APPLYING SPECTRAL CLUSTERING TO GUT MICROBIOME: A CLASSIFICATION APPROACH

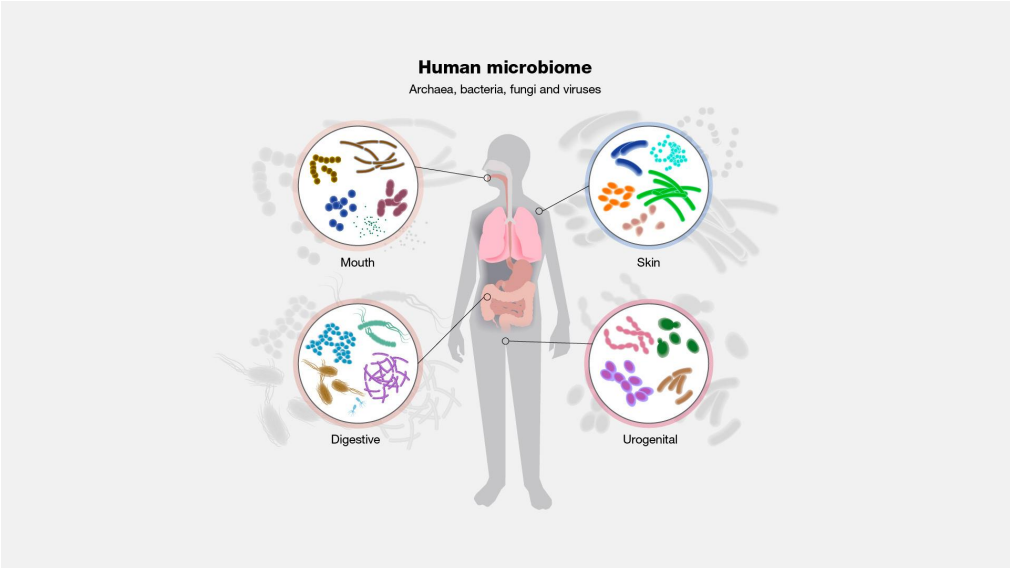
Name: *Francesco Canonaco*
ID: *781239*
Date: *April 2024*

OBJECTIVE OF THE PROJECT

In this study, I explored the **adaptation** of the **spectral clustering algorithm** to serve as a **classifier**, investigating its effectiveness in classifying **gut microbiome data** by employing different thresholds to compute the affinity matrix (graph).

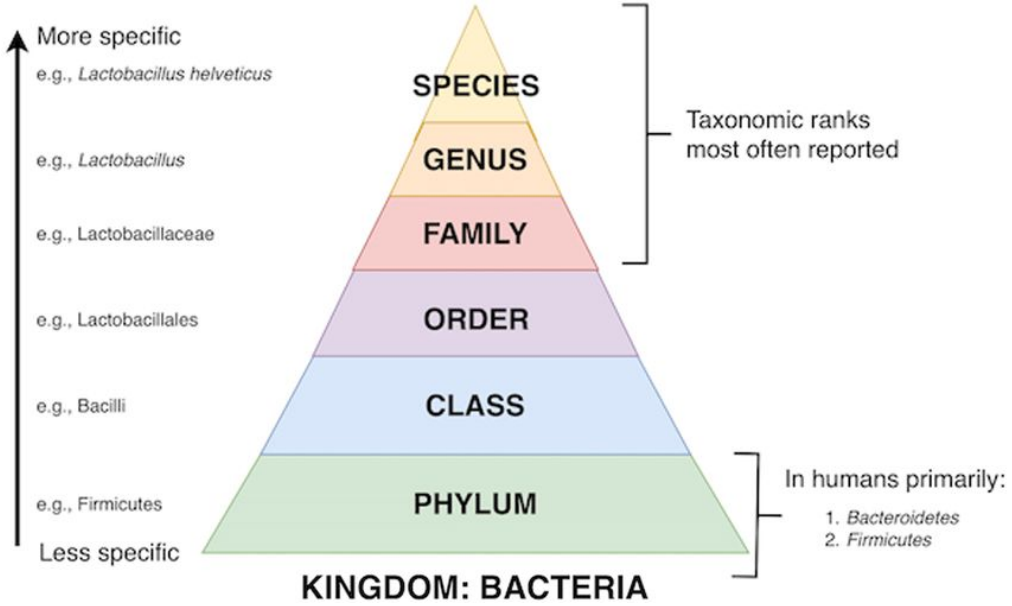
INTRODUCTION

The **human microbiome** is the collection of microorganisms, that reside on or within human tissues and biofluids along with the corresponding anatomical sites.



INTRODUCTION

Microbiome can be **analyzed** at many **levels**, the **higher** the level the **more specific** the classification of the microorganisms.



MICROBIOME DATA: CURATED METAGENOMIC DATA

The **curatedMetagenomicData** package offers standardized **human microbiome data** from various body sites, including *gene families*, *marker abundance*, and *pathway information*. It provides curated sample *metadata* and *metagenomic* data in (Tree)SummarizedExperiment objects.

[Repo]

waldronlab/
curatedMetagenomicData

Curated Metagenomic Data of the Human Microbiome

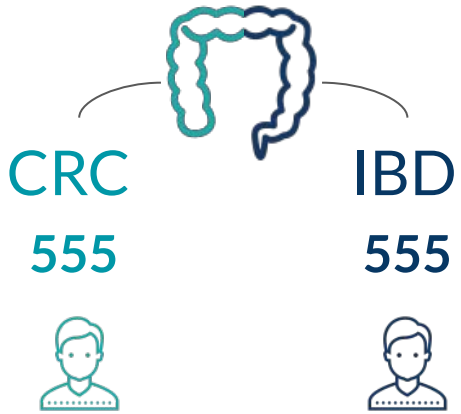
Authors: Lucas Schiffer, Levi Waldron

GUT MICROBIOME DATA: IBD (INFLAMMATORY BOWEL DISEASE) & CRC (COLORECTAL CANCER) PATIENTS

The analysis was conducted using the **family level**, using all the **555 subjects** with **CRC** and selecting **555 subjects** with **IBD.**

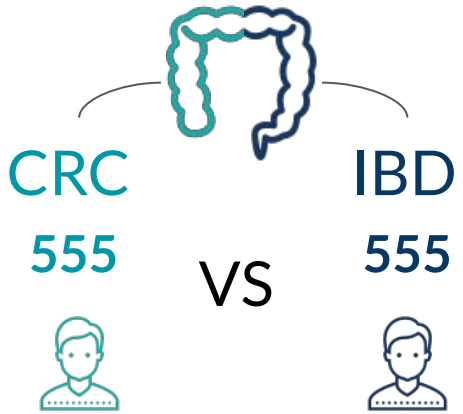
Each **sample** is characterized by **99 columns**, representing various **microbial families** and their **relative abundances**, and a column representing the class of the disease.

These **relative abundance** values indicate the **prevalence** of specific **microbial families** within each sample's gut microbiome.



INSTANCE OF THE PROBLEM: CLASSIFICATION OF CRC VS IBD PATIENTS

Spectral clustering was adjusted for a classification task and compared to a **Neural Network** for performance assessment.

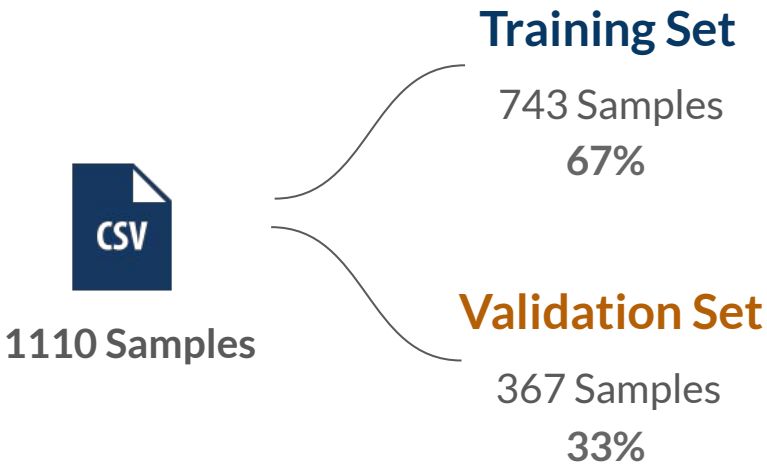


SPECTRAL CLUSTERING AS A CLASSIFIER

Given a dataset ***D*** with ***N*** examples having ***m*** columns the unsupervised learning was adapted to solve a classification task using the following procedure:

1. The **graph (affinity matrix)** was computed using the *Bray-Curtis distance* resulting in a **N**x**N** matrix representing a graph.
2. **Distance matrix** was transformed in a **similarity matrix**.
3. **Clusters** were ***learned*** by the Spectral Clustering procedure and each cluster was **assigned a class** based on the **majority class** within it.
4. Given a **new sample**, the **class** was assigned by *transforming it to align with the computed affinity matrix* and by assigning it to the **nearest cluster**.

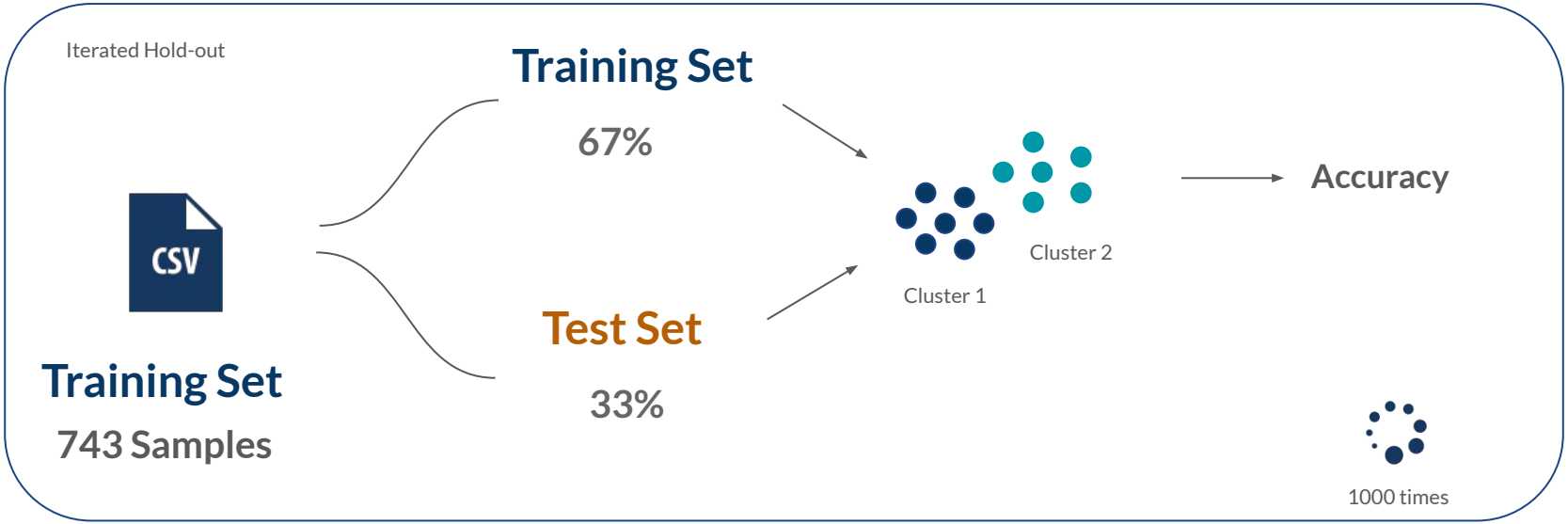
TRAINING AND VALIDATION SET SPLIT



The dataset was split in **training set** and **validation set**, maintaining the same proportion for the two classes (IBD, CRC).

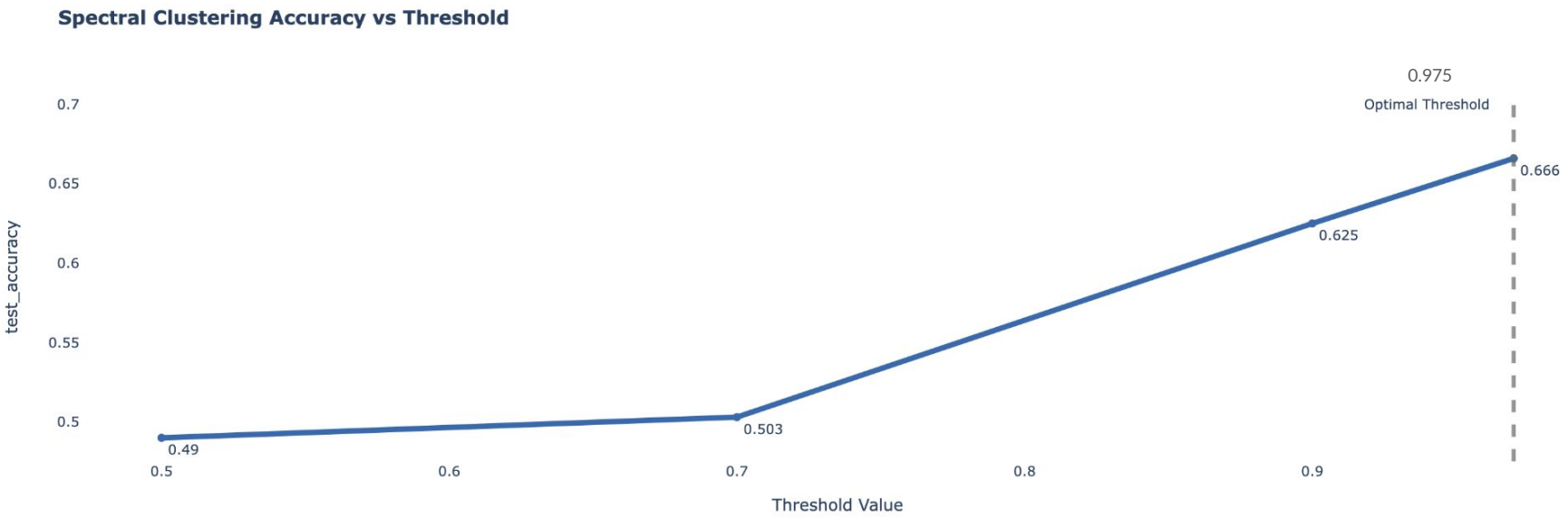
The **training set** was used for the **model selection procedure** whilst the **validation set** was used for the **final test**.

OPTIMIZATION OF THE SPECTRAL CLUSTERING BY THRESHOLDING THE GRAPH (AFFINITY MATRIX)



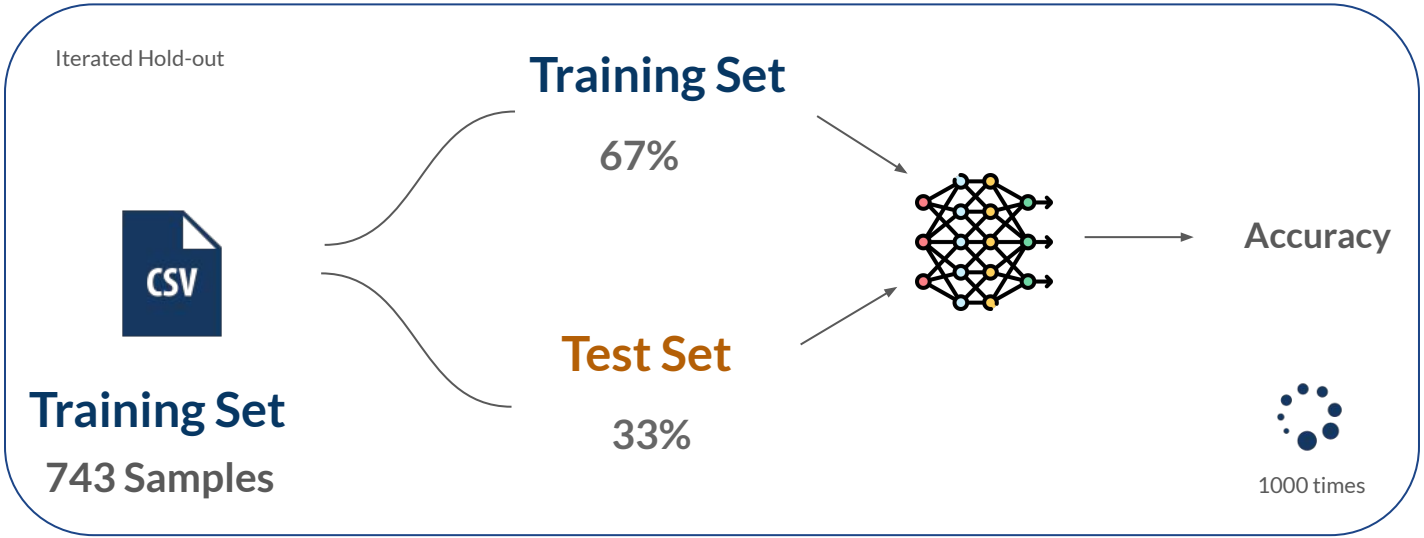
**Before feeding the graph (affinity matrix) to the clustering algorithm a threshold was applied (before transforming it in a similarity matrix) in order to optimize the accuracy of the model.*

OPTIMIZATION OF THE SPECTRAL CLUSTERING BY THRESHOLDING THE GRAPH (AFFINITY MATRIX)



*The optimal threshold was 0.975, indicating that edges with a similarity below 0.025 were assigned a weight of zero.

OPTIMIZATION OF A MULTILAYER PERCEPTRON

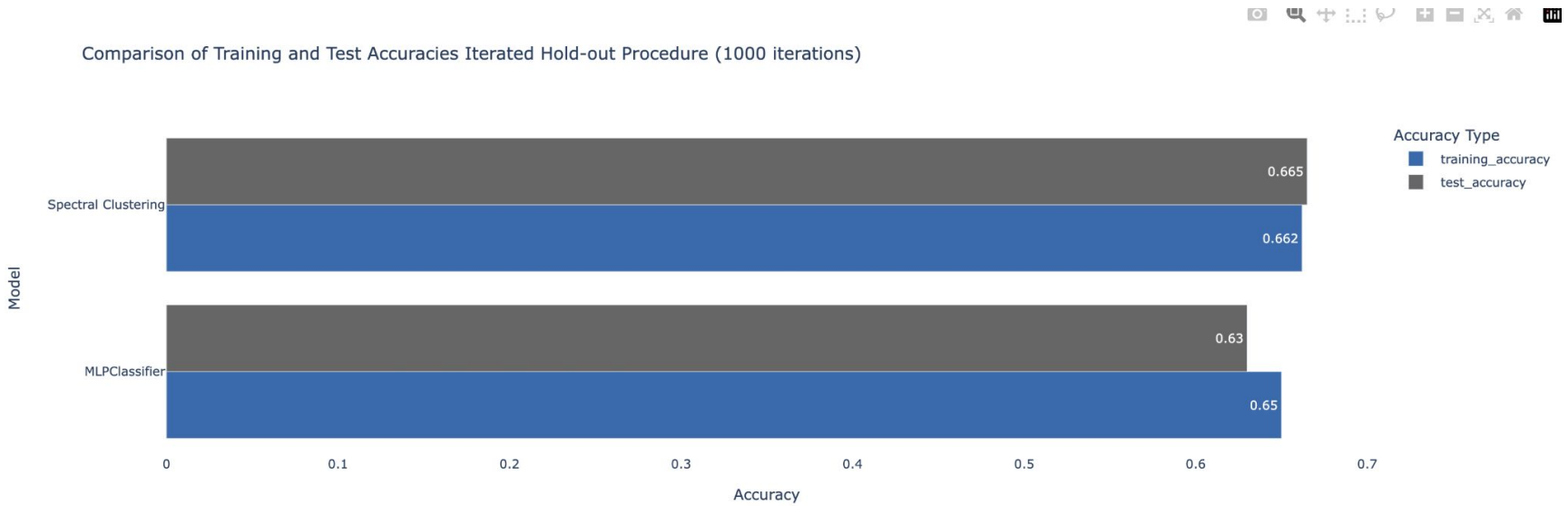


The **MLPClassifier** underwent **optimization** with respect to various parameters, including ***hidden layers***, ***alpha***, and ***max iterations***. Each parameter was optimized individually through an iterated hold-out process consisting of 1000 iterations. The resulting optimal parameters are as follows:

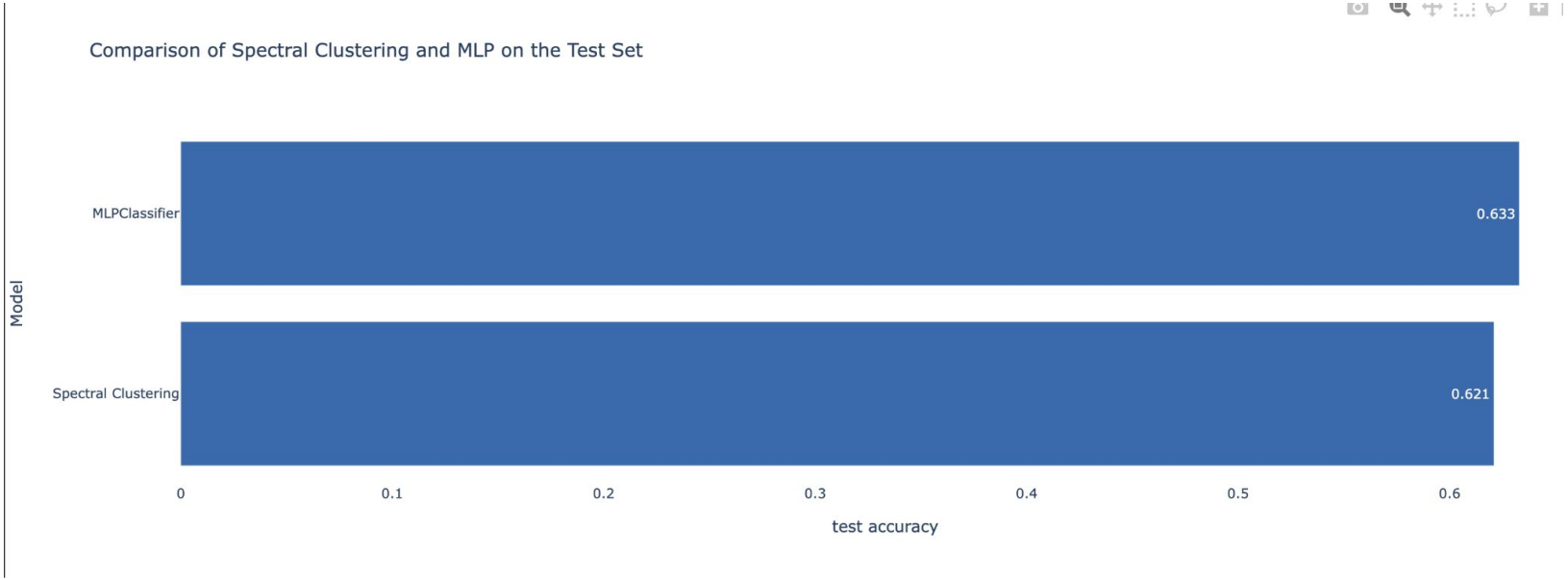
- **Hidden Layers**: (99,50)
- **Alpha**: 1e-5
- **Max Iterations**: 3

The activation function used was **ReLU**, and the solver employed was **Adam**.

COMPARING SPECTRAL CLUSTERING AGAINST MULTILAYER PERCEPTRON: ITERATED HOLD-OUT



COMPARING SPECTRAL CLUSTERING AGAINST MULTILAYER PERCEPTRON: VALIDATION SET



LET'S WRAP IT UP

- Gut microbiome **data were collected** and processed using the *curated metagenomics repository*.
- **Spectral clustering** was **adapted** to be used for a classification task.
- The new approach was **optimized** using a *thresholding approach*
- A **MLP** (multi layer perceptron) was used to **assess the performance** of the new classifier.

CONCLUSION

- Both models showed **comparable performance**, surpassing *random guessing*.
- However, **performance was not remarkable**, likely due to *numerous variables* and *limited sample size*.
- Despite challenges, **Spectral Clustering Classifier** performed **quite good**, given its unsupervised nature and **data limitations**.