Multiomics integration in PANS

ACCIARO MORISCO

Multiomics integration in PANS

Integrate **metabolomics** and **proteomics** data from PANS. Perform classification of patients based on the two types of data, and extract biomarkers that maximise classification performance.

OUR PROJECT IN A NUTSHELL

01

METHODOLOGY

A-Posteriori
VS
A-Priori
data integration

02

FEATURE SELECTION

Isomap VS **PCA** VS **RFE**
with three different scalers

03

MODEL SELECTION

Grid Search among **seven** different classifiers

04

RESULTS

For the a-posteriori methodology, results were obtained through **Ensemble** methods

A-POSTERIORI VS A-PRIORI

A-POSTERIORI

We consider the two datasets initially independent, performing two separate Feature/Model Selection processes, and then merge the final models through Ensemble techniques.

A-PRIORI

On the contrary, in this case we merge the two datasets before proceeding with the Feature and Model Selection processes.

A-POSTERIORI

Proteomics

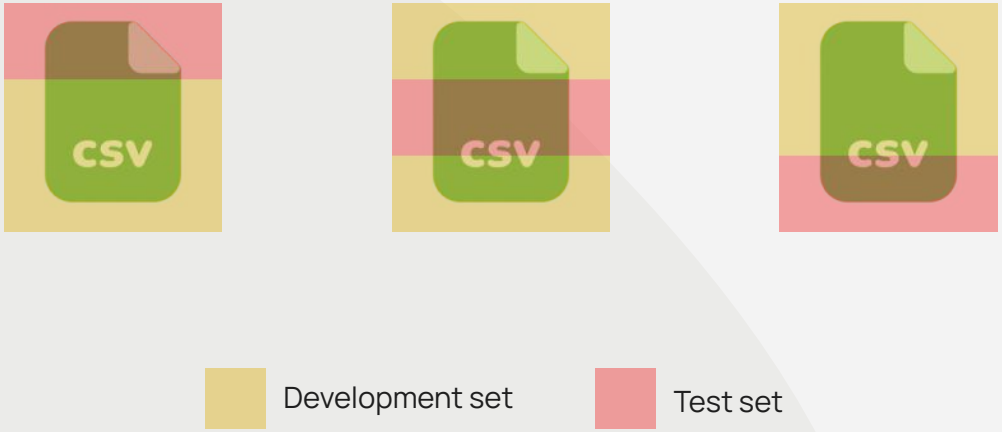
Metabolomics

A-PRIORI

A-Priori

HOW WE AVOID LUCK

To avoid lucky cases, we split each dataset into three separate parts, alternating the test set. The effect obtained therefore at the Model Selection phase, as we shall see will be that of a **Nested Cross Validation**.



FEATURE SELECTION

The metabolomics dataset contains 931 features, whereas, the proteomics dataset contains 1317 features. It was first necessary to find a way to reduce the number of features.

Therefore, we compared three different methods to reduce features:

ISOMAP

PCA

RFE

	Isomap	PCA	RFE
	wait for it	wait for it	wait for it
	wait for it	wait for it	wait for it
	wait for it	wait for it	wait for it

SCALERS

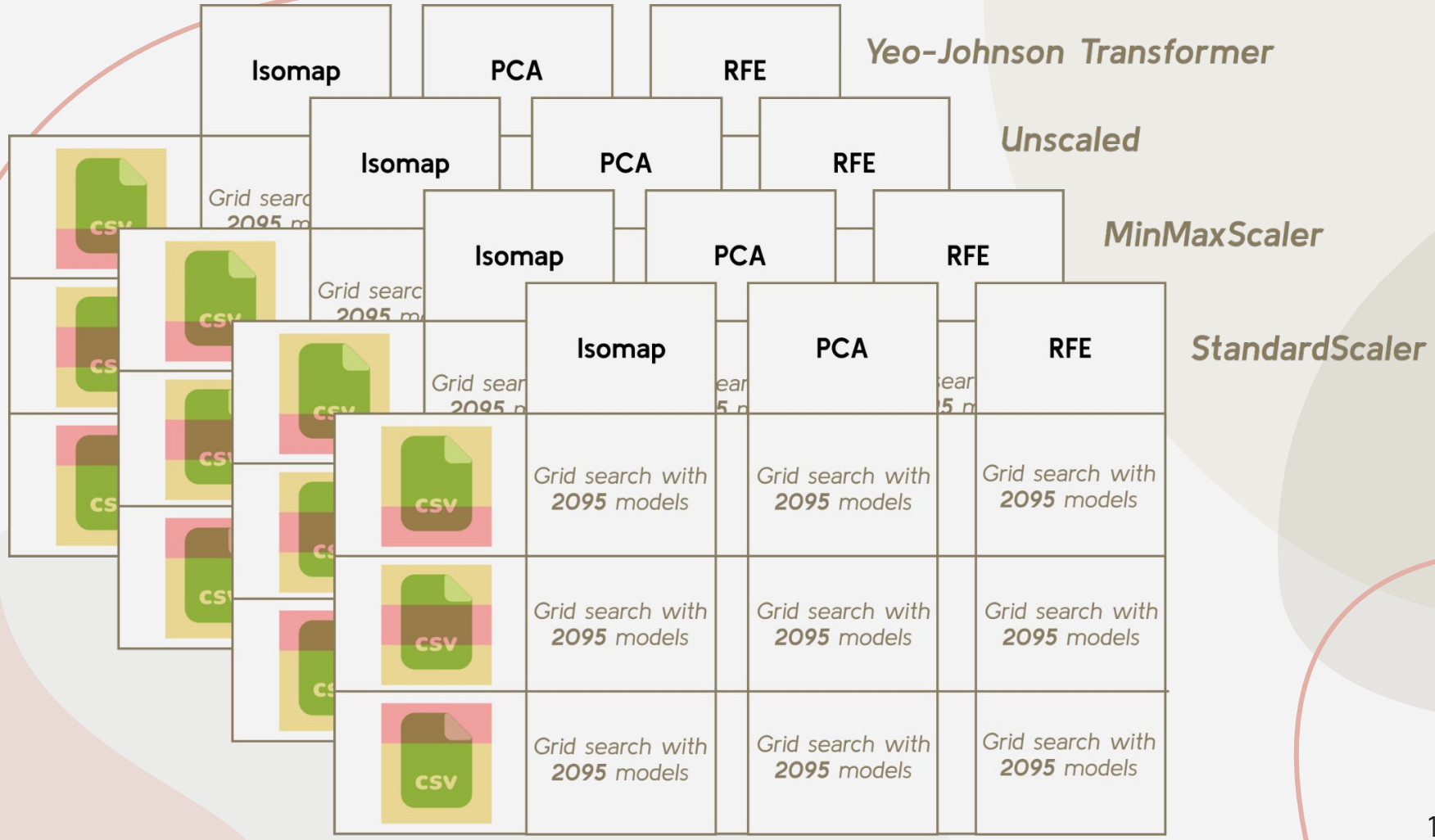
For each Feature Selection method, we apply the following scalers in order to normalize data

1) MinMaxScaler
2) StandardScaler
3) Yeo-Johnson Transformer
4) Unscaled

MODEL SELECTION

For each possible combination of Feature Selection and Scaler method, we performed a Grid Search with the following models (and all their possible hyperparameters)

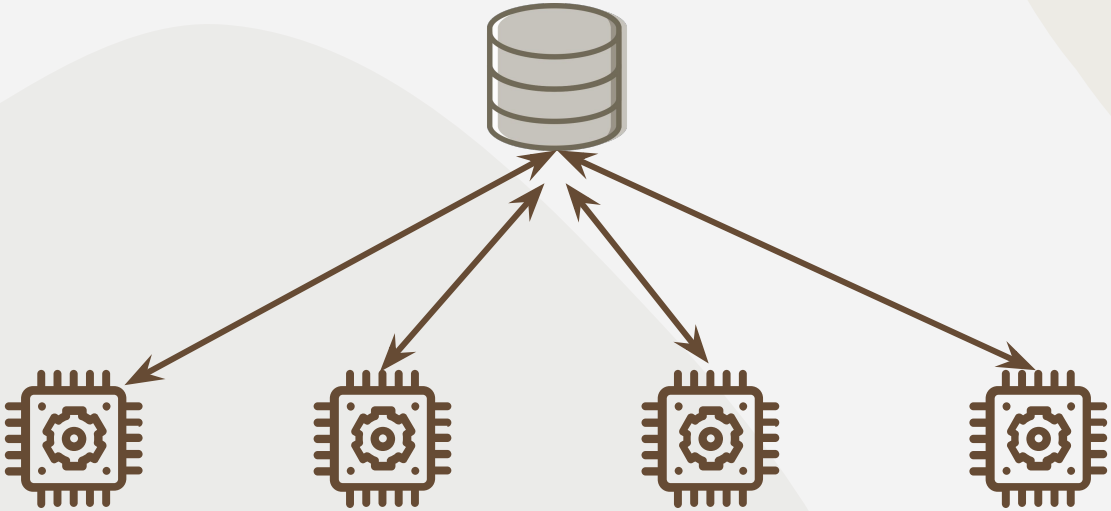
1) Naive Bayes
2) Decision Tree
3) KNN
4) Logistic Regression
5) MLP
6) Random Forest
7) SVM


Isomap
PCA
RFE
Yeo-Johnson Transformer
Unscaled
MinMaxScaler
StandardScaler
Grid search with 2095 models
csv

226,260

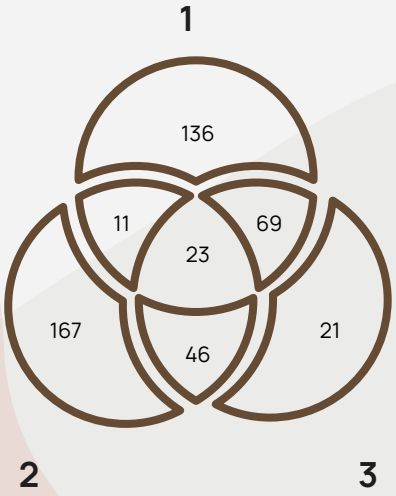
The total number of models we tested

(DISTRIBUTED) MODEL SELECTION

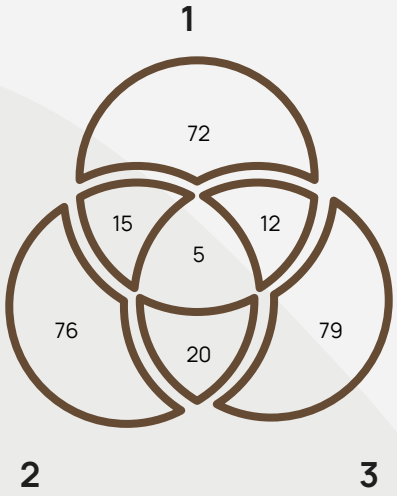


FEATURE SELECTION **RESULTS**

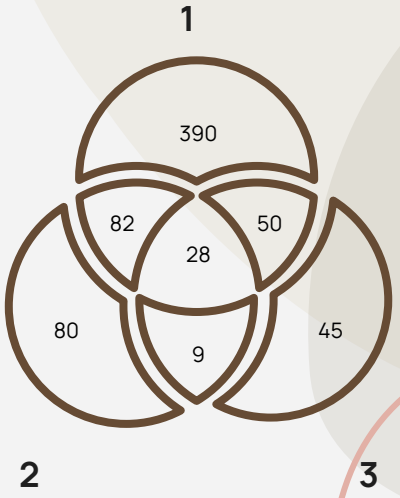
metabolomics



proteomics

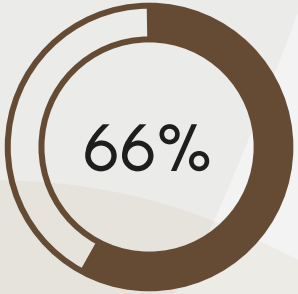


a-priori

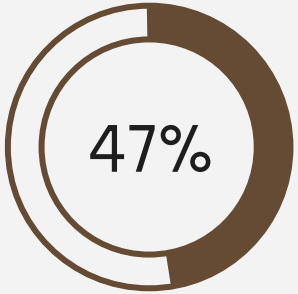


FINAL **RESULTS**

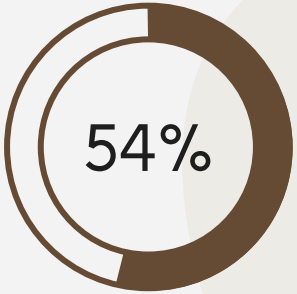
Metabolomics



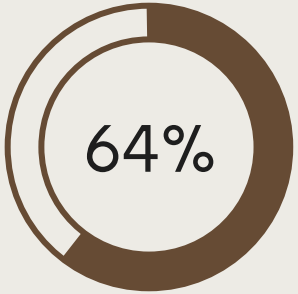
Proteomics



A-Posteriori



A-Priori



Results obtained based on the Test Set

LESSONS **LEARNT**

TIME

Grid searches take time

BIOLOGY CONCEPTS

For the study of omics data

FEATURE SELECTION

Sometimes, most of the data are useless

A REAL-LIFE ML PROJECT

Using Sklearn is simple, knowing which with parameters no

Thank you for your time