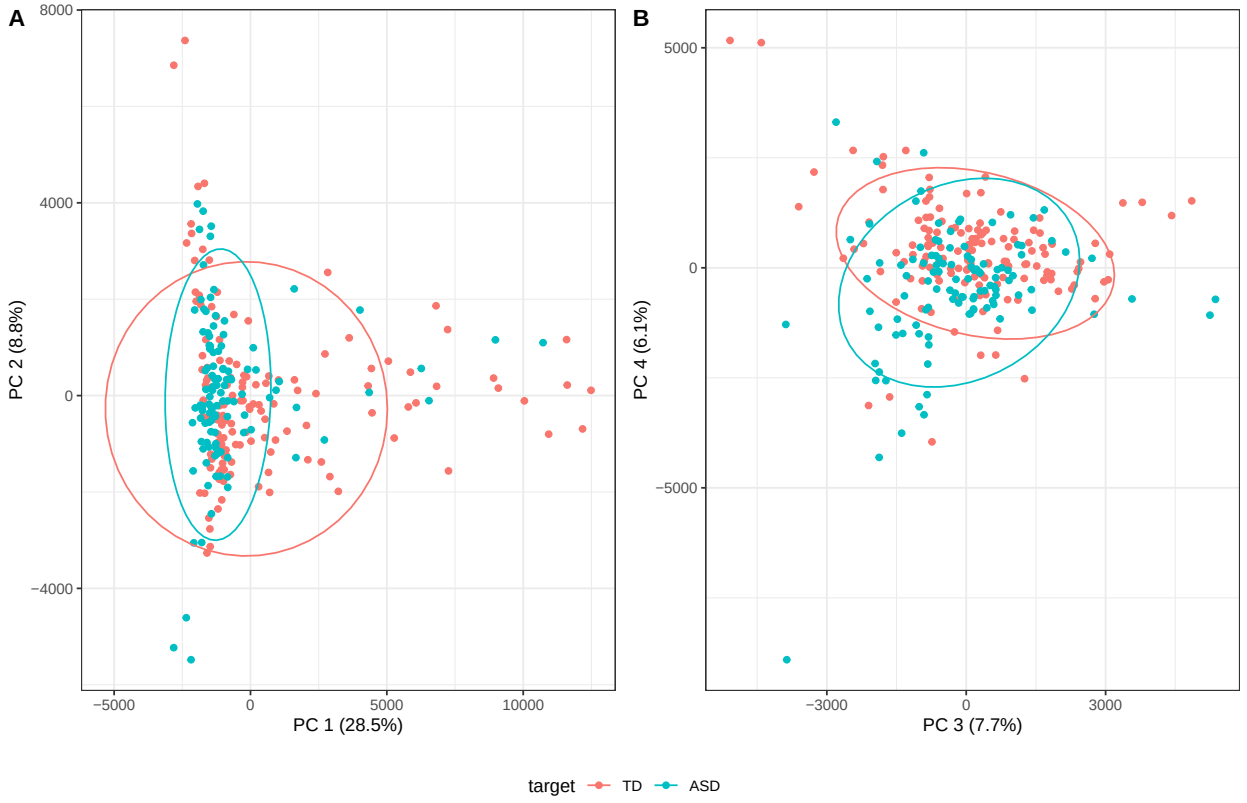

Figure 3: Principal Components Analysis of microbiome data. A: First and second principal components analysis axes. B: Third and fourth principal components analysis axes

To reduce the dimensionality of ecological data, we need to use dissimilarity measures, not Euclidean distances, to compare samples. Several dissimilarity indices are available, but the most common ones are the Jaccard dissimilarity, which considers the occurrence of species to quantify the similarity of the samples, and the Bray-Curtis dissimilarity, which also considers the abundance of species. We also need to use another ordination technique that accepts dissimilarity data. In this case we will use Principal Coordinate Analysis (PCoA). Compare the results of these ordination techniques (figure 4) with the PCA result (figure 3). We can see a clearer separation of the two groups, and when we compare the result from Jaccard dissimilarities with Bray-Curtis, the first one appears to separate better the two groups in the first two axys. This indicates that the difference between them lies in the occurrence of less abundant bacteria, and that the more abundant ones, which contribute more to Bray-Curtis dissimilarities, are more similar between the two groups.

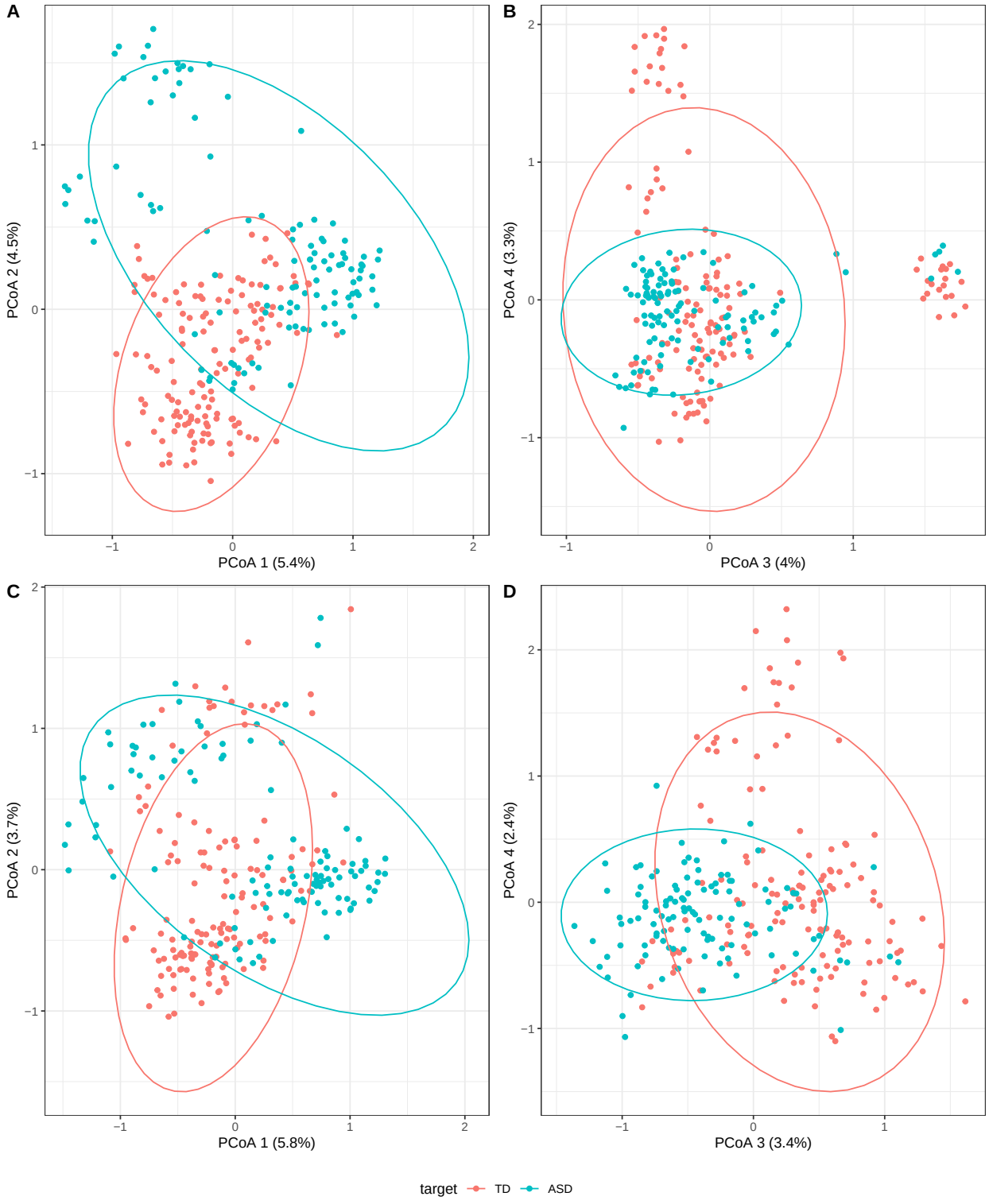

Figure 4: Principal Coordinates Analysis of microbiome data. A: First and second principal coordinate analysis axes for Jaccard dissimilarity. B: Third and fourth principal coordinate analysis axes for jaccard dissimilarity. C: First and second principal coordinate analysis axes for Bray-Curtis dissimilarity. D: Third and fourth principal coordinate analysis axes for Bray-Curtis dissimilarity

These results shows us that we can't use machine learning algorithms that relies in euclidian distances, because they won't be able to separate the two groups, but models that look for the co-occurrence of predictors have a good chance of correctly classifying individuals. This is one of the reasons that Random Forest works well with microbiome data, it looks for the best variables to split the data, and not the similarity between the samples.

Modeling approach

The data was split in two groups, a test set with 20% of the samples and a train set. The test set will only be used to evaluate the models, and the models will be build trained with the train set, with 5-fold cross-validation with 80% of the data. Since the data are sparse, any bacteria not present in the train set will be removed from the analysis. Since for this dataset the PCoA based on Jaccard dissimilarity appears to separate the two groups, all models will also be applied in the dataset transformed in a presence/ausence table. a benefit of this transformation is that it will facilitate the use of euclidean distances, as all distances will be either 0 or 1. The models that we will use are the following:

Random Forest (Rborist): This algorithm builds several decision trees, and presents the class selected by most of the trees as a result for each sample. In this way, the result of this algorithm is a majority vote ensemble of several models.

Extreme Gradient Boosted Decision Trees (xgbTree): Like RF, Extreme Gradient Boosted Decision Trees is an ensemble of models, but each new model is built to predict the errors of the previous model. In this way, each new decision tree in the model improves the predictions of the previous model.

Neural Networks with a Principal Component Step (pcaNNet): Neural networks involves a number of processors operating in parallel and arranged in tiers, with each tier operating in the output of the previous tier. This specific method performs a PCA previous to the model construction to reduce the dimensionality of the data, so it can be applied to datasets with more features than samples. But since it relies on PCA, it probably won't be efficient for microbiome data.

Logistic regression (glm): A simple model for predicting binary outcomes, it is an extension of linear regression that assures that the estimate is between 0 and 1, using the logistic transformation for this. As the data are not normal and the variables are not independent, this model is likely to perform poorly.

Naive Bayes (naive_bayes): A model similar to logistic regression, it uses Bayes' theorem to find the probability of a class given the probability of a feature.

k-nearest neighbors (knn): This model calculates the distance between the samples to find the neighborhood that a sample belongs. Since it relies on distance between the features, is another model that is likely to perform poorly.

Results

The performance metrics of all the models are presented in table 1 and figure 5. As expected, the tree-based models were the ones with the highest accuracy, with Extreme Gradient Boosted Decision Trees surpassing Random Forest. All models that expect normal distribution of data and independence of variables had low accuracy, with Neural Networks with a Principal Component Step presenting the best result among them. In general, the specificity was greater than the sensitivity, which indicates that the models had greater ease in classifying TD children as TD, but some ASD children were classified as TD.

The transformation of the dataset to presence/ausence did not interfere with tree-based models, but greatly improved the two models based in distance, Neural Networks with a Principal Component Step and k-nearest neighbors. k-nearest neighbors even reached the same accuracy score as Extreme Gradient Boosted Decision Trees, but training in only one second, versus the ~50 seconds that xgbTree took to train.

Table 1: Performance metrics of the six different models trained in the microbiome dataset.

	Accuracy		Sensitivity		Specifcity		Time (seconds)	
	Raw	P/A	Raw	P/A	Raw	P/A	Raw	P/A
Rborist	0.885	0.885	0.739	0.739	1.000	1.000	86.251	92.452
xgbTree	0.904	0.904	0.826	0.783	0.966	1.000	47.076	51.163
pcaNNet	0.769	0.885	0.652	0.739	0.862	1.000	28.923	28.402
glm	0.500	0.462	0.435	0.435	0.552	0.483	7.087	7.385
naive_bayes	0.692	0.558	0.348	0.000	0.966	1.000	11.723	11.865
knn	0.673	0.904	0.565	0.783	0.759	1.000	1.026	1.033

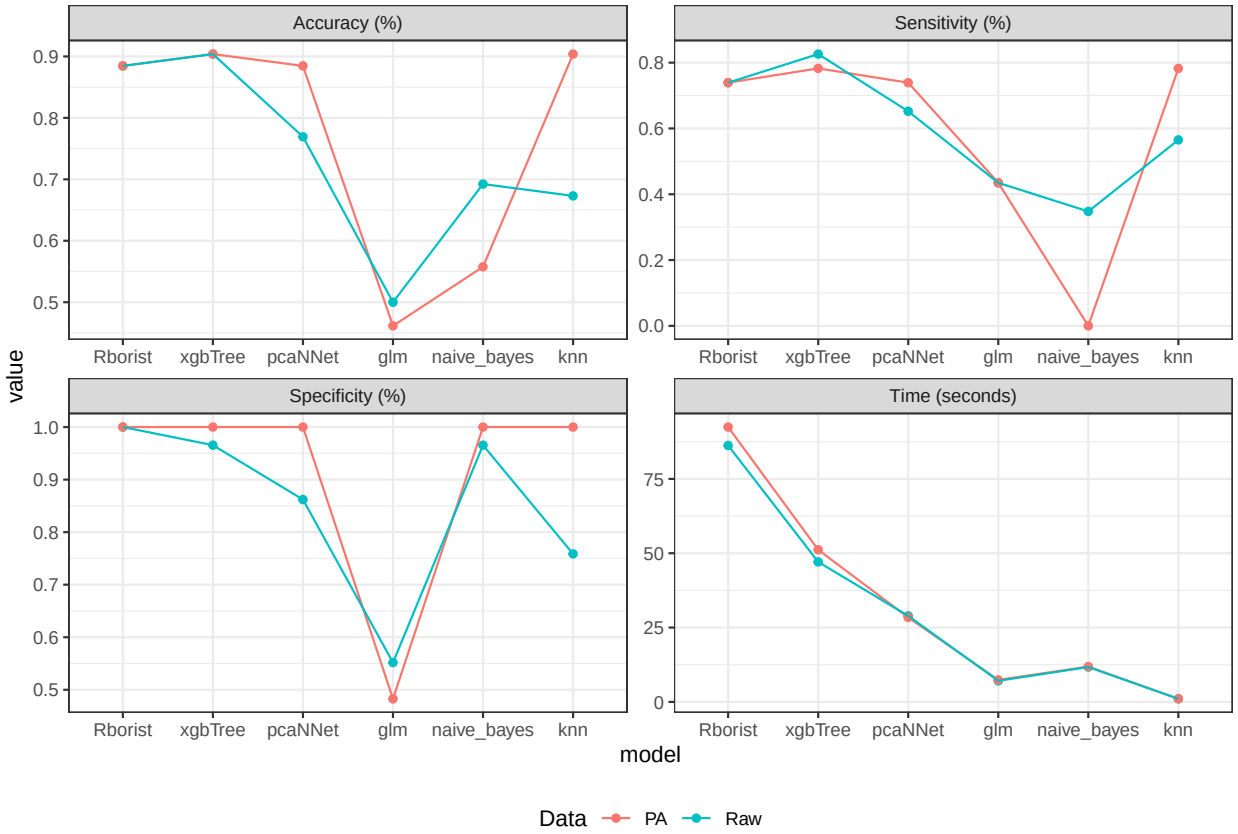

Figure 5: Performance metrics of the tested models

Conclusion

These results showcase the power of decision trees in problematic data, such those of microbiome, with very sparse data and more features than samples. It also shows that with knowledge about how to transform the data, even simple models like knn can be applied with great accuracy, with only 5 samples misclassified.

```
##           Reference
## Prediction TD ASD
##        TD  29   5
##        ASD  0  18
```

A care that must be taken with this data is that it does not establish cause and effect. It is not an altered microbiome that leads to the development of ASD, but more likely the other way around. In the paper the authors also analysis changes in the microbiota with age, and found that ASD patients do not alter the microbiota with age, unlike TD children, whose microbiota changes over time.