Improving the performance of Machine Learning models with features selection techniques in the context of signal/background discrimination for the AGILE telescope

Leonardo Baroncelli (PhD student in Data Science and Computation),

Andrea Bulgarelli, Nicolò Parmiggiani

(INAF/OAS Bologna)

leonardo.baroncelli@inaf.it

Deep Learning @ Inaf, F2F meeting, 16-19 September 2019

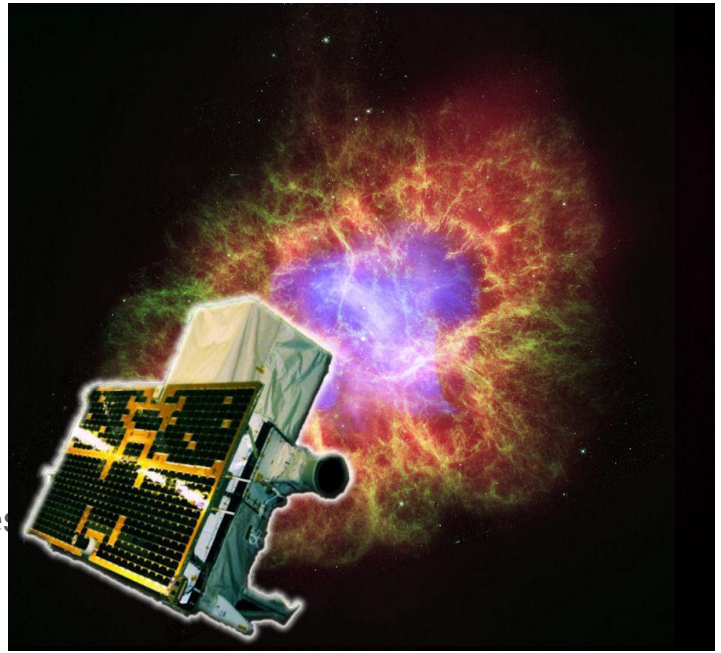
The task: signal/background discrimination

INAF OAS BOLOGNA

AGILE (Astrorivelatore Gamma ad Immagini LEggero) is a scientific mission of the Italian Space Agency (ASI), launched on April 23, 2007 for Gamma Ray Astrophysics.

The AGILE-GRID data acquisition system is composed of three main detectors:

- a Tungsten-Silicon Tracker designed to detect and image photons in the 30 MeV-50 GeV energy band,
- a Mini-Calorimeter that detects gamma-rays and charged particles energy deposits between 300 keV and 100 MeV,
- an anti-coincidence (AC) system that surrounds the Silicon Tracker and the Mini-Calorimeter.



On-ground signal-to-background discrimination: **FM3.199** filter.

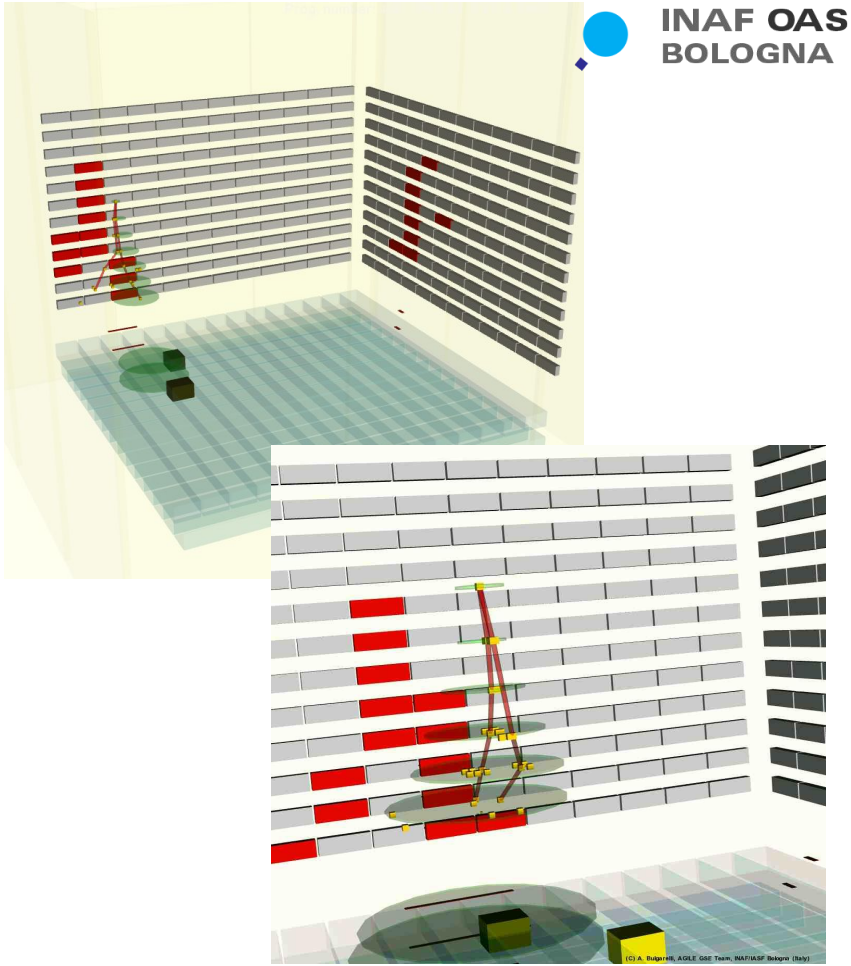
The FM3.119 uses AdaBoost and a **subset** of discriminant variables (57 features) selected by a domain expert (manual feature selection).

⇒ Will automatic features selection techniques bring to better results?

⇒ Which is the best features selection method for this task? How many features?

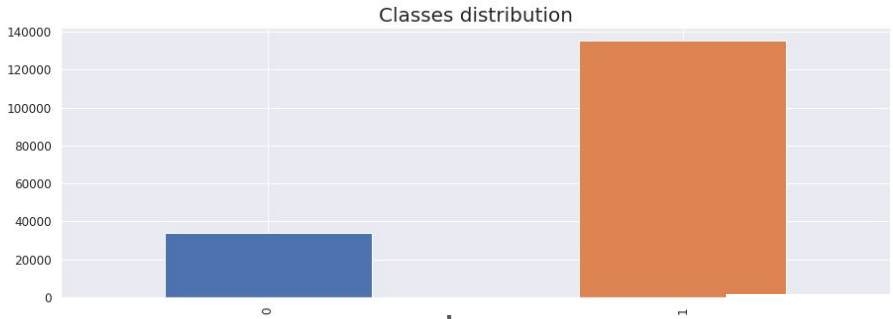
The dataset

- It has been generated from Monte Carlo simulations of the AGILE on-fly data acquisition systems.
- It describes the particle interactions with the AGILE instruments (silicon tracker, calorimeter, anti-coincidence system).
- It is a supervised dataset: the particles are divided into two classes: gamma photons and background particles.
- It is composed by 169.813 rows (particles) and 260 columns (interaction features) in csv format.



Dataset preprocessing

Random (uniform) subsampling of the gamma examples to obtain a balanced dataset.



From 169.185 to 67.952 examples.

Dropped ~60% of the dataset.

Initial shape = (169.813, 260)

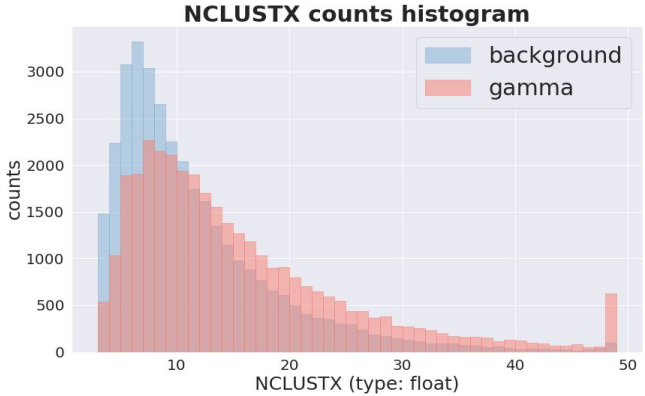
Cleaning + subsampling

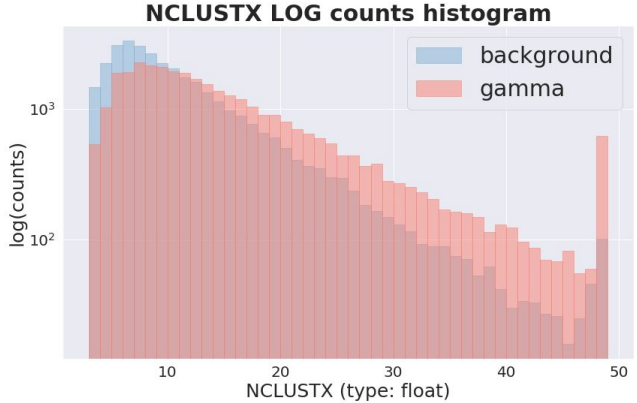
Final shape = (67.952, 241)

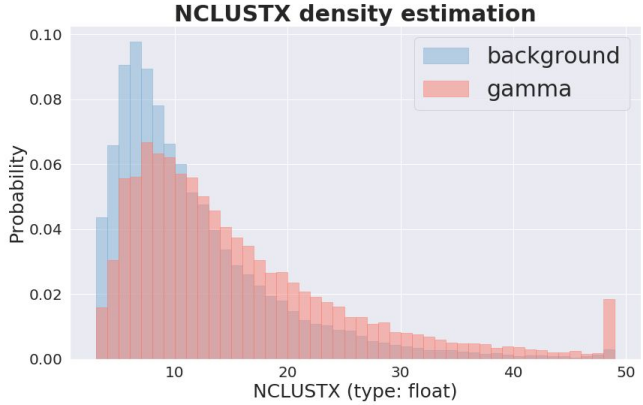
Dataset exploration

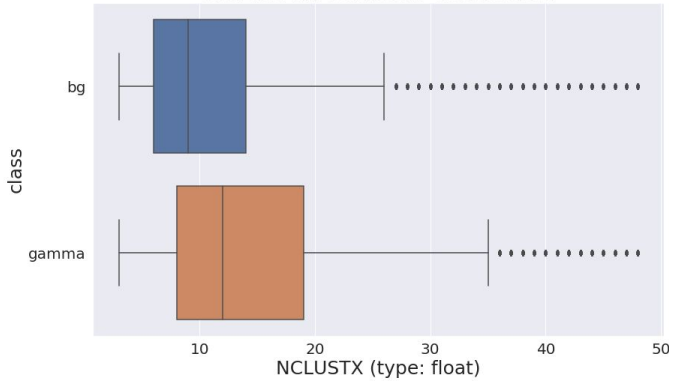
For each feature:

- Counts histogram
- Density histogram
- Counts histogram in log scale
- Box plot



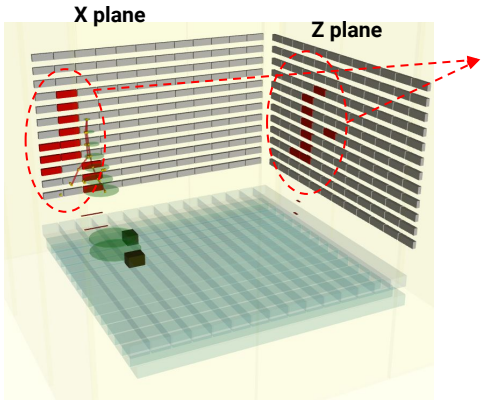




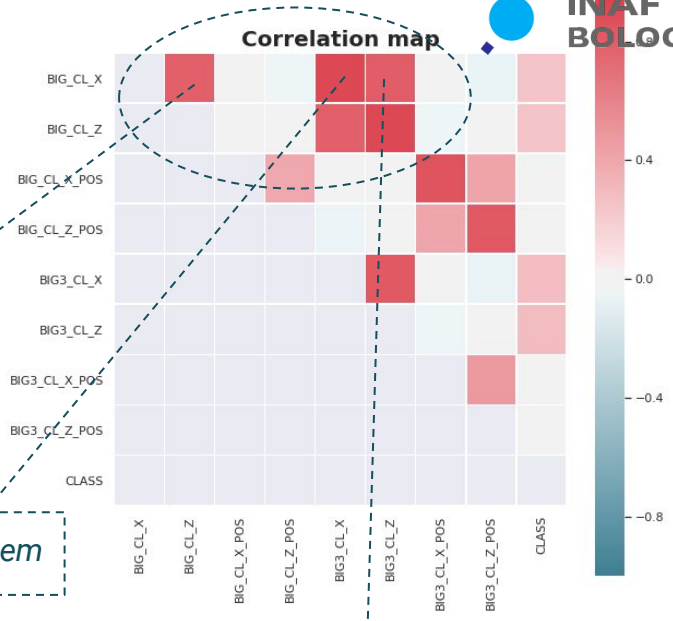


Dataset exploration

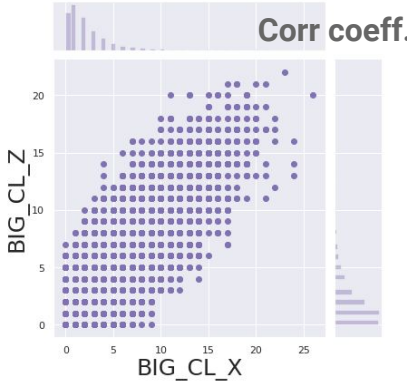
For each group of related features: correlation matrix

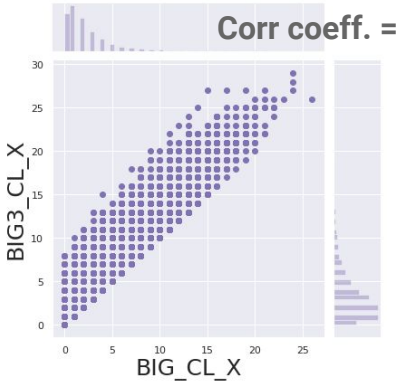


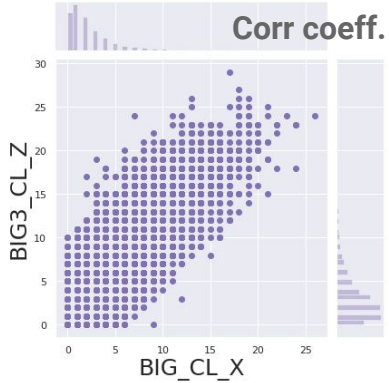
Example: BIG_CL = big cluster



Multicollinearity problem

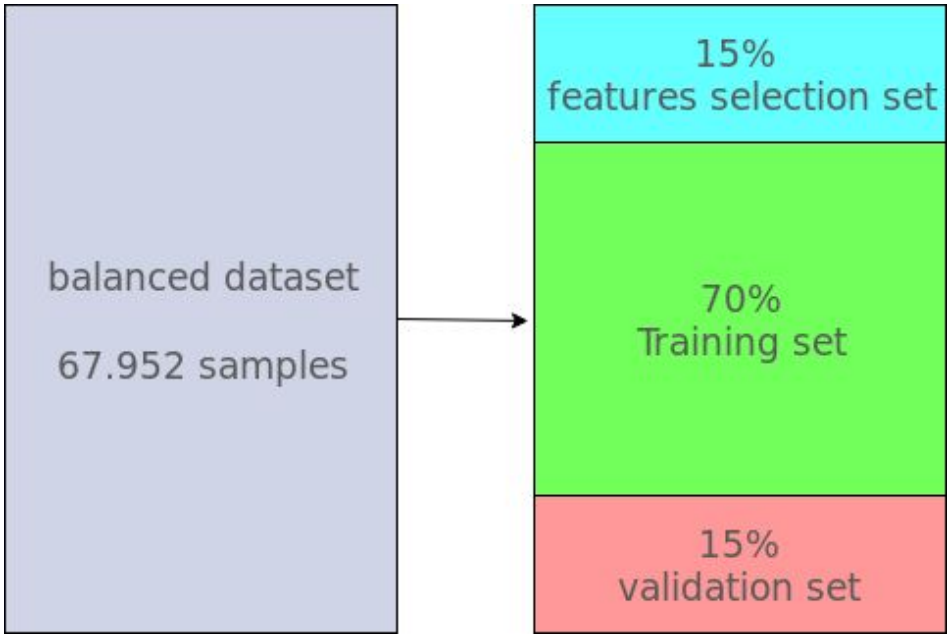






Dataset splitting

The dataset has been random splitted in three parts.



Shape: (10193, 241)
- class gamma = 5104
- class bg = 5089

Shape: (47566, 241)
- class gamma = 23760
- class bg = 23806

Shape: (10193, 241)
- class gamma = 5112
- class bg = 5081

Dimensionality reduction

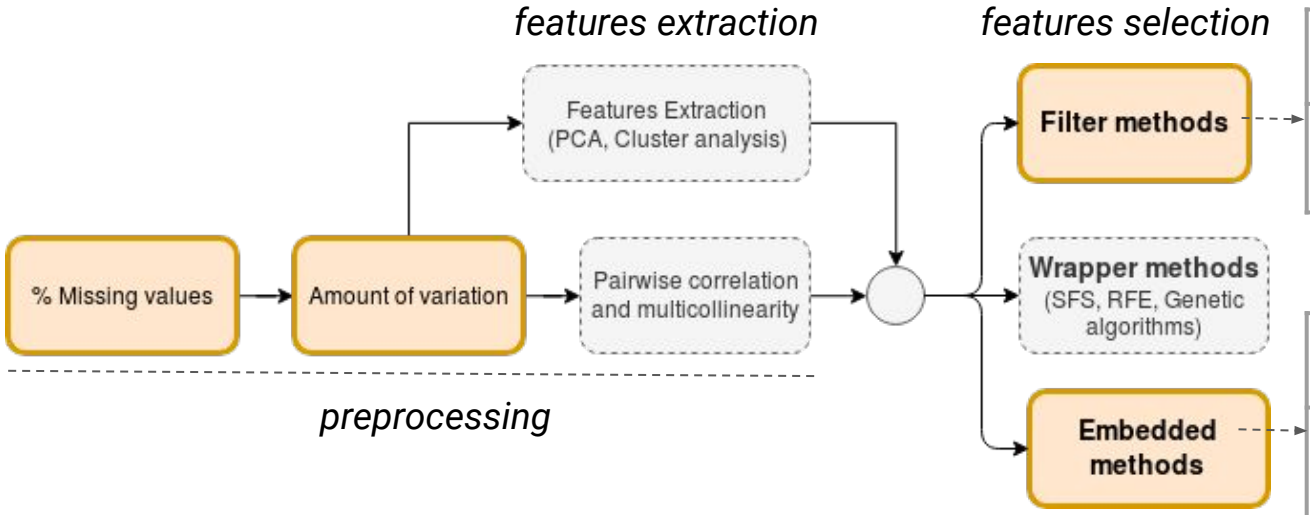
Why?

- Data storage is reduced.
- Machine learning models are simplified which can lead to increased generalization capability.
- Computational complexity for training and testing machine learning models is reduced.

Dimensionality reduction techniques are generally divided into two categories:

- **Features extraction** ⇒ linear or non-linear projection of data into a lower-dimensional subspace.
- **Features selection** ⇒ subset of the original features based on some performance criterion.

Dimensionality reduction workflow



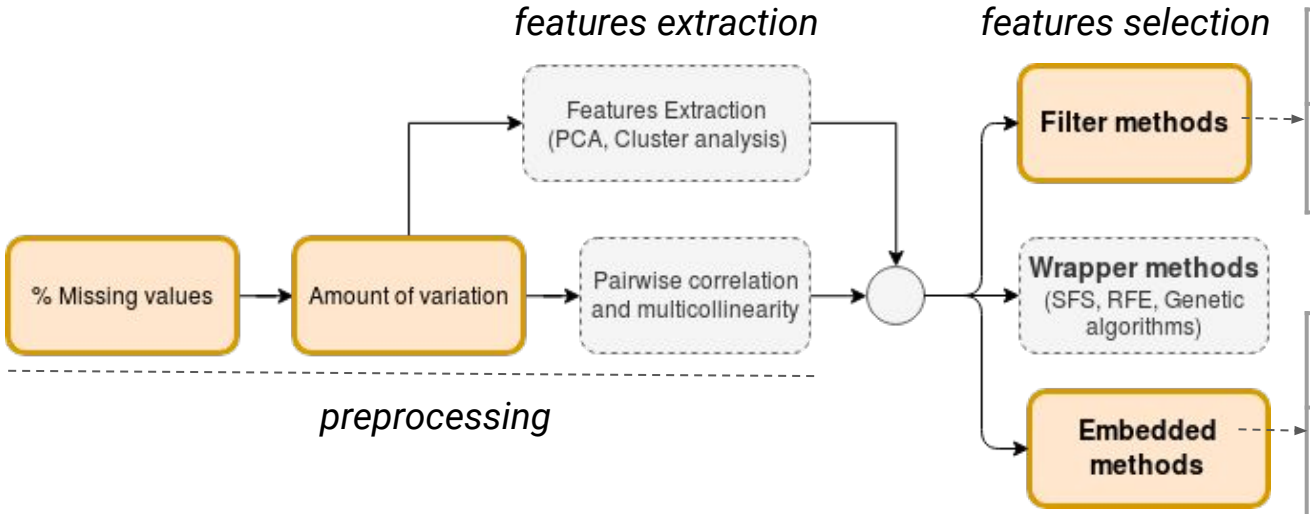
PEARSON CORR. COEFF.	Correlation with the target.
MUTUAL INFORMATION	Based on the concept of entropy of a random variable.

LASSO	It constrains or shrinks the model coefficients towards zero.
EXTRA TREES	It fits a number of randomized decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

..and many more techniques:

Bachu, Venkatesh & Anuradha, J.. (2019). A Review of Feature Selection and Its Methods. Cybernetics and Information Technologies. 19. 3. 10.2478/cait-2019-0001.

Dimensionality reduction workflow



PEARSON CORR. COEFF.	Correlation with the target.
MUTUAL INFORMATION	Based on the concept of entropy of a random variable.

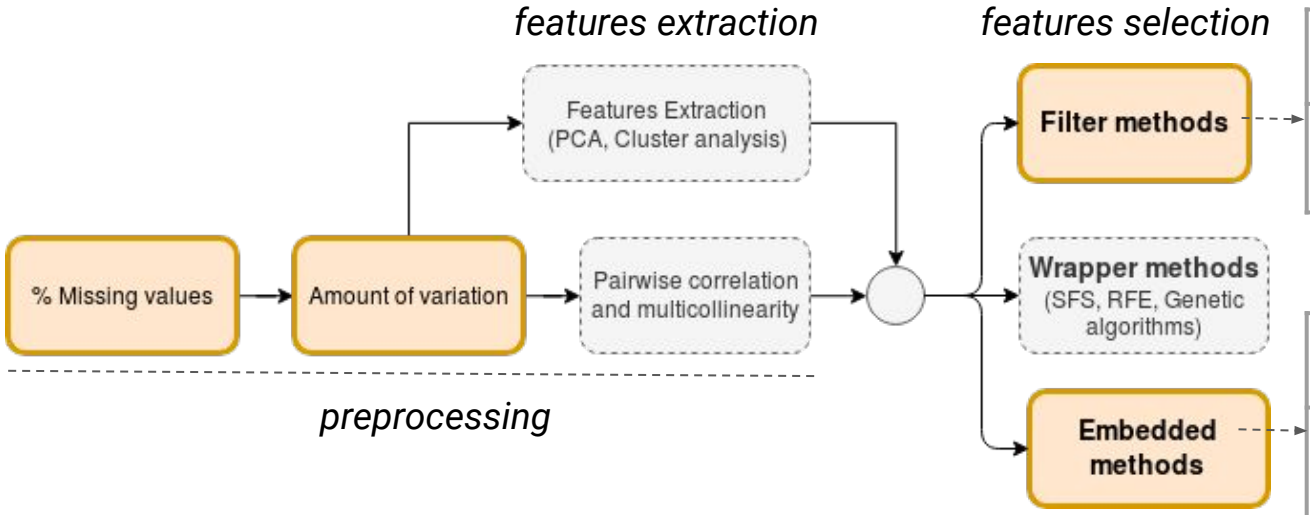
LASSO	It constrains or shrinks the model coefficients towards zero.
EXTRA TREES	It fits a number of randomized decision trees on various sub-samples of the dataset and uses averaging to improve the e accuracy and control ng.

$$\min_{w,c} \|w\|_1 + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1)$$

..and many more te

Bachu, Venkatesh ernetics and Information Technologies. 19. 3. 10.2478/cait-2019-0001.

Dimensionality reduction workflow



PEARSON CORR. COEFF.	Correlation with the target.
MUTUAL INFORMATION	Based on the concept of entropy of a random variable.

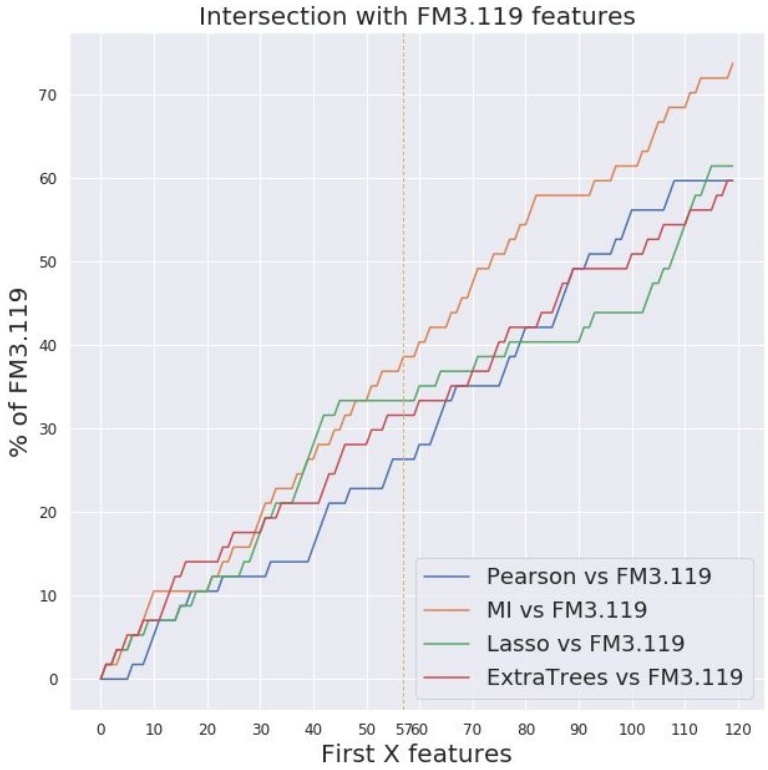
LASSO	It constrains or shrinks the model coefficients towards zero.
EXTRA TREES	It fits a number of randomized decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

..and many more techniques:

Bachu, Venkatesh & Anuradha, J.. (2019). A Review of Feature Selection and Its Methods. Cybernetics and Information Technologies. 19. 3. 10.2478/cait-2019-0001.

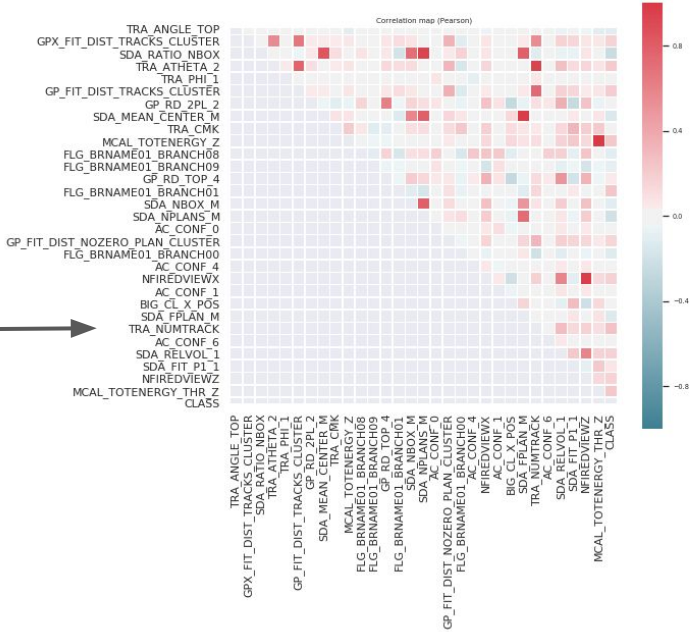
Automatically selected features vs FM3.119 features

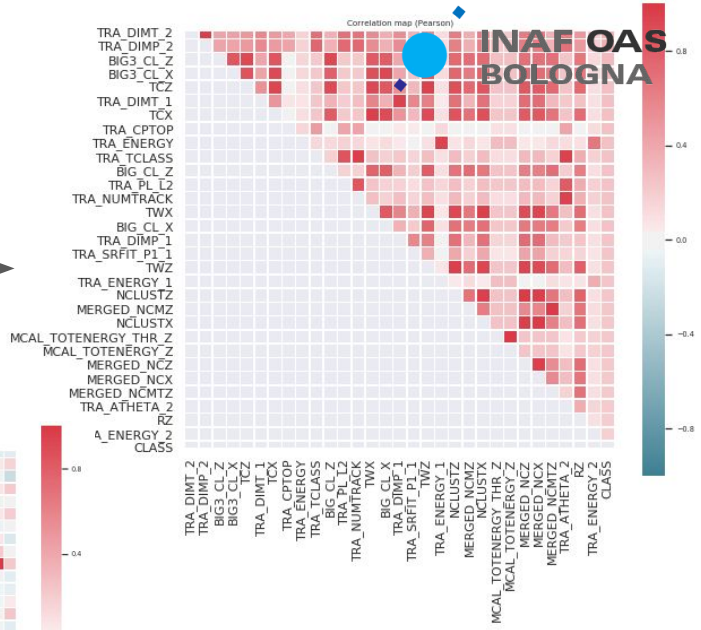
First 57 most discriminant features	
Method	**Intersection with FM3.119**
PEARSON CORR. COEFF.	26.31%
MUTUAL INFORMATION	**38.6 %**
LASSO	35.09 %
EXTRA TREES	33.33 %



Pairwise correlations

First 30 most discriminant features	
Method	**Pairwise correlations (corr > 0.6)**
PEARSON CORR. COEFF.	26.6 %
MUTUAL INFORMATION	2.6 %
LASSO	**2.1 %**
EXTRA TREES	4.8 %





Training of the ML algorithms

INAF OAS BOLOGNA

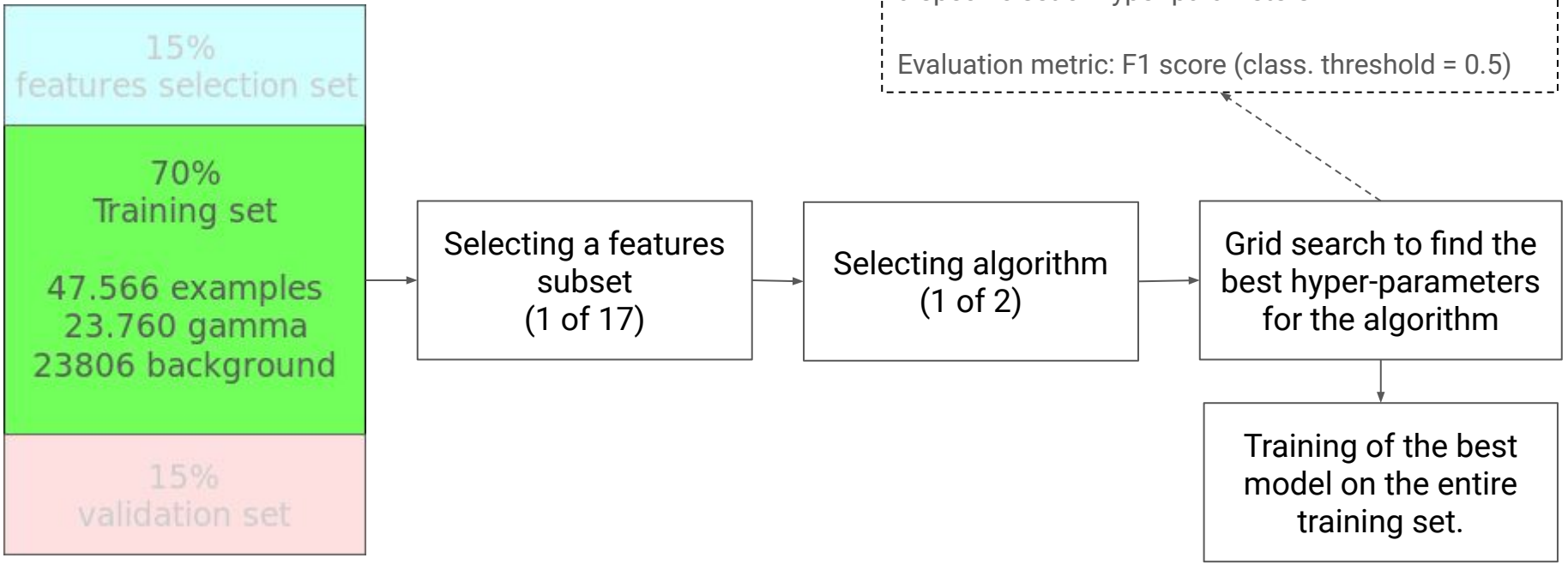
Features subsets		Algorithm	Hyper-parameters space
Features selection method	**Number of features**		
FM3.119	57	Random Forest	'criterion': ['gini', 'entropy'], 'n_estimators': [900, 1100, 1300, 1500, 1700], 'max_depth': [15, 20, 25, 30, 35], 'max_features': ['sqrt', 'log2'] **⇒ 100 models**
ALL features	241		
Pearson	30, 60 ,120, 180	Gradient Boosting	'learning_rate': [1, 0.1], 'n_estimators': [300 ,500, 700, 900, 1100], 'max_depth': [5, 10, 15, 20, 25], 'max_features': ['sqrt', 'log2'] **⇒ 100 models**
Mutual Information	30, 60 ,120, 180		
Lasso	30, 60 ,120		
Extra Trees	30, 60 ,120, 180		
⇒ 17 features subsets		**Total models to train** = 17 * (100+100) = 3.400	

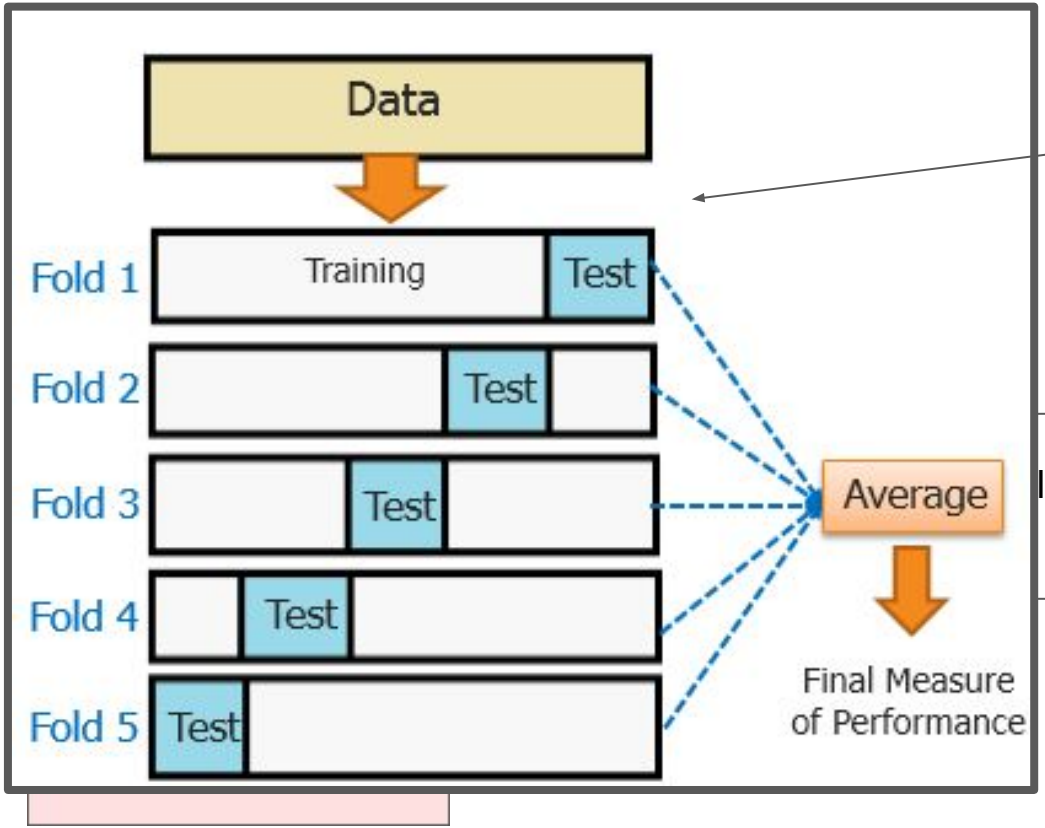
Training/Testing workflow

INAF OAS BOLOGNA

The **5-Fold cross-validation** method has been used to assess the performance of each model built with a specific set of hyper-parameters.

Evaluation metric: F1 score (class. threshold = 0.5)





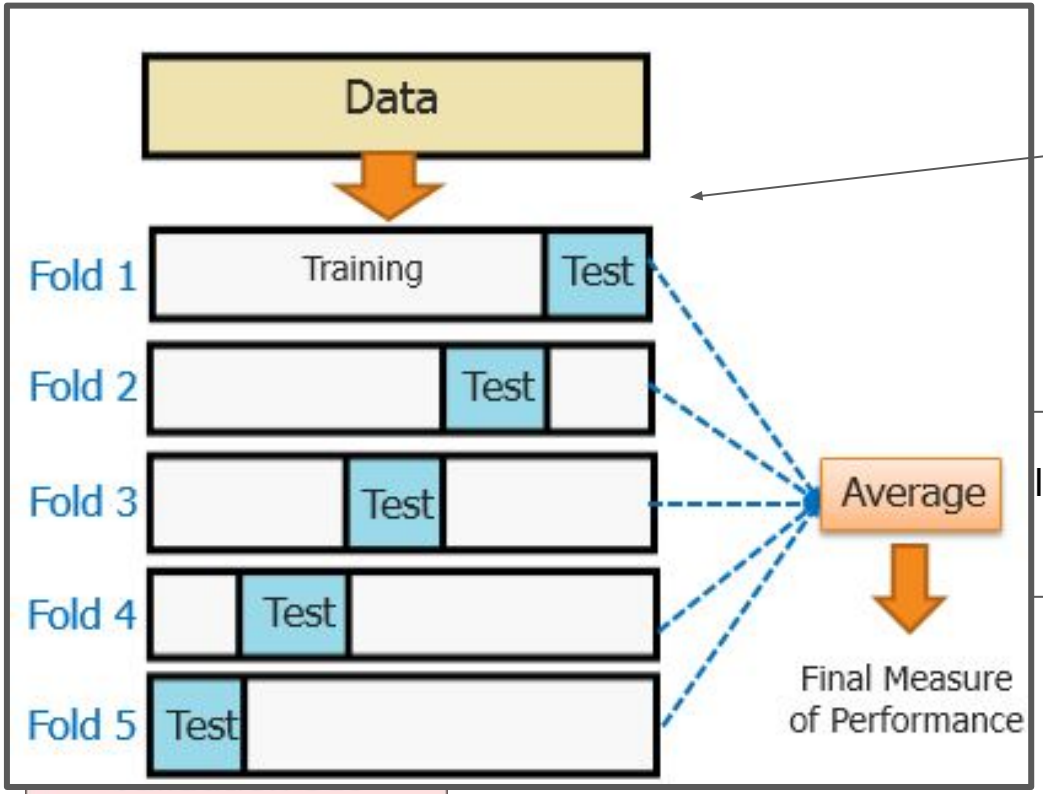
The **5-Fold cross-validation** method has been used to assess the performance of each model built with a specific set of hyper-parameters.

Evaluation metric: F1 score (class. threshold = 0.5)

lecting algorithm
(1 of 2)

Grid search to find the best hyper-parameters for the algorithm

Training of the best model on the entire training set.



The **5-Fold cross-validation** method has been used to assess the performance of each model built with a specific set of hyper-parameters.

Evaluation metric: F1 score (class. threshold = 0.5)

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

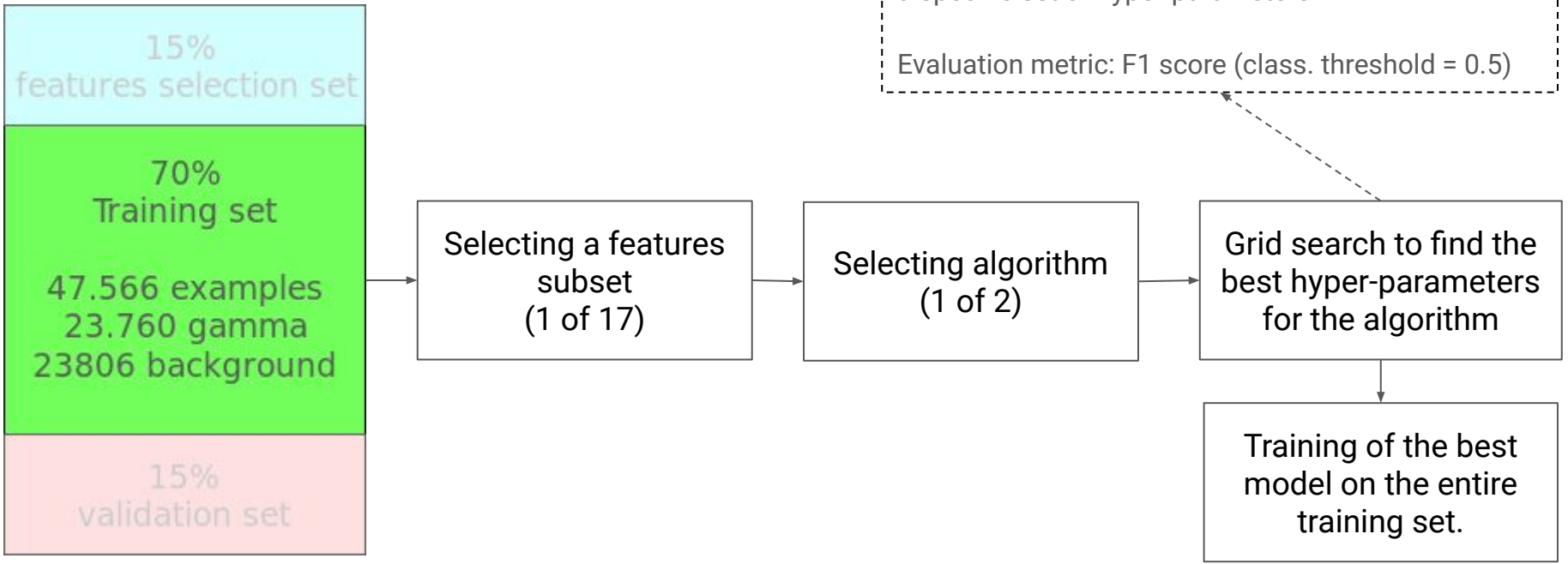
$$\text{Recall} = \frac{tp}{tp + fn}$$

Training/Testing workflow

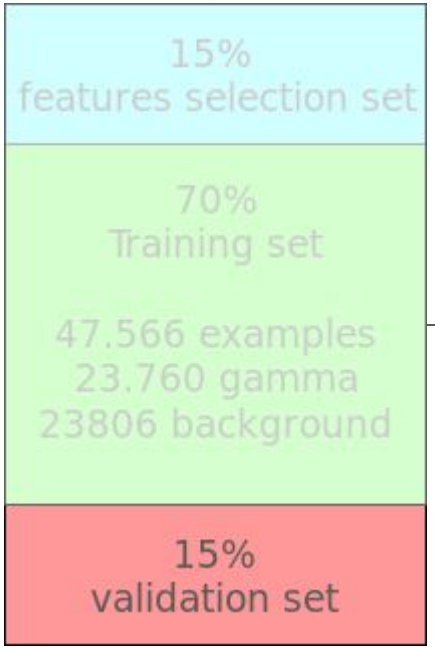
INAF OAS BOLOGNA

The **5-Fold cross-validation** method has been used to assess the performance of each model built with a specific set of hyper-parameters.

Evaluation metric: F1 score (class. threshold = 0.5)



Training/Testing workflow



The **5-Fold cross-validation** method has been used to assess the performance of each model built with a specific set of hyper-parameters.

Evaluation metric: F1 score (class. threshold = 0.5)

Selecting a features subset (1 of 17)

Selecting algorithm (1 of 2)

Grid search to find the best hyper-parameters for the algorithm

Training of the best model on the entire training set.

Evaluation of the best model on the validation set.

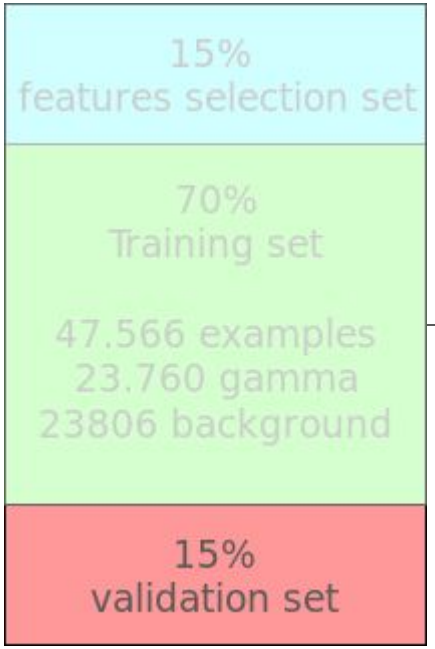
Evaluation metric: F1 score (class. threshold = 0.5)

Training/Testing workflow

INAF OAS BOLOGNA

The **5-Fold cross-validation** method has been used to assess the performance of each model built with a specific set of hyper-parameters.

Evaluation metric: F1 score (class. threshold = 0.5)



Selecting a features subset (1 of 17)

Selecting algorithm (1 of 2)

Grid search to find the best hyper-parameters for the algorithm

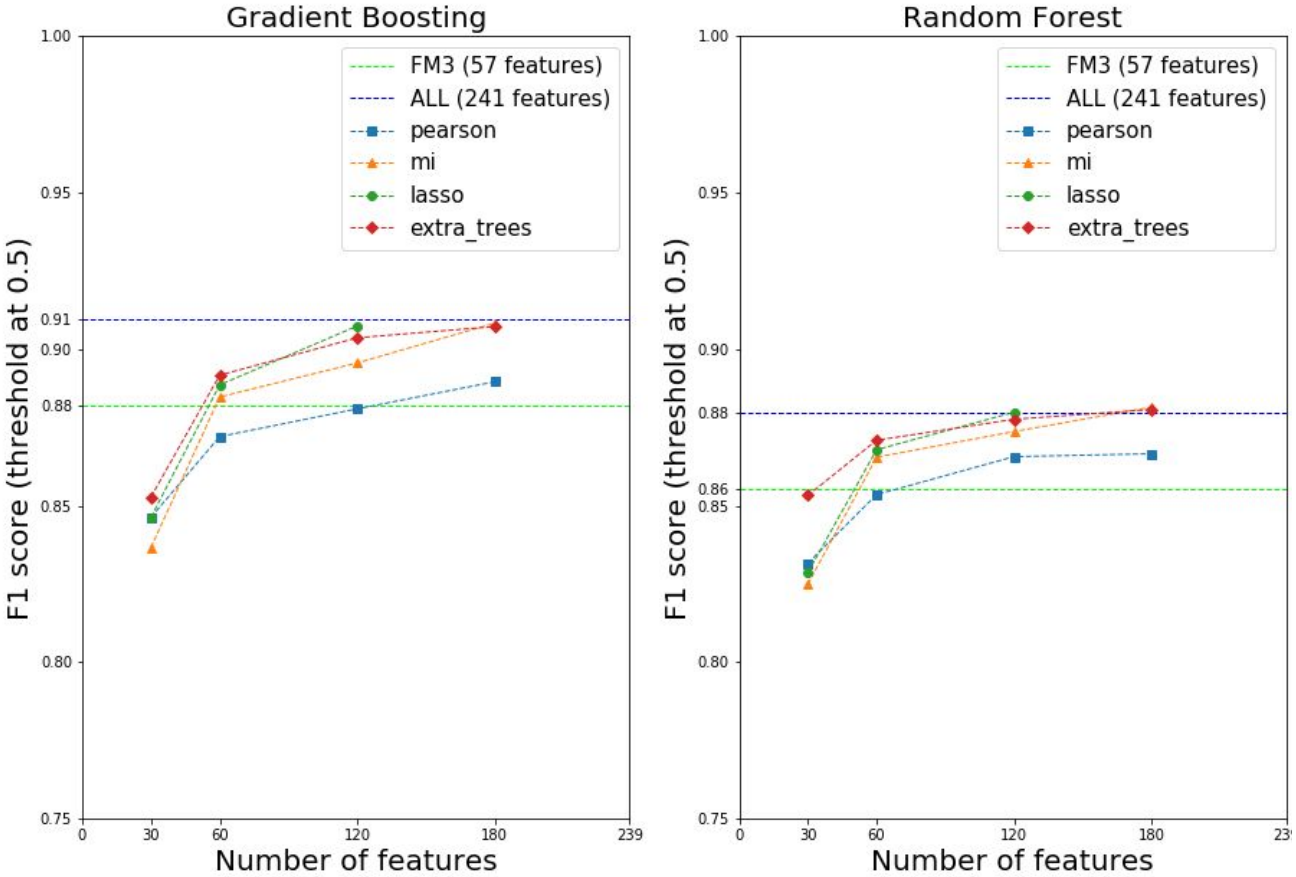
Training of the best model on the entire training set

Evaluation of the best model on the validation set

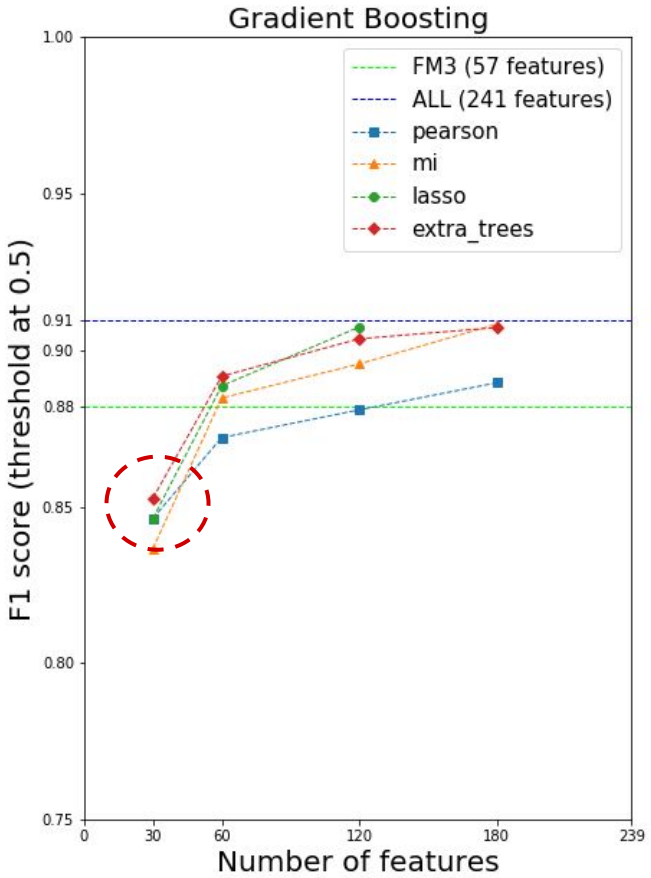
34 fine-tuned models

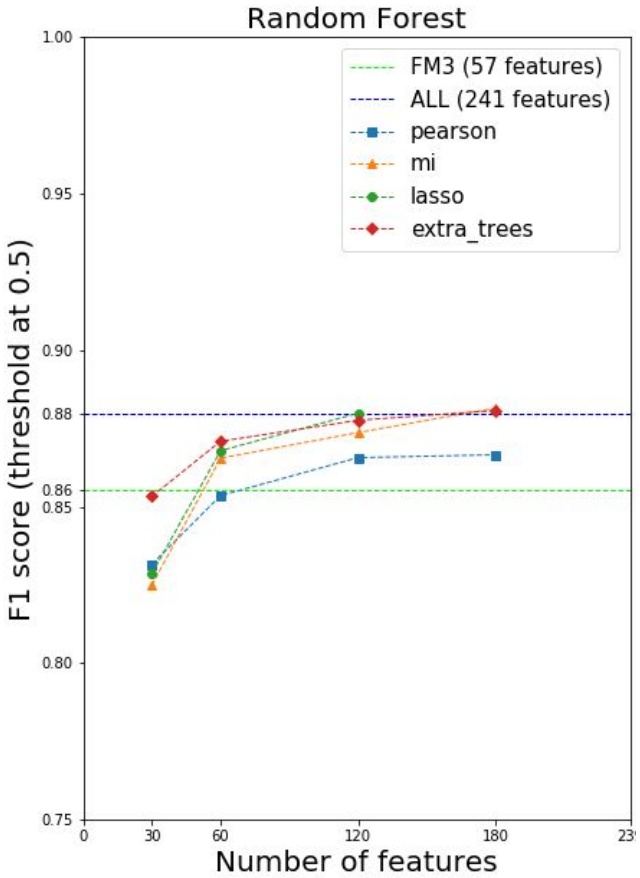
Evaluation metric: F1 score (class. threshold = 0.5)

Performance of the 34 models



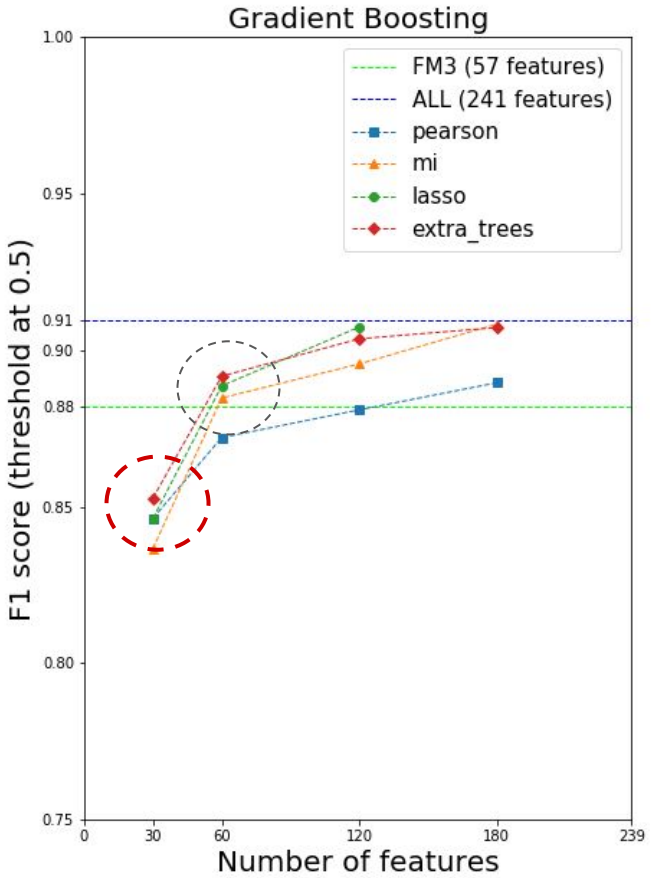
Performance of the 34 models

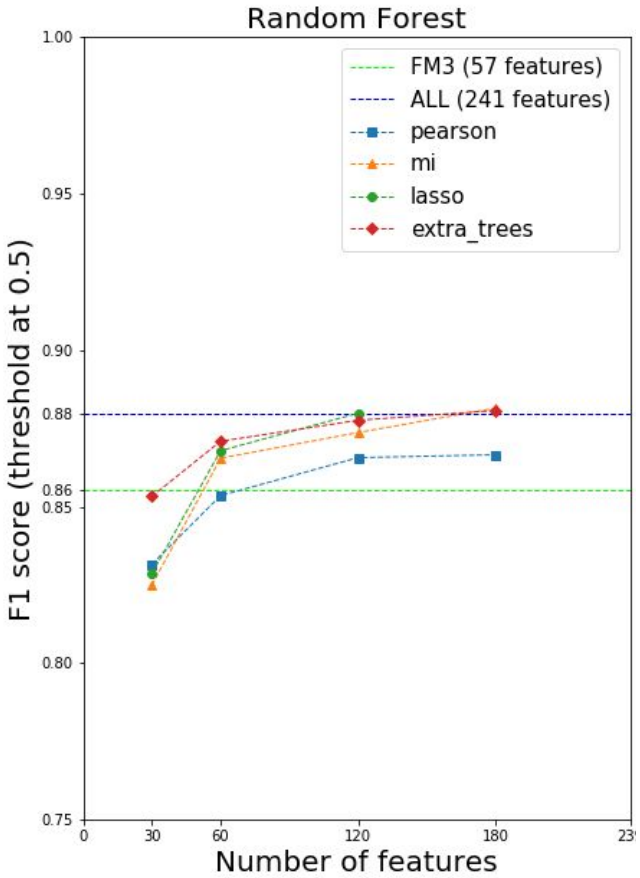




4451	661
869	4212

Performance of the 34 models

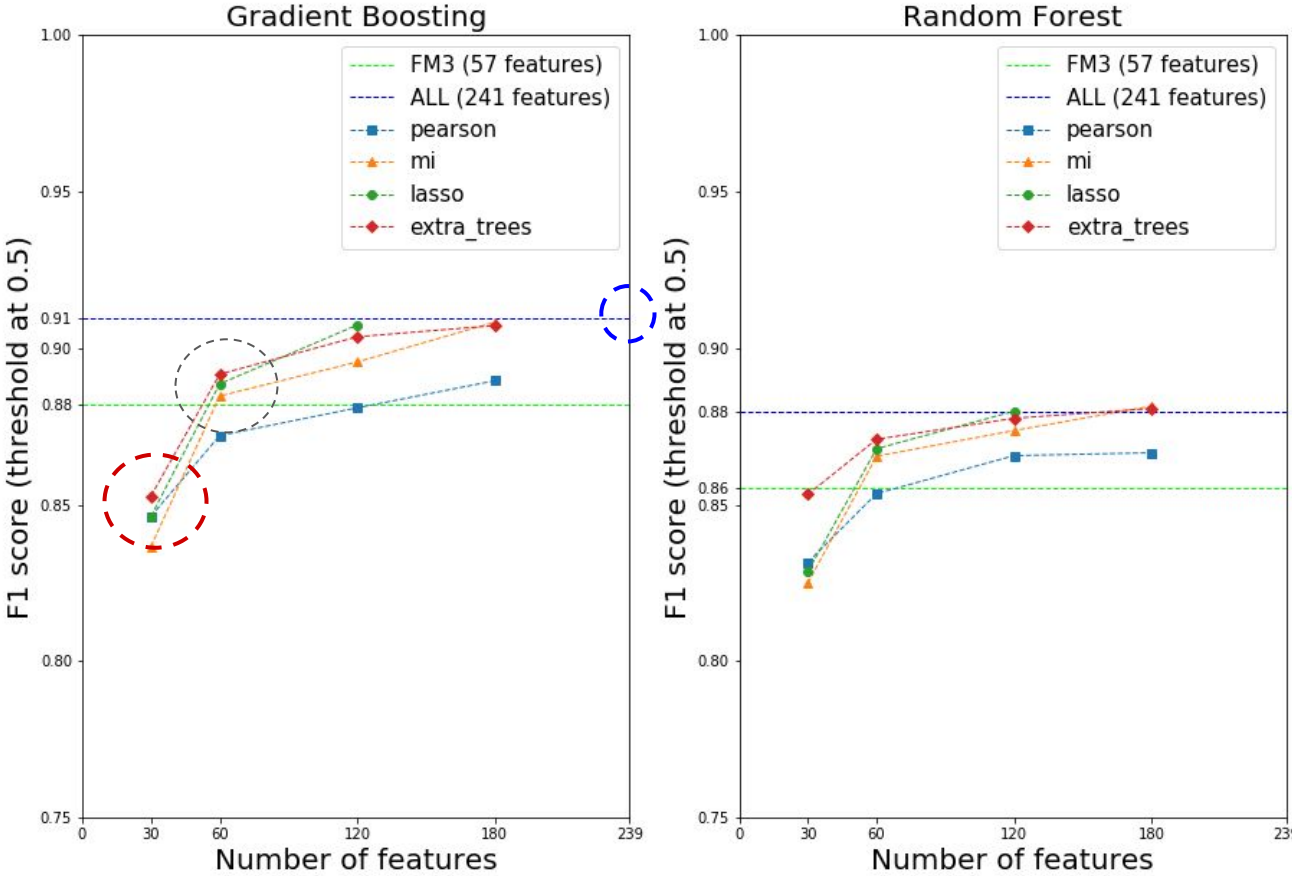




4451	661
869	4212

Performance of the 34 models

4741	371
534	4547

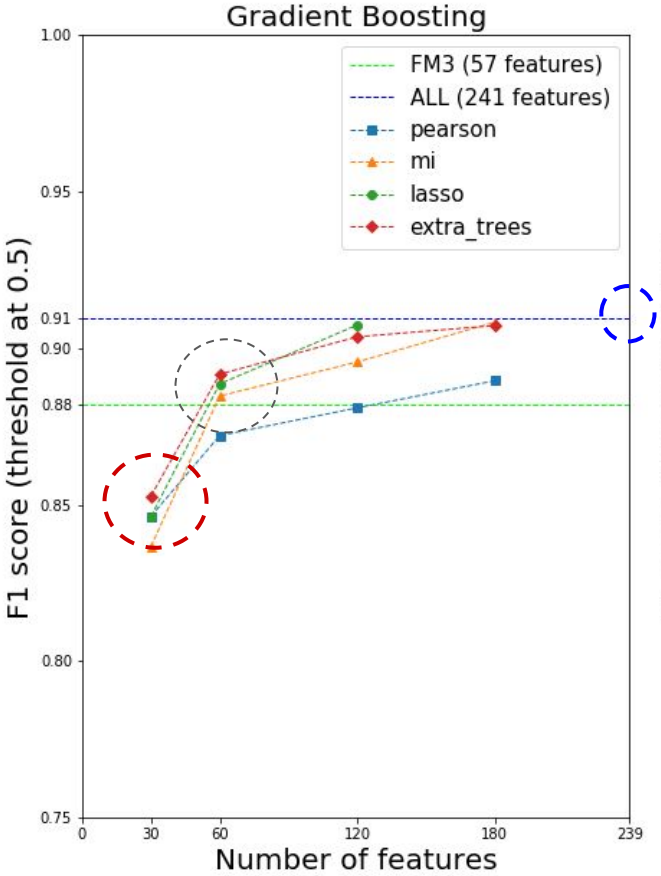


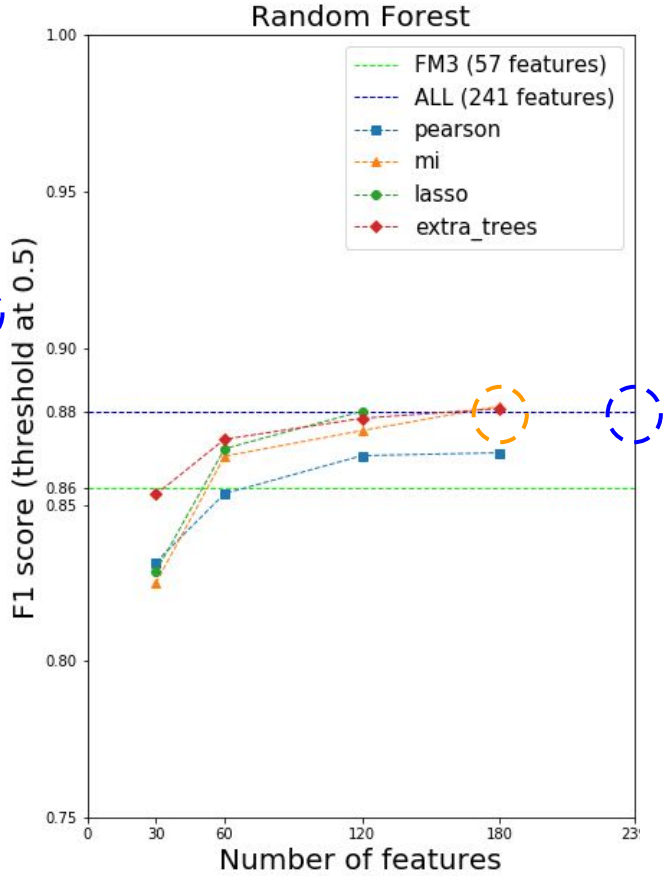
4451	661
869	4212

Performance of the 34 models

INAF OAS BOLOGNA

4741	371
534	4547





4710	402
760	4321

4701	411
770	4311

4451	661
869	4212

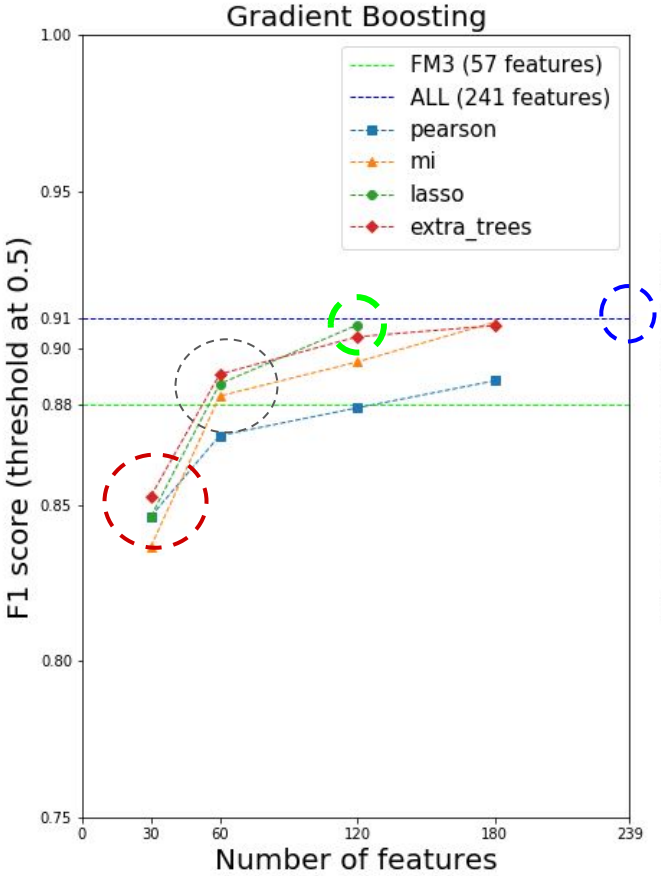
Performance of the 34 models

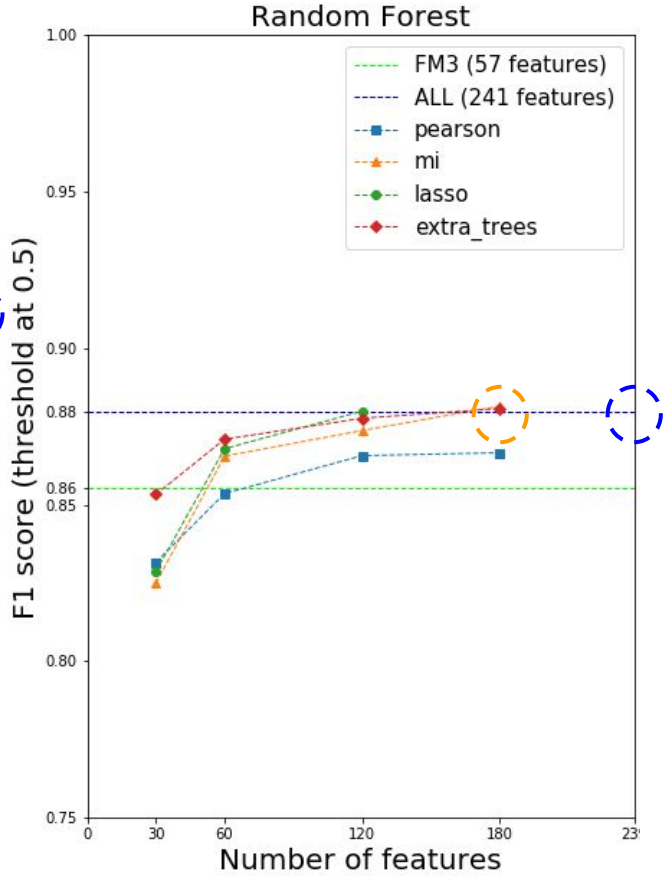
INAF OAS BOLOGNA

4741	371
534	4547

4783	329
588	4493

4451	661
869	4212





4710	402
760	4321

4701	411
770	4311

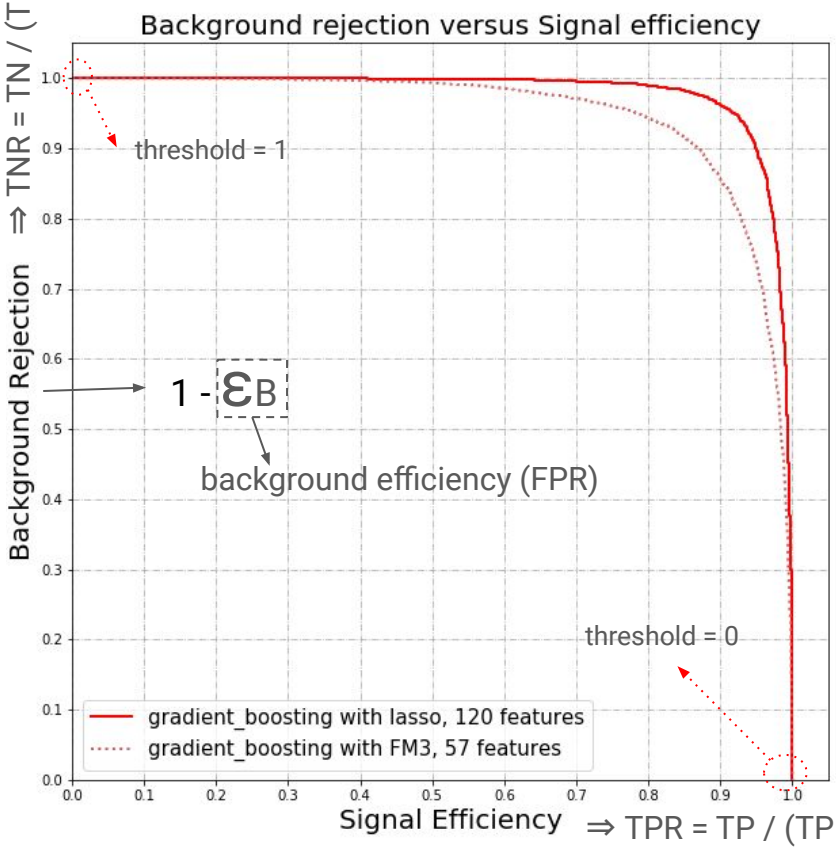
Choosing the final model

INAF OAS BOLOGNA

Algorithm	Fs method	Number of features	F1 score ↓	ROCAUC
Gradient Boosting	ALL features	239	0.909491	0.911164
Gradient Boosting	Mutual Information	180	0.908302	0.910746
Gradient Boosting	Lasso	120	0.907402	0.909958
...	...	...	...	...

Algorithm	Fs method	Number of features	F1 score ↓	ROCAUC
Random Forest	Mutual Information	180	0.881477	0.885892
Random Forest	Extra Trees	180	0.880734	0.885109
Random Forest	Lasso	120	0.879837	0.883844
...	...	...	...	...

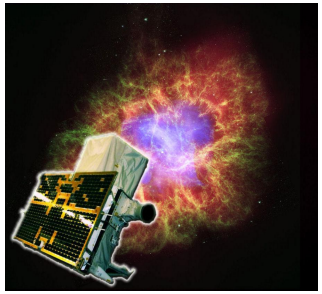
BG rejection vs Signal efficiency



Background rejection vs Signal efficiency	$\varepsilon_{B=0.01}$	$\varepsilon_{B=0.1}$	$\varepsilon_{B=3}$
Gradient Boosting with Lasso (120 features)	0.798	0.949	0.983
Gradient Boosting with FM3.119 (57 features)	0.560	0.869	0.958
(production filter) AdaBoost with FM3.119 (57 features)	0.747	0.925	0.976

Conclusions

- A multi-collinearity analysis is required to completely remove the correlation among the features.
- Features selection techniques MI, Lasso and Extra Trees selected a subset of features that generated a better model with respect to manual selection.
- The best performances were NOT achieved using a subset of automatic selected features but instead they were achieved training the Gradient Boosting model using all the features (Gradient Boosting performs features selection implicitly).
- With ~50% of the total features (120) selected by Lasso, it has been possible to obtain performance really close to the best results. Data storage requirement is reduced.
- With a ~12% of the total features (30) selected by Extra Trees, it has been possible to obtain a quite good 0.85 F1 score and potentially an easy interpretable model (I need the domain expert, now!).



INAF OAS BOLOGNA

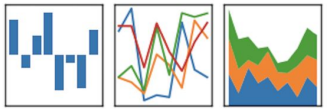
Thank you for your attention!
Any questions?

jupyter

NumPy

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



scikit learn

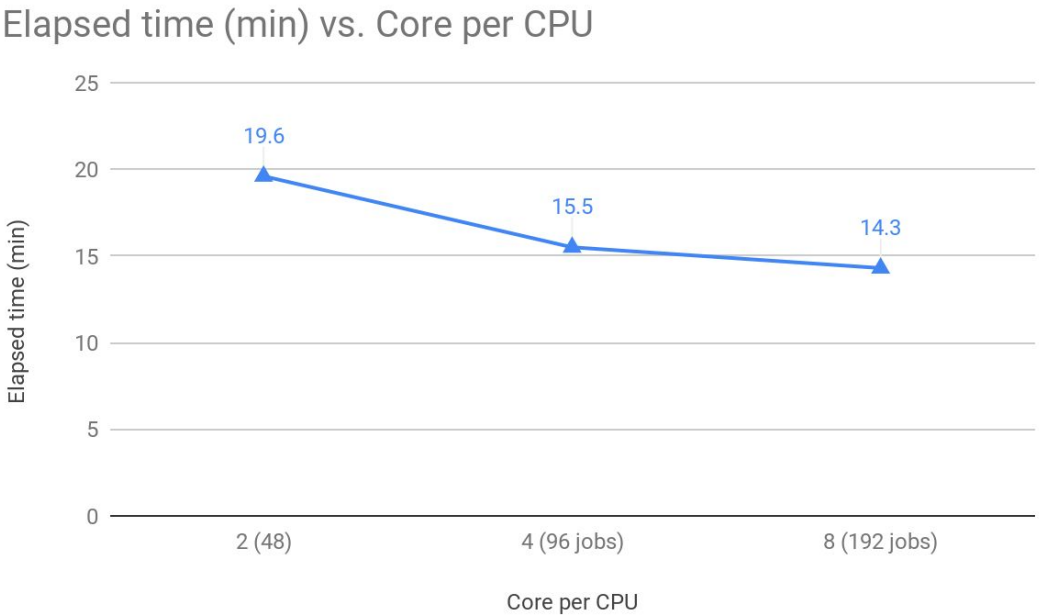
matplotlib

EXTRA: CVGridSearch mt scalability

Feature selection hyper-parameters space
("mi", 10)

Random forest hyper-parameters space
grid_params_rf = [{ 'rfc__criterion': ['gini', 'entropy'],
'rfc__n_estimators': [300, 500, 700, 900],
'rfc__max_features': ['sqrt', 'log2'],
'rfc__max_depth': [10, 15, 20],
}]

CPU(s)	192
Thread(s) per core	8
Core(s) per socket	6
Socket(s)	4



⇒ 96 will be used to run the Grid Search