UNIVERSITÀ DI PISA

RANDOM FORESTS

Project Presentation of Paper n.11

4th September 2023

Bombino Biancamaria
Fabbri Lucia
Frasson Andrea

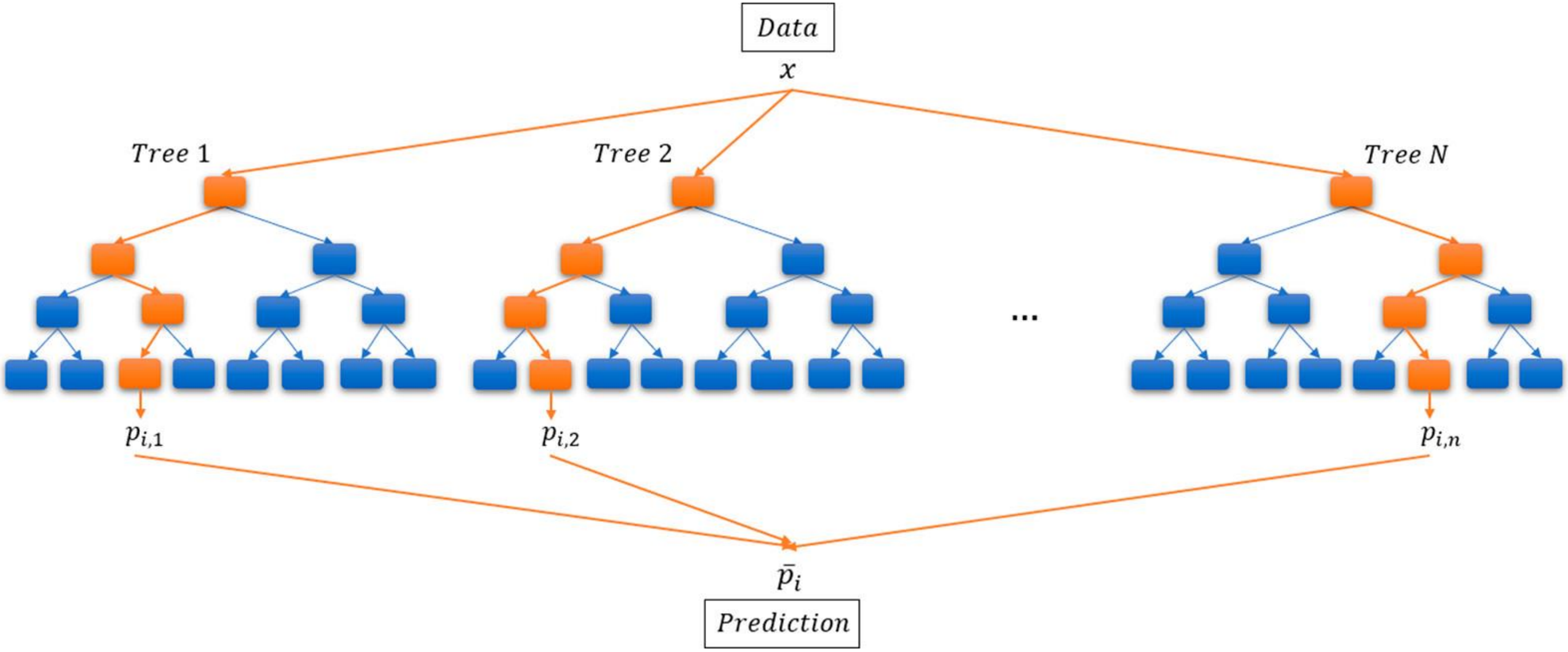
STATISTICS FOR DATA SCIENCE
A.Y. 2022/2023

WORKFLOW

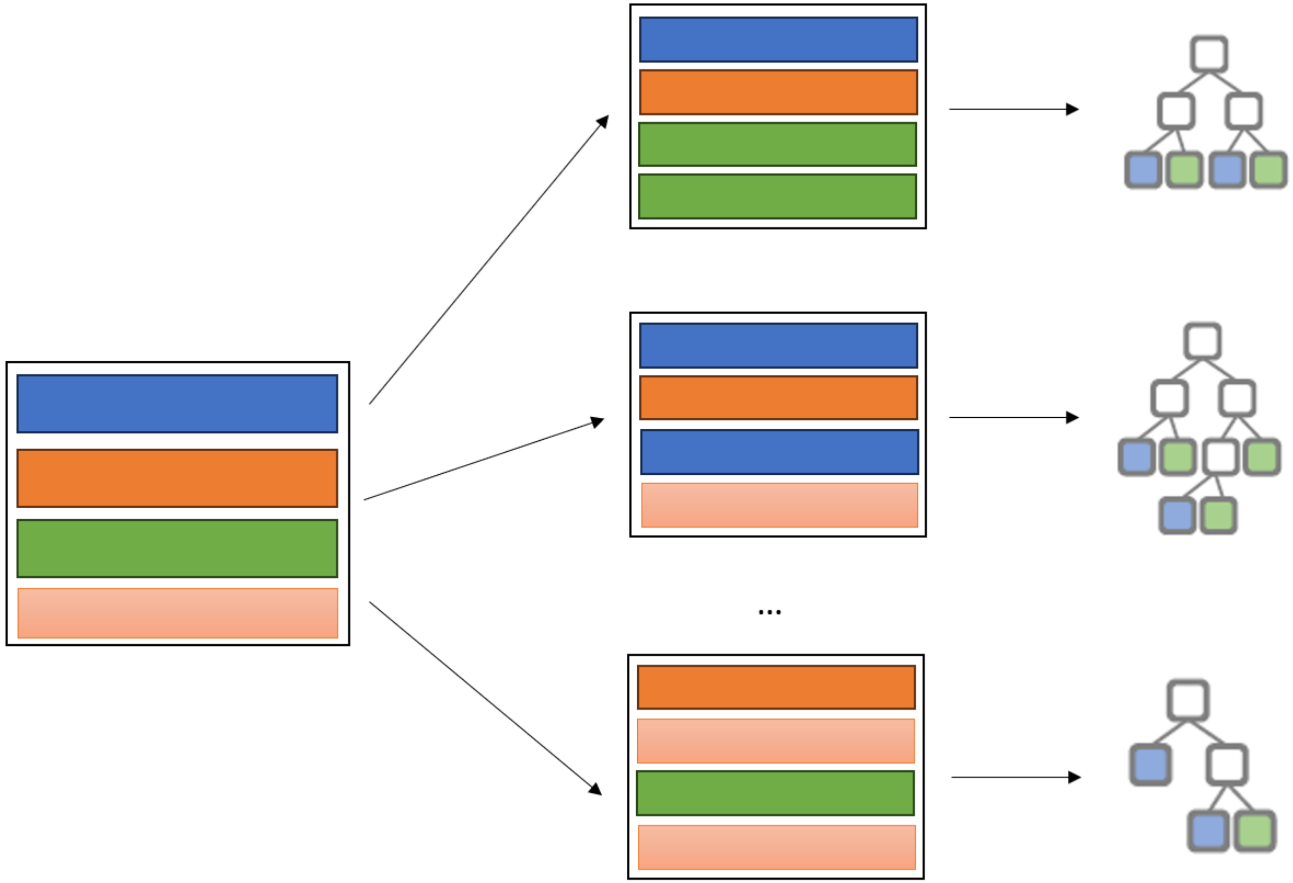
- Introduction to Random Forest
- Convergence
- Strength and Correlation
- Datasets
- Experiments
- Results
- Estimates Strength and Correlation
- Graphs
- Conclusions

WHAT ARE RANDOM FORESTS

Combinations of tree predictors such that each tree depends on the values of random vectors sampled independently and with the same distribution for all trees in the forest.



HOW DO THEY WORK



For the k-th tree, a random vector Θ_k is generated, independent of the past random vectors $\Theta_1, \ldots, \Theta_{k-1}$ but with the same distribution.

Then a tree is grown using the training set and Θ_k, resulting in a classifier $h_k(\boldsymbol{X})$.

After a large number of trees (K) is generated, the forest is built.

The random forests algorithm works by aggregating the predictions made by the trees, considering the most popular class predicted for a record **x**.

CONVERGENCE

Given an ensamble of K classifiers $h_1(\boldsymbol{x}), h_2(\boldsymbol{x}), \ldots, h_K(\boldsymbol{x})$, the margin function is

$$mg(\mathbf{X},Y) = av_k I(h_k(\mathbf{X}) = Y) - \max_{j \neq Y} av_k I(h_k(\mathbf{X}) = j)$$

The margin function is used to measure classification confidence.
The greater the margin, the more confidence the classification.

Then the generalization error is given by $PE^* = P_{\mathbf{X},Y}(mg(\mathbf{X},Y) < 0)$.

An important result is that, when the number of trees increases

$$PE^* \longrightarrow P_{\mathbf{X},Y}\left(P_\Theta(h(\mathbf{X},\Theta) = Y) - \max_{j \neq Y} P_\Theta(h(\mathbf{X},\Theta) = j) < 0\right)$$

This result explains why random forests do not overfit as more trees are added, but produce a limiting value of the generalization error.

STRENGTH AND CORRELATION

An upper bound of the generalization error can be derived in terms of two parameters, the strength of the whole forest and the correlation between different members of the forest.

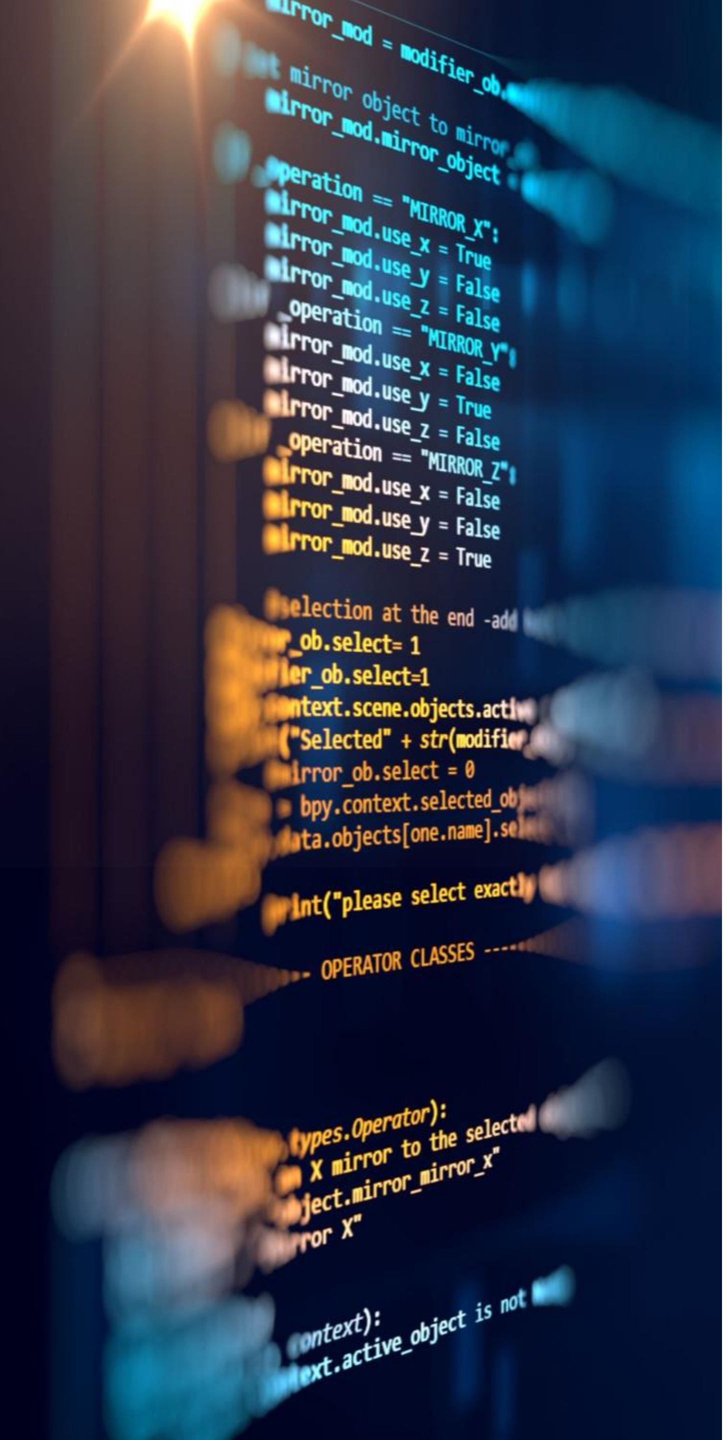
$$\boxed{PE^* \leq var(mr)/s^2}$$

Given the margin function for a random forest as

$$mr(\mathbf{X}, Y) = P_\Theta(h(\mathbf{X}, \Theta) = Y) - \max_{j \neq Y} P_\Theta(h(\mathbf{X}, \Theta) = j)$$

The strength of the set of classifiers is $s = E_{\mathbf{X},Y}[mr(\mathbf{X}, Y)]$

While the *var(mr)* can be replaced by $\bar{\rho}(1 - s^2)$, where $\bar{\rho}$ is the correlation between two different members of the forest, in terms of the raw margin functions.



DATASETS

DATASET	TRAIN SIZE	TEST SIZE	FEATURES	CLASSES
Biopsy	512	171	11	2
Wine	3673	1225	12	7
Stroke	3681	1228	11	2
Ionosphere	263	88	35	2
HeartDisease	688	230	7	2
PimaIndiansDiabetes	576	192	9	2

EXPERIMENTS

We decided to implement different strategies that could improve the performance of random forests.

Forest-RI

Select at random, at each node a small group of input variables to split on. Choice of F.

Single Feature

Trees grow using only one random feature.

Forest-RC

Define linear combinations of input features, select a random subset and use the best split.

RESULTS

DATASET	Forest-RI	Single Tree	Forest-RC	Adaboost
Biopsy	0.029	0.017	0.028	0.052
Wine	0.317	0.317	0.328	0.44
Stroke	0.036	0.036	0.037	0.048
Ionosphere	0.043	0.120	0.029	0.056
HeartDisease	0.230	0.312	0.305	0.283
PimaIndiansDiabetes	0.240	0.311	0.249	0.218

ESTIMATES FOR STRENGTH & CORRELATION

To estimate strength and correlation, we decided to use out-of-bag methods. Given a specific training set *T*, a classifier is built based on a bootstrap sample T_k. To obtain the out-of-bag classifier, for each training sample $(y, \boldsymbol{x})$ aggregate the votes only over those classifiers for which T_k doesn't contain $(y, \boldsymbol{x})$.

These quantities give internal estimates that clarify accuracy and how to improve it.

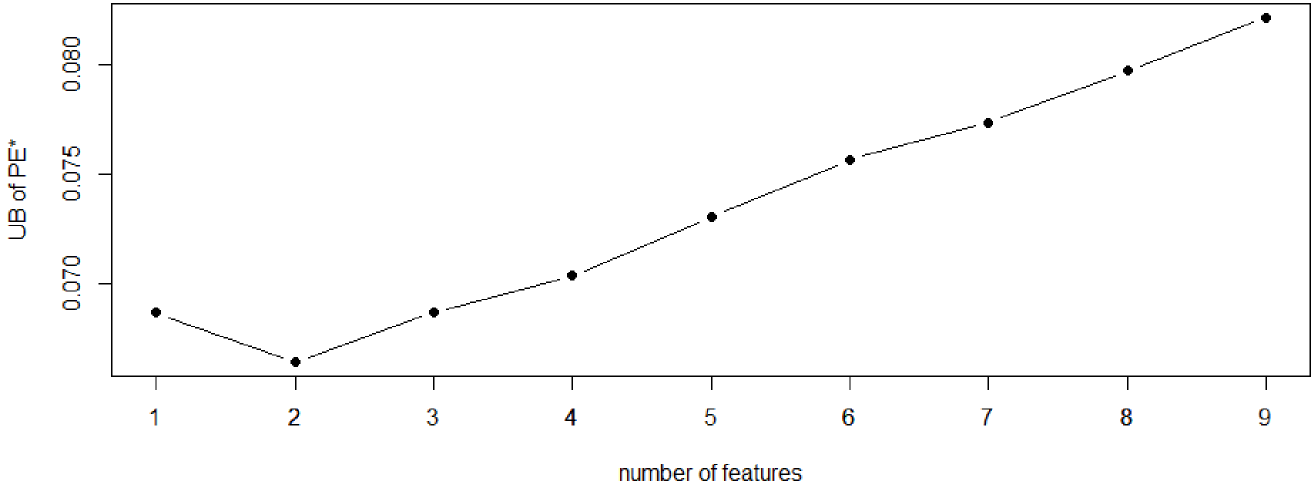
The first thing is to compute an estimate of $P_\Theta(h(\boldsymbol{x}, \Theta) = j)$ for every class j. In the out-of-bag setting we use $Q(\boldsymbol{x}, j) = \sum_k I(h(\boldsymbol{x}, \Theta) = j; (y, \boldsymbol{x}) \notin T_k) \,/\, \sum_k I\big((y, \boldsymbol{x}) \notin T_k\big)$. Thus, $Q(\boldsymbol{x}, j)$ is the out-of-bag proportion casted at *j*, substituting the right quantities and averaging over the training set gives the strength estimate.

The estimation of the correlation is computationally more expensive. The proposed procedure set $\bar{\rho} = var(mr)/E_\Theta[sd(\Theta)]^2$, where $sd(\Theta)$ is the standard deviation of the raw margin of the random forest.

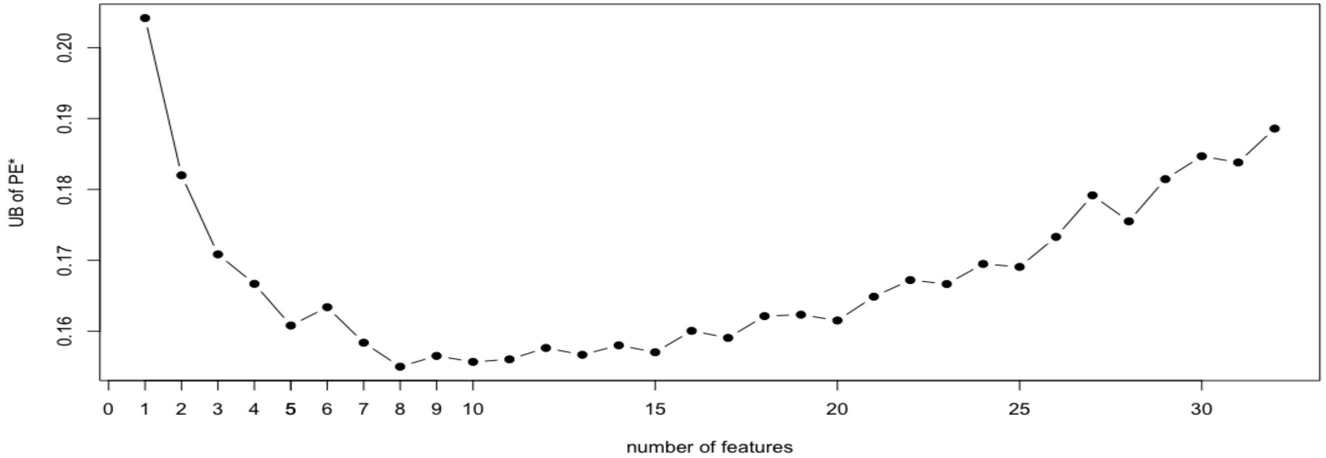
For the calculation of $sd(\Theta)$ we refer to the paper, however, the algorithm described involves calculating $sd_k(\Theta_\text{k})$ for each classifier and calculating the expected value. This procedure is expensive and, although it could be parallelized, we were able to implement it only for few examples.

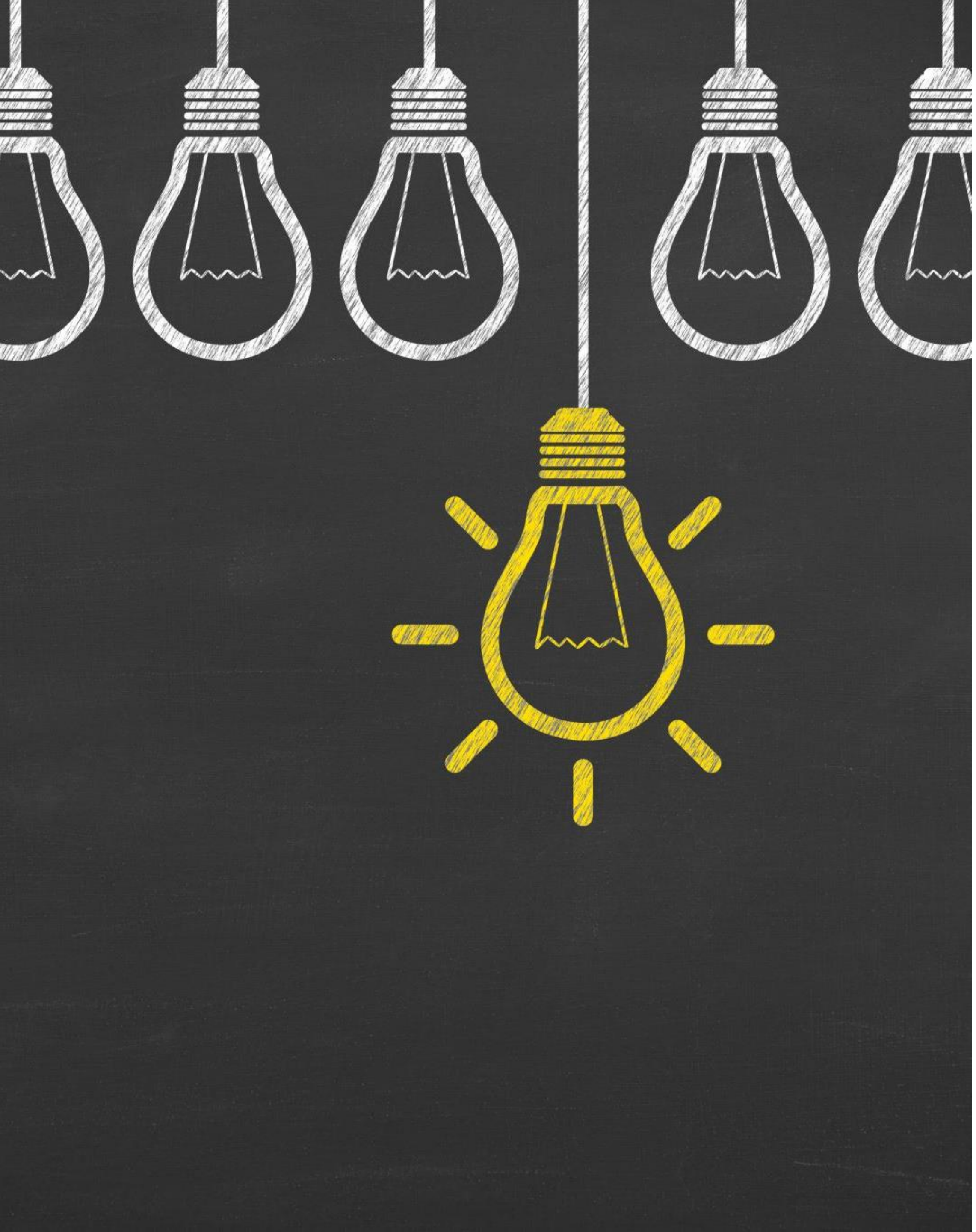
GRAPHS’ RESULTS

- PE* as the number of features changes for dataset 'biopsy’.



- PE* as the number of features changes for dataset 'Ionosphere'.





CONCLUSIONS

- Random forests are an effective tool in prediction, the results are competitive with boosting and other bagging method, without changing the train set.
- Because of the Law of Large Numbers, they do not overfit.
- Using out-of-bag estimation makes concrete the otherwise theoretical values of strength and correlation, giving insights on the behaviour of the classifier. This result is critical in those application where understand the interactions of variables is critical.

IN SUPREMÆ DIGNITATIS
1343

THANK YOU FOR YOUR ATTENTION