WG3 progress reporting & future plan

8th July 2021

WG3 Objectives and major deliverables

Objectives:

To optimise and standardize the use of state-of-the-art ML techniques, resulting in **best practice SOPs** specific to various microbiome data types, human body ecosystems and research questions. The WG3 will also investigate opportunities for **automating** the established SOPs into pipelines for translational use by clinicians and non-experts.

Major Deliverables:

D3.1: [oct - dec 2021] A decision tree of ML/Stats methods along with optimised parameters suitable for various data types, ecosystems and research questions (disseminated through Web-portal and GitHub).

D3.2: [april - jun 2022] A publication and white-paper describing the SOPs emanating from D3.1.

D3.3: [july - sept 2022] A report outlining areas suitable for automation

Summary of WG3 progress (2019-2021)

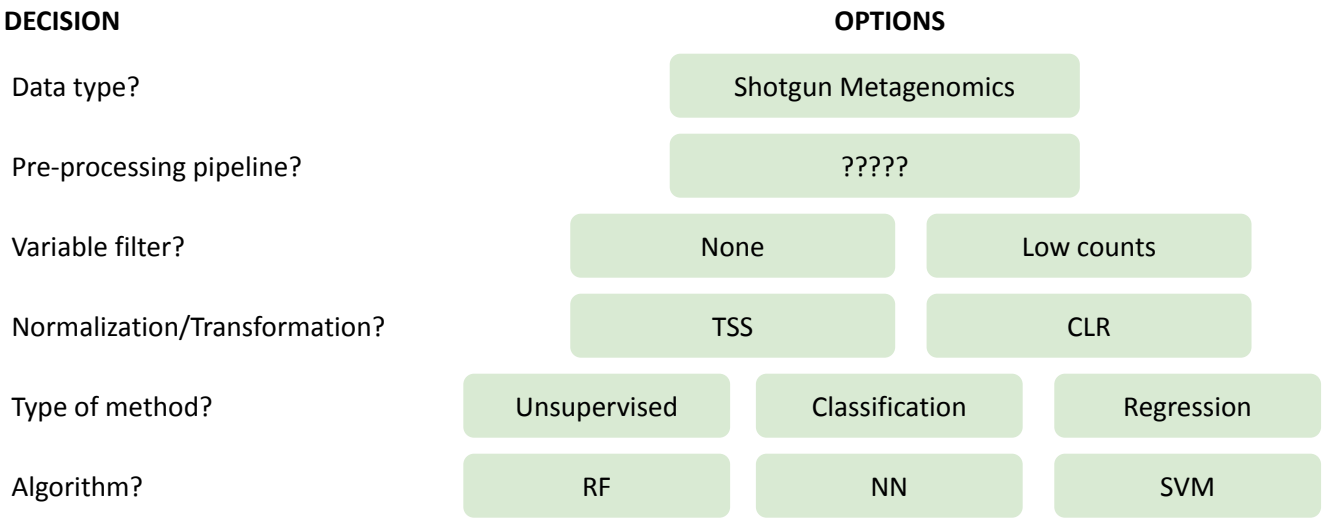
Several threads of research, from different groups and collaborations. For the moment, mainly analysis on publicly available datasets (mainly 16s). (**...this is not an exhaustive list**)

- Explain the observed diversity in human microbiome (*University of Turku, Finland*)
- Predicting the onset of Type2 diabetes with AutoML using microbiome data *(Dept. of Computer Science, University of Bari Aldo Moro, Bari, Italy - Institute of Genomics, University of Tartu, Tartu, Estonia*)
- Probabilistic distribution of taxonomic units (*Ss. Cyril and Methodius University in Skopje, North Macedonia*)
- Clustering and classification of human microbiome data (*University of Novi Sad, Serbia- University of Ljubljana, Slovenia*)
- Comparing different normalization strategies and ML classification methods on 6 different datasets for 5 diseases (*Universitat Politècnica de València, Spain*)
- Analysis of human microbiome data with JADBio (*Department of Computer Science, University of Crete, Greece, FORTH*)
- Statistical and ML analysis of microbiome data using the logratio methodology of compositional data (*Palacký University, Czech Republic)*

Summary of WG3 progress (2019-2021)

From the studies of the members we started to define a decision tree for SOP, showing in different data/normalization/pre-processing/algorithms what is the best approach according to their experience

Example:



MICROBIOME

Summary of WG3 progress (2019-2021)

Two main approaches to choose the Operating Procedures to be adopted in the studies:

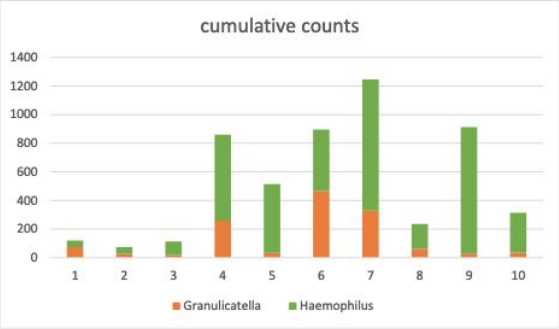
- Classical experimental & explanatory approach
- Automatic, based on AutoML

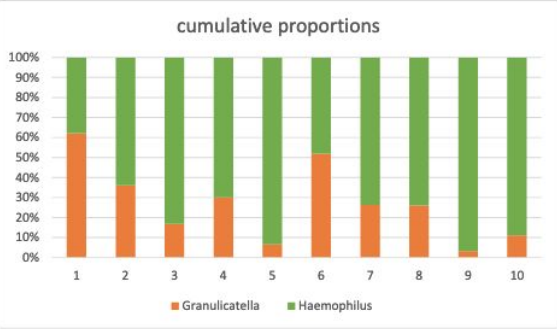
A joint work with WG1 has also been conducted to identify and analyze relevant papers. The standards steps from existing literature will also be included in the tree when relevant.

Short talks (07/07/21)

- Karel Hron: ***Why are microbiome data compositional?***
- Andrea Mihajlovic (Tatjana Loncar Turukalo): ***Inflammatory bowel disease prediction based on metagenomics data***
- Magali Berland: ***Extensive benchmark of machine learning methods for microbiome data***
- Michelangelo Ceci: ***Predicting the onset of Type2 diabetes through the analysis of microbiome data***

Karel Hron: Why are microbiome data compositional?





Do both representations carry the same information?

- **NOT** in absolute scale, **YES** in relative scale
- counts can not be estimated from proportions
- but proportions can be estimated from counts

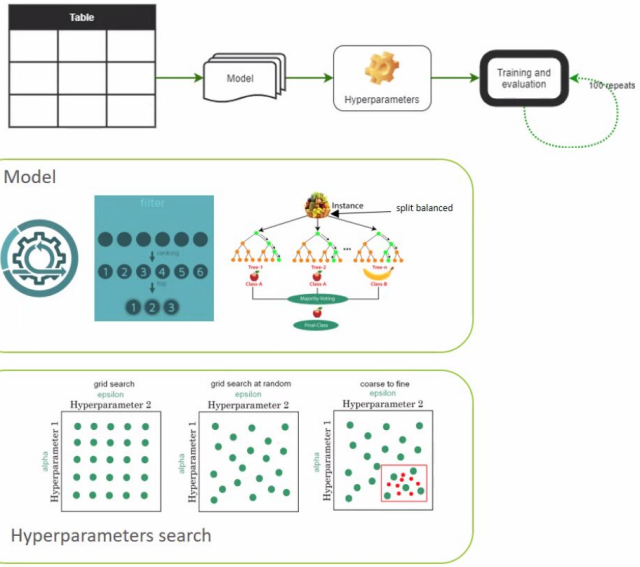
Karel Hron: Why are microbiome data compositional?

microbiome data are compositional!!!

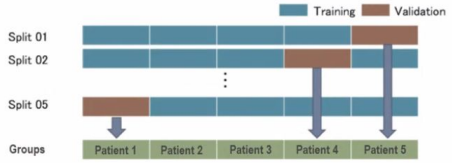
- interest is (or should be) in the relative information carried by proportions
- the simplex corresponds to the set of possible observations
- an interpretable measure of difference and scale of variables is available
- a suitable, well known algebraic-geometric structure allows building coherent models
- for CoDa, it is better to think in terms of ratios

Andrea Mihajlovic: Inflammatory bowel disease prediction based on metagenomics data

Methodology



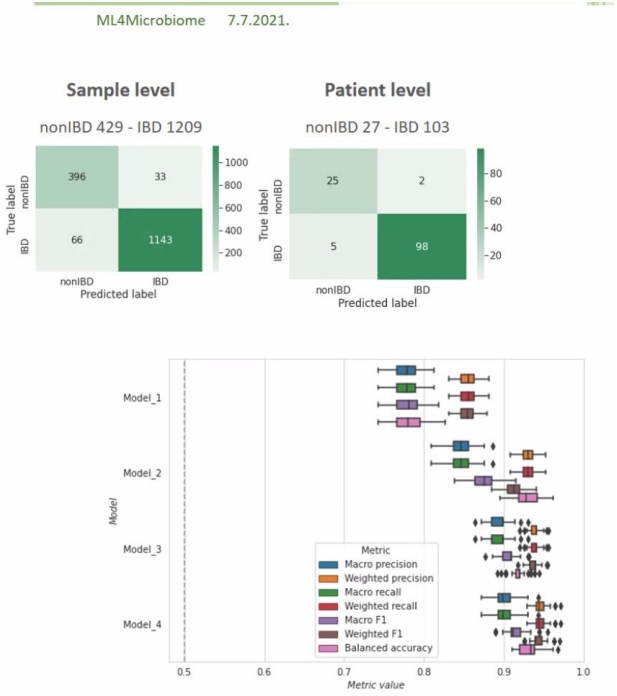
- Machine learning based binary classification
- Two levels: sample and participant
- 10 fold group cross validation
- Evalution – 100 repeats



Andrea Mihajlovic: Inflammatory bowel disease prediction based on metagenomics data

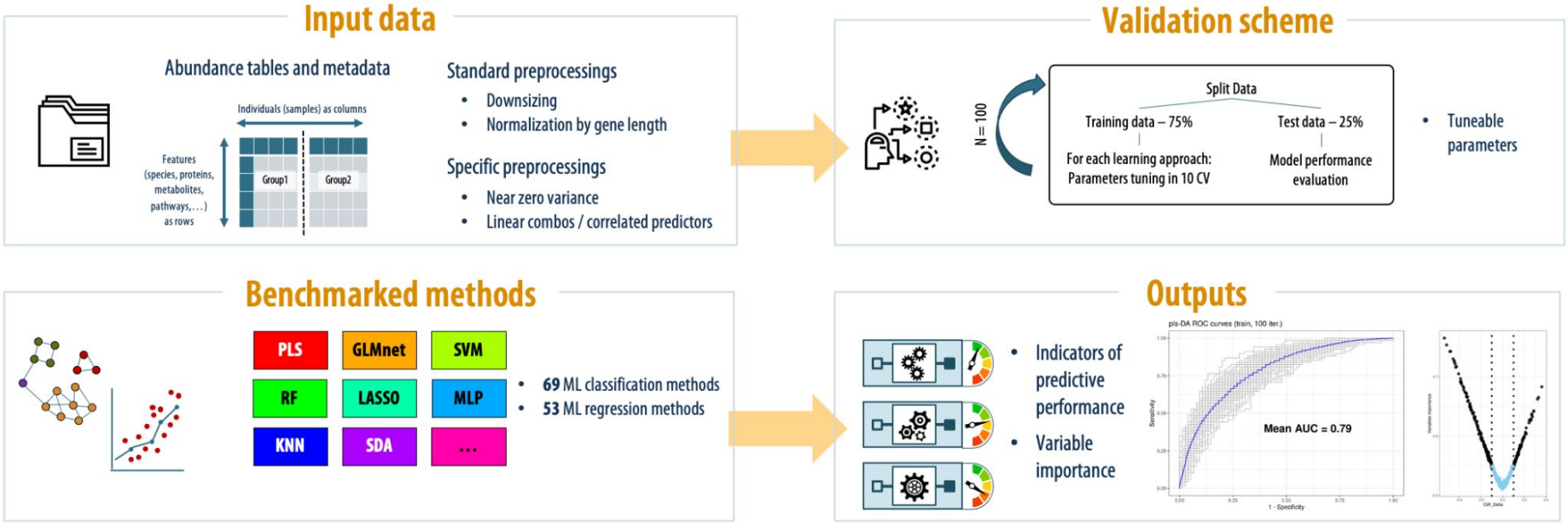
Results

Estimator	Parameter	Model 1	Model 2	Model 3	Model 4
$f(x)$	feature selection	all	bacteria only, strains removed	bacteria only, strains removed	all
SKB	k	300	150	500	no reduction
BRF	class weigth	None	None	None	None
	criterion	gini	gini	gini	entropy
	max depth	3	25	25	15
	n estimators	100	150	150	150
	min sample leaf	2	1	1	1
	bootstrap	True	True	True	True
Decision	threshold	0.4	0.24	0.24	0.31



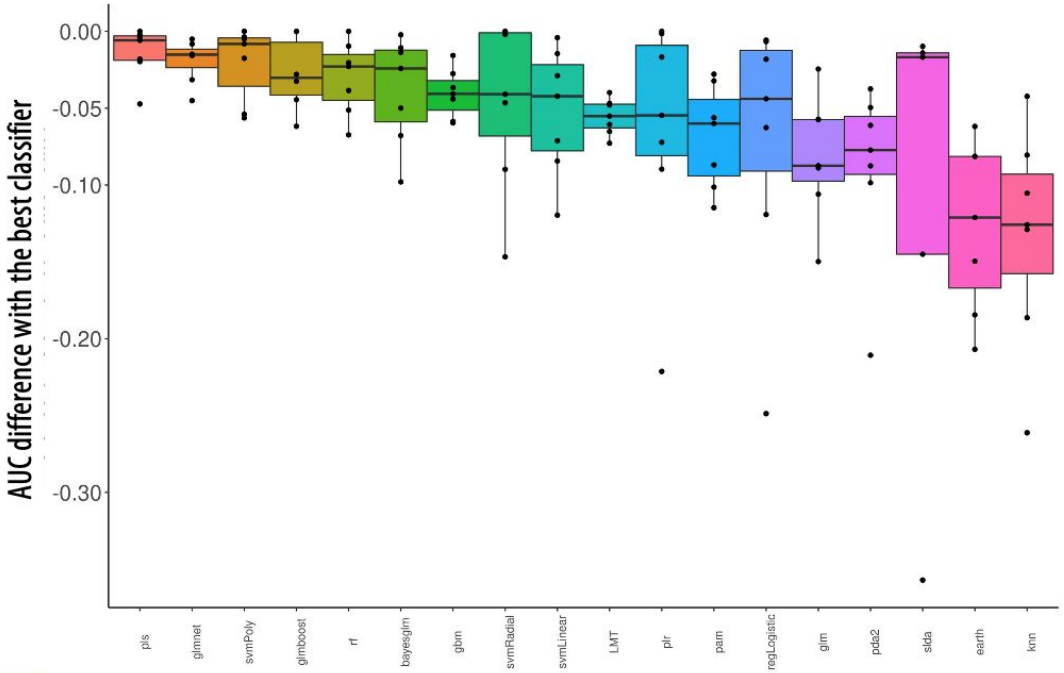
MICROBIOME

Magali Berland: Extensive benchmark of machine learning methods for microbiome data



MICROBIOME

Magali Berland: Extensive benchmark of machine learning methods for microbiome data



Microbiome database

MICROBIOME

Michelangelo Ceci: Predicting the onset of Type2 diabetes through the analysis of microbiome data

Data Source: data from METSIM cohort[1]

Pre-processing steps on microbiome data:

- Total Sum Scaling (TSS)
- Centered Log-ratio Transformation (CLR)
- Isometric Log Ratio (ILR)

Settings:

- **Supervised** (887 patients, excluding those with missing values in the target attributes)

Task:

- Learn multiple **classification** models (one for each target attribute)

Considered approaches:

- AutoWEKA[2] (AutoML approach)

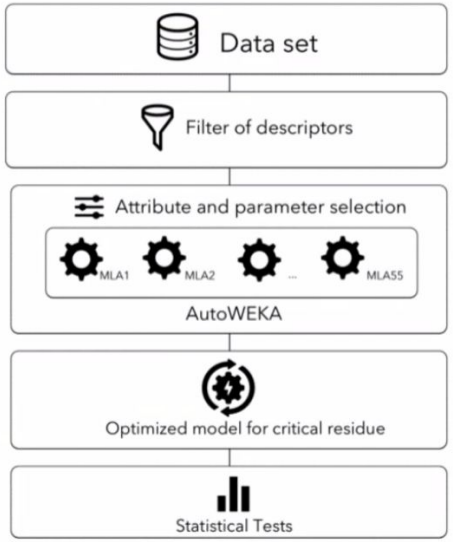
1. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000743.v1.p1
2. https://www.cs.ubc.ca/labs/beta/Projects/autoweka/

Target Variables:

- Insulin AUC measured after the OGTT test (**auc_ins**)
- Disposition index (**disposition**)
- Post-load glucose (**gl120**)
- Fasting insulin (**ins0**)
- Post-load insulin (**ins120**)
- Matsuda insulin sensitivity index (**matsuda**)

Michelangelo Ceci: Predicting the onset of Type2 diabetes through the analysis of microbiome data

Why AutoML?



Feature selection

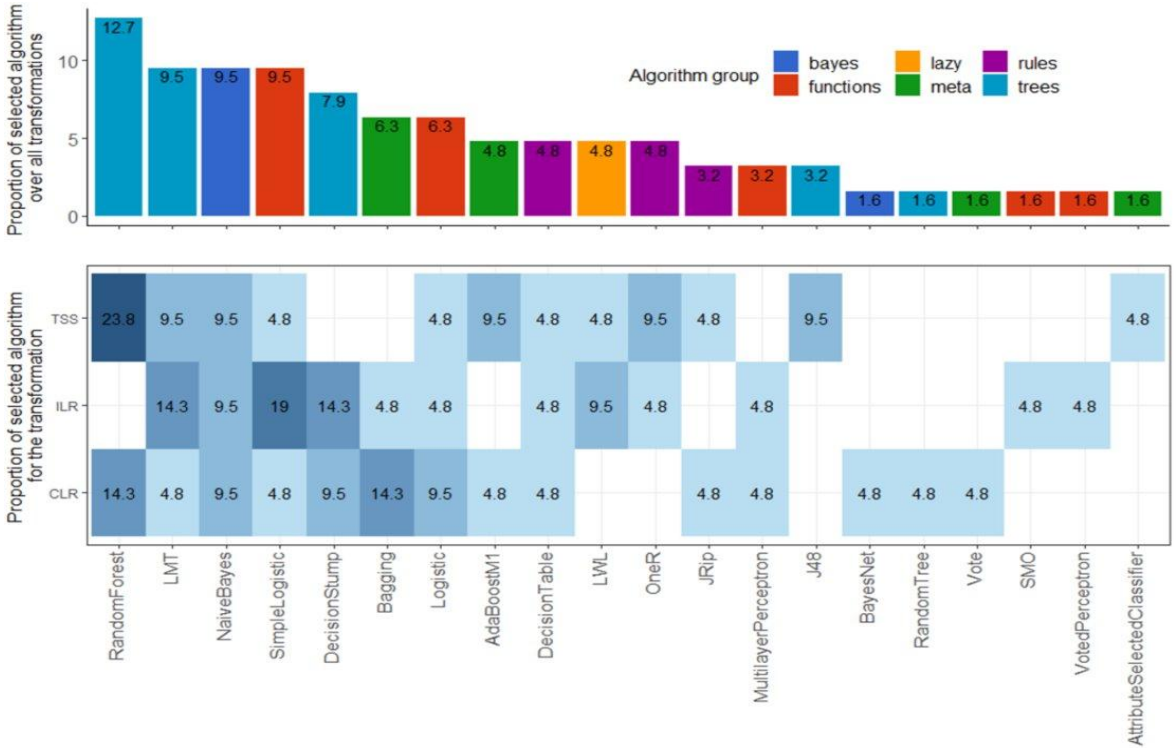
Attribute search: BestFirst; GreedyStepwise

Attribute Evaluation: CfsSubsetEval

Algorithms	**Paramenters**
• J48	
• DecisionTable	
• GaussianProcessed	
• M5P	
• LogisticModelTrees	
• PART	
• SMO	
• BayesNet	
• NaiveBayes	
• Jrip	
• SimpleLogistic	
• LinearRegression	
• SGD	
• …	

MICROBIOME

Michelangelo Ceci: Predicting the onset of Type2 diabetes through the analysis of microbiome data



MICROBIOME

Discussion and future actions/collaborations (2021-2022)

- WG3 actions
  - Decision tree (oct 21) ⇨ Monthly meetings to complete the tree from our experience
    - ***Next meeting: September 10th, 10:00-12:00 CET***
  - SOPs & publication (april 22) ⇨ Need for standard dataset from WG2 to bring reproducible standards out
  - Automation report (july 22)
- WG3 group (12): Magali Berland, Michelangelo Ceci, Sonia Tarazona
  - Andrea Mihajlovic, Tatjana Loncar Turukalo, Jill O'Sullivan, Julia Eckenberger, Karel Hron, Yorgos Papoutsoglou, Enrique Carillo (WG1), Kanita Karadjuzovic Hadziabdic (WG1), Marta Belchior Lopes