Andrey Lange

Skolkovo Institute of Science and Technology
a.lange@skoltech.ru

Abstract:
https://docs.google.com/document/d/1I88gZsei62NKkIATo5DnK0B8xPMzkHeWdr2NGhlRvXU/edit?usp=sharing

Отбор релевантных признаков в условиях мультиколлинеарности на примере биомаркеров рака легких

Skoltech

OPEN ACCESS Freely available online

PLoS one

Inferring Regulatory Networks from Expression Data Using Tree-Based Methods

Vân Anh Huynh-Thu[1,2]*, **Alexandre Irrthum**[1,2], **Louis Wehenkel**[1,2], **Pierre Geurts**[1,2]

1 Department of Electrical Engineering and Computer Science, Systems and Modeling, University of Liège, Liège, Belgium, **2** GIGA-Research, Bioinformatics and Modeling, University of Liège, Liège, Belgium

Abstract

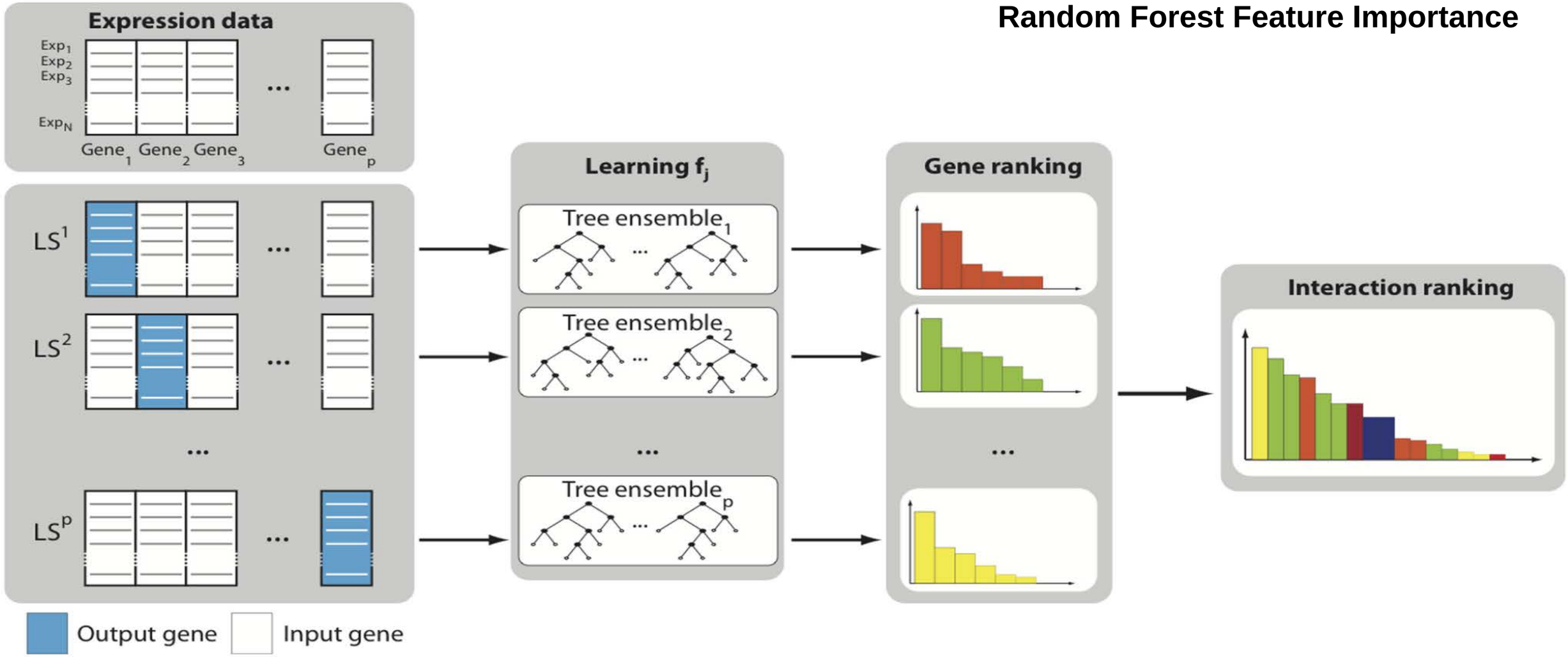
One of the pressing open problems of computational systems biology is the elucidation of the topology of genetic regulatory networks (GRNs) using high throughput genomic data, in particular microarray gene expression data. The Dialogue for Reverse Engineering Assessments and Methods (DREAM) challenge aims to evaluate the success of GRN inference algorithms on benchmarks of simulated data. In this article, we present GENIE3, a new algorithm for the inference of GRNs that was best performer in the DREAM4 *In Silico Multifactorial* challenge. GENIE3 decomposes the prediction of a regulatory network between p genes into p different regression problems. In each of the regression problems, the expression pattern of one of the genes (target gene) is predicted from the expression patterns of all the other genes (input genes), using tree-based ensemble methods Random Forests or Extra-Trees. The importance of an input gene in the prediction of the target gene expression pattern is taken as an indication of a putative regulatory link. Putative regulatory links are then aggregated over all genes to provide a ranking of interactions from which the whole network is reconstructed. In addition to performing well on the DREAM4 *In Silico Multifactorial* challenge simulated data, we show that GENIE3 compares favorably with existing algorithms to decipher the genetic regulatory network of *Escherichia coli*. It doesn't make any assumption about the nature of gene regulation, can deal with combinatorial and non-linear interactions, produces directed GRNs, and is fast and scalable. In conclusion, we propose a new algorithm for GRN inference that performs well on both synthetic and real gene expression data. The algorithm, based on feature selection with tree-based ensemble methods, is simple and generic, making it adaptable to other types of genomic data and interactions.

Citation: Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLoS ONE 5(9): e12776. doi:10.1371/journal.pone.0012776

Skoltech

Gene Regulatory Networks

Random Forest Feature Importance



Ranking of RNA-Protein relationship

RNA	ENSG00000121410_A1BG	ENSG00000268895_A1BG-AS1	ENSG00000175899_A2M	ENSG00000245105_A2M-AS1
cell_id				
45006fe3e4c8	0.0	0.0	0.0	0.0
d02759a80ba2	0.0	0.0	0.0	0.0
c016c6b0efa5	0.0	0.0	0.0	0.0
ba7f733a4f75	0.0	0.0	0.0	0.0
fbcf2443ffb2	0.0	0.0	0.0	0.0
...	...	...	...	...
650ee456f0f3	0.0	0.0	0.0	0.0
cc506e7707f5	0.0	0.0	0.0	0.0
a91f1b55a520	0.0	0.0	0.0	0.0
3a9882c98205	0.0	0.0	0.0	0.0
c91b6b2ccd3d	0.0	0.0	0.0	0.0

70988 rows × 22050 columns

Protein	CD86
cell_id	
45006fe3e4c8	1.167804
d02759a80ba2	0.818970
c016c6b0efa5	-0.356703
ba7f733a4f75	-1.201507
fbcf2443ffb2	-0.100404
...	...
650ee456f0f3	0.905420
cc506e7707f5	2.101247
a91f1b55a520	1.221313
3a9882c98205	-0.151433
c91b6b2ccd3d	-0.439299

70988 rows × 140 columns

Problem: Feature Importance of RNA for given Protein -?

Skoltech

Genome-Wide Association Studies

Genes (predictors)

	AC005884.1	AKAP2	CASS4	DPEP2	SHE	SPAAR	TGFBR2	RXFP2	DONSON	MYBL2	POLQ	WDR12
1	1.319020	0.306308	4.844284	7.162875	6.912610	2.253120	99.781881	0.146702	1.288607	0.796111	0.048998	1.615651
2	1.200746	0.853465	6.243791	4.687499	8.255939	3.801440	106.915618	0.036281	1.805294	1.035882	0.026332	1.101498
3	0.947373	0.274206	5.961070	8.961572	4.766719	4.001089	82.597865	0.388657	1.066370	1.081741	0.045116	1.951958
4	0.278170	0.212502	5.610866	7.054260	5.040815	1.168014	96.891945	0.151810	1.419171	1.986289	0.062065	1.734307
5	0.477756	0.644286	5.354550	7.240387	6.007856	3.378418	137.802210	0.167615	1.791950	1.254690	0.158858	1.805437
...	...	...	...	...	...	...	...	...	...	...	...	...
547	0.061612	0.045386	0.816050	0.646057	0.546420	0.204960	9.021273	0.000000	7.471296	44.949094	1.533881	3.570515
548	0.100758	0.000000	0.162020	0.333642	0.192032	0.093851	9.823080	0.000000	7.351857	27.971096	1.374930	5.511240
549	0.050332	0.008239	0.621920	0.376543	0.754627	0.294688	7.853928	0.000000	9.791333	43.363962	4.192173	6.678553
550	0.105428	0.069034	0.707861	0.586155	0.921614	0.252518	25.806540	0.000000	9.930669	26.574808	5.873322	2.946315
551	0.000000	0.050019	0.604174	0.589356	0.399525	0.192203	12.349765	0.007181	5.719141	67.706048	2.757292	3.305939

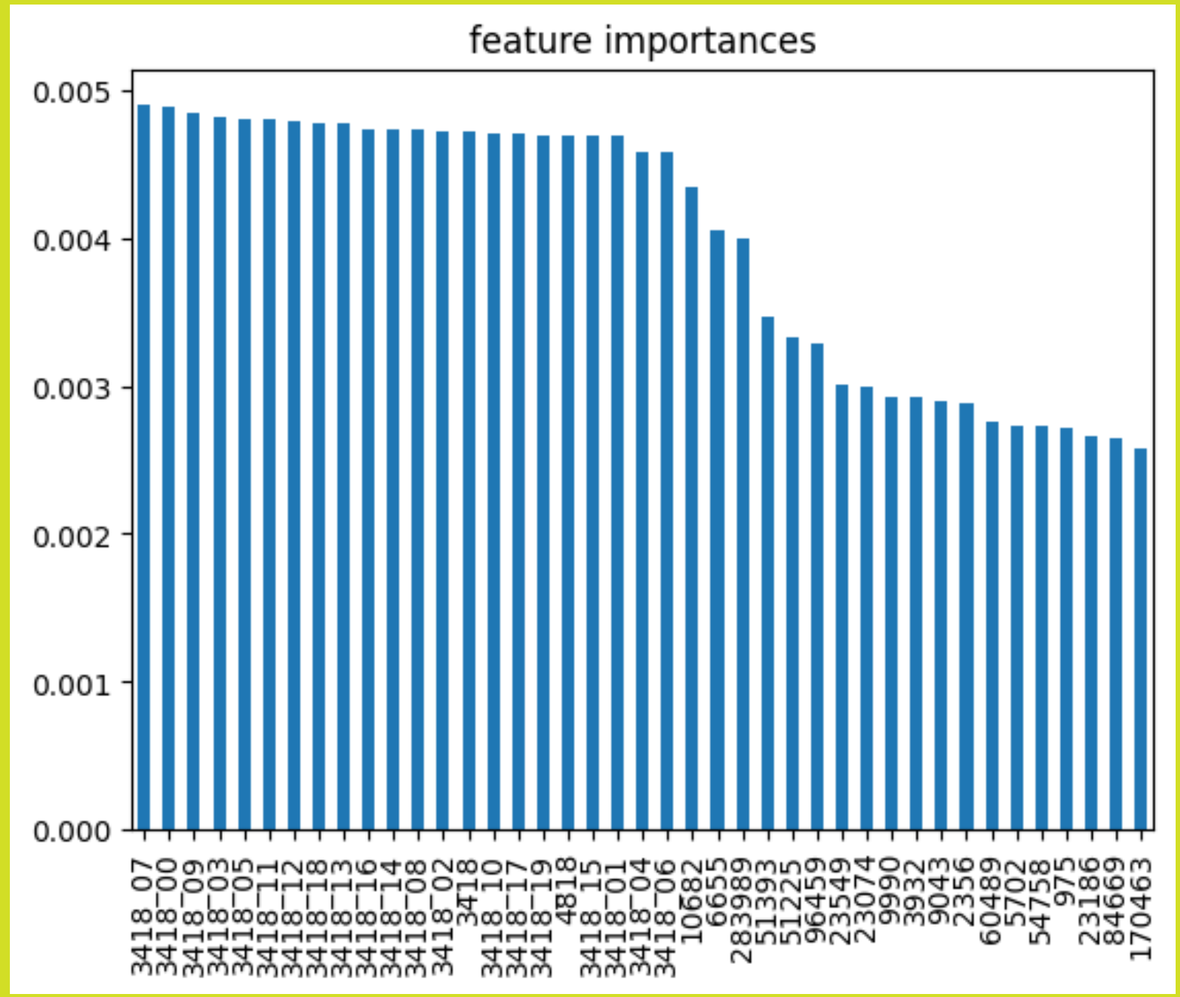
551 rows × 12 columns

Cancer label (target variable)

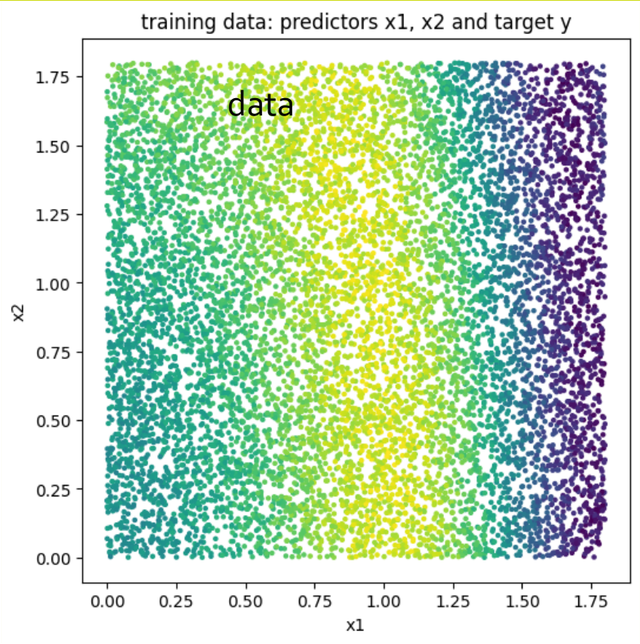
1	normal
2	normal
3	normal
4	normal
5	normal
	...
547	tumor
548	tumor
549	tumor
550	tumor
551	tumor

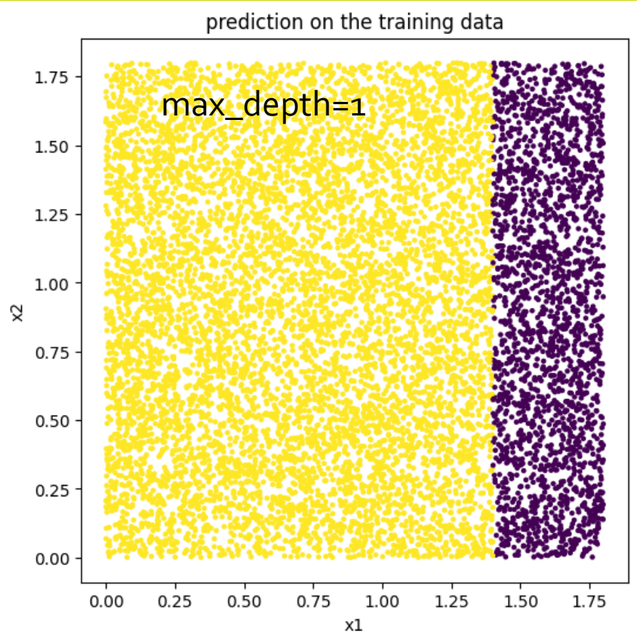
Methods for Feature Importance

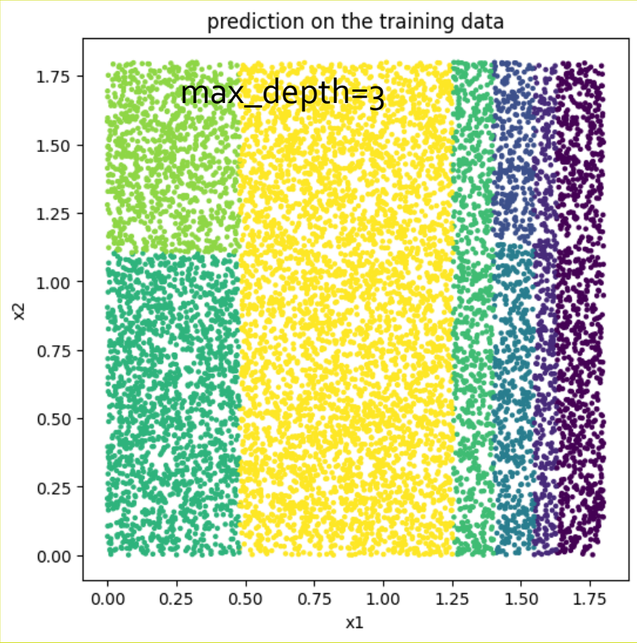
- Gradient Boosting
- Random Forest (Gini Importance)
- Permutation Importance
- Shapley Values (SHAP)
- Conditional Importance

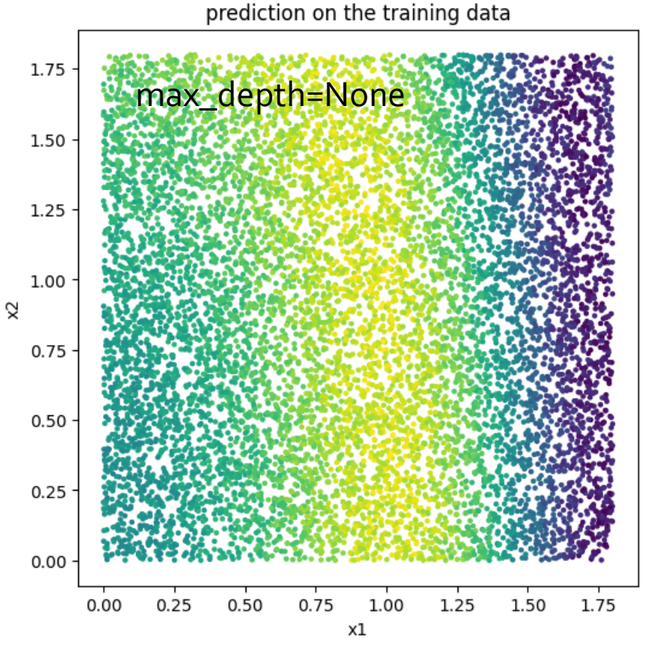


Decision tree



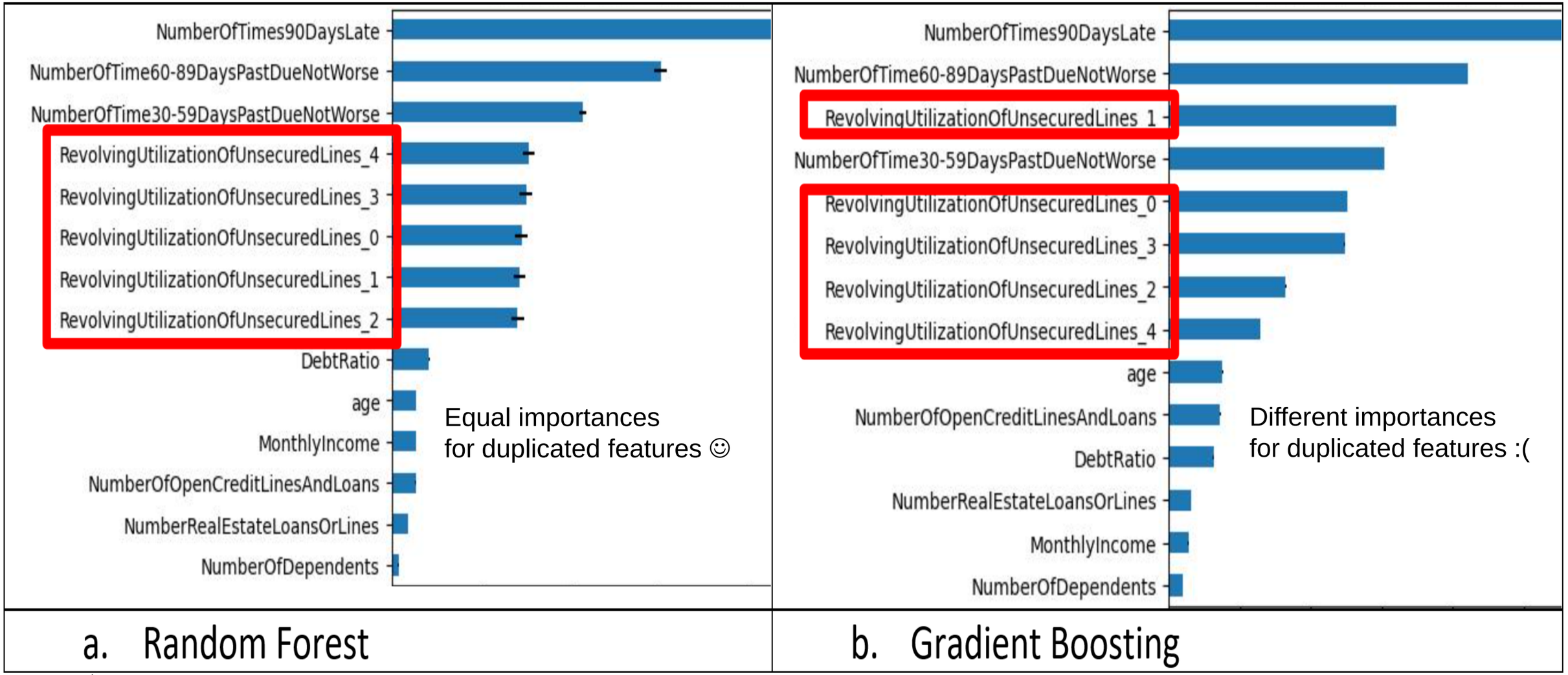




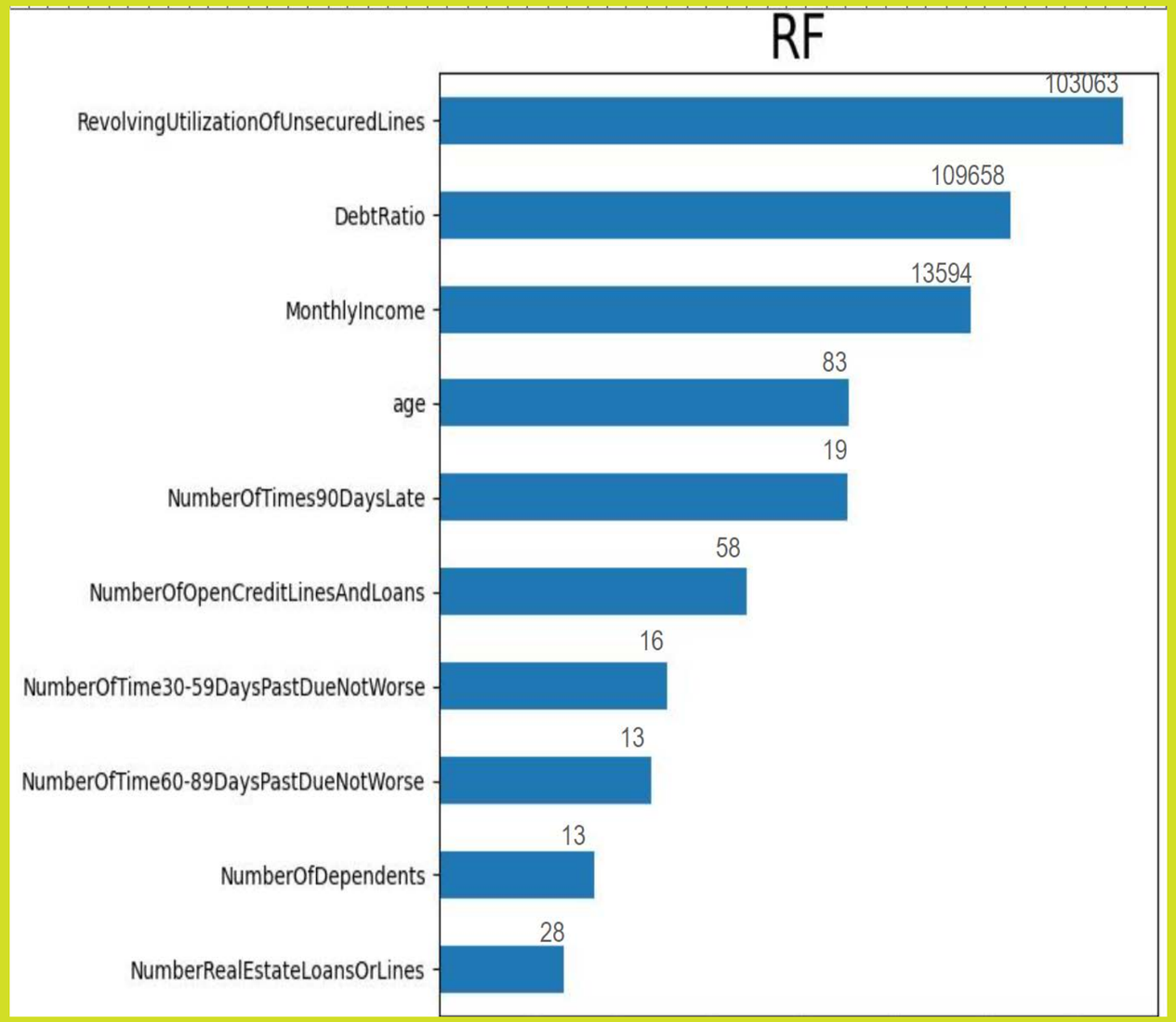


Skoltech

Random Forest - stable!
Gradient Boosting – No!



Gini Bias in Random Forest Feature Importance



Importance depends on feature cardinality!

Binary features – underestimated …

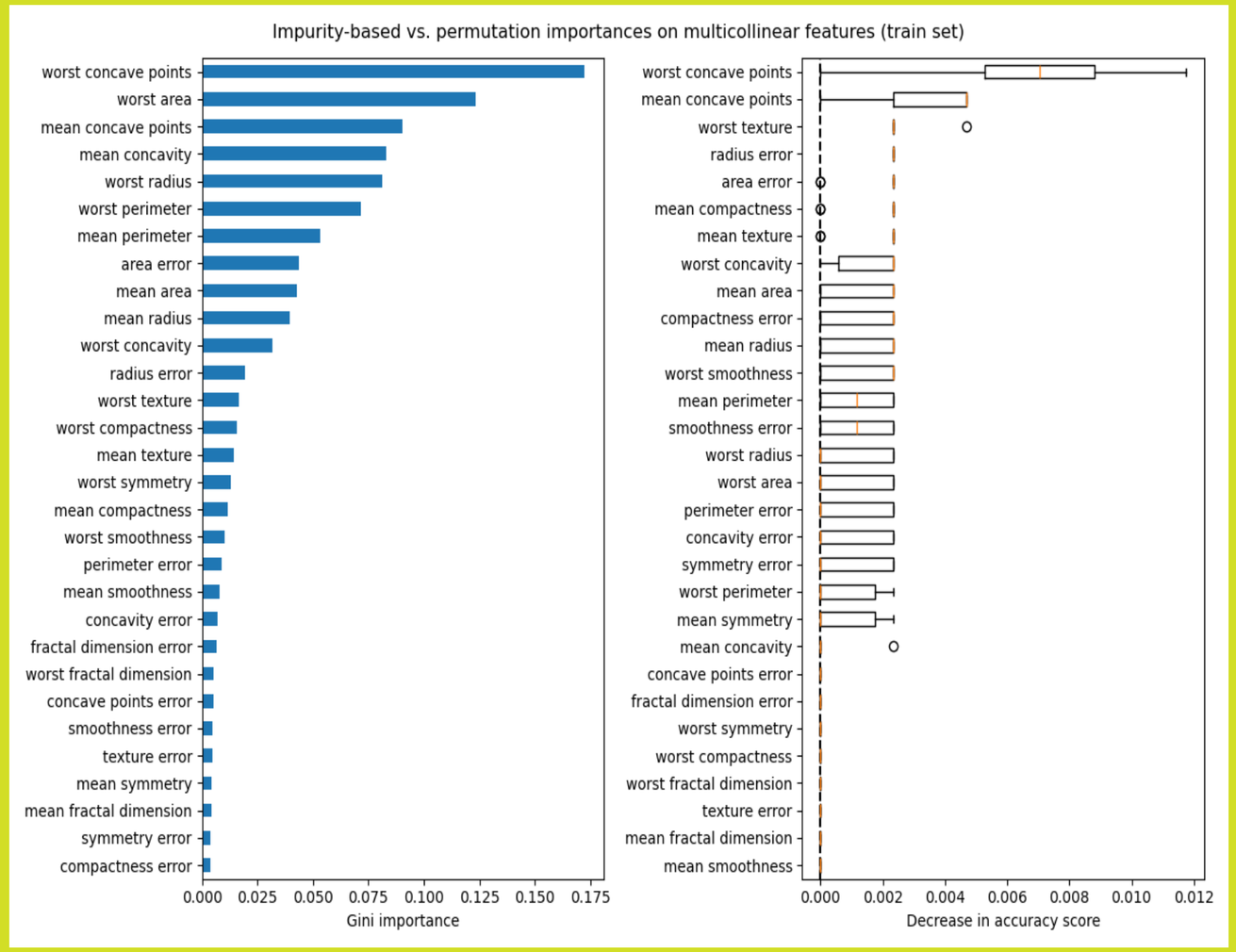
Continuous features – overestimated…

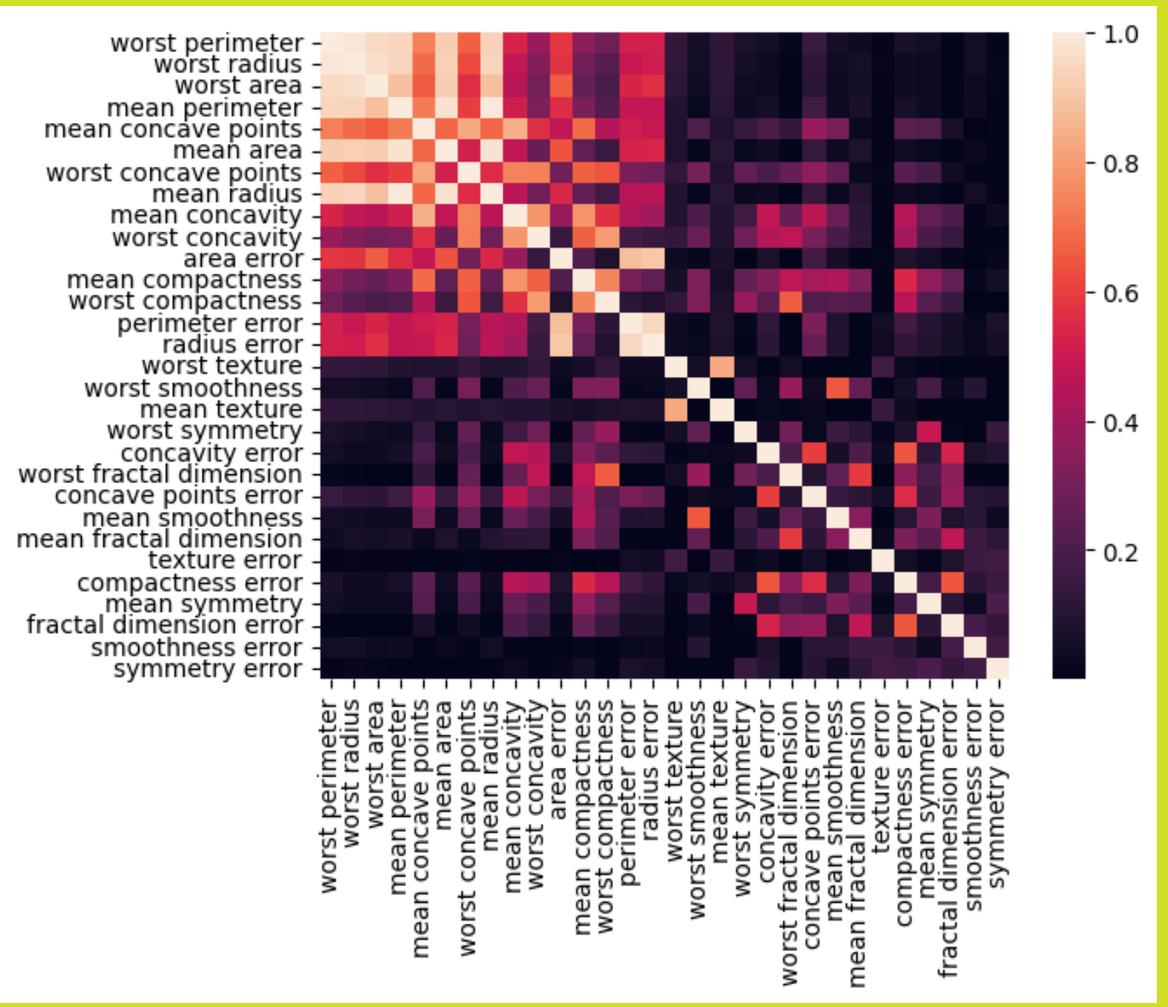
Correlation Bias (Multicollinearity)

In Random Forest Gini Importance:

In Permutation Importance, SHAP:

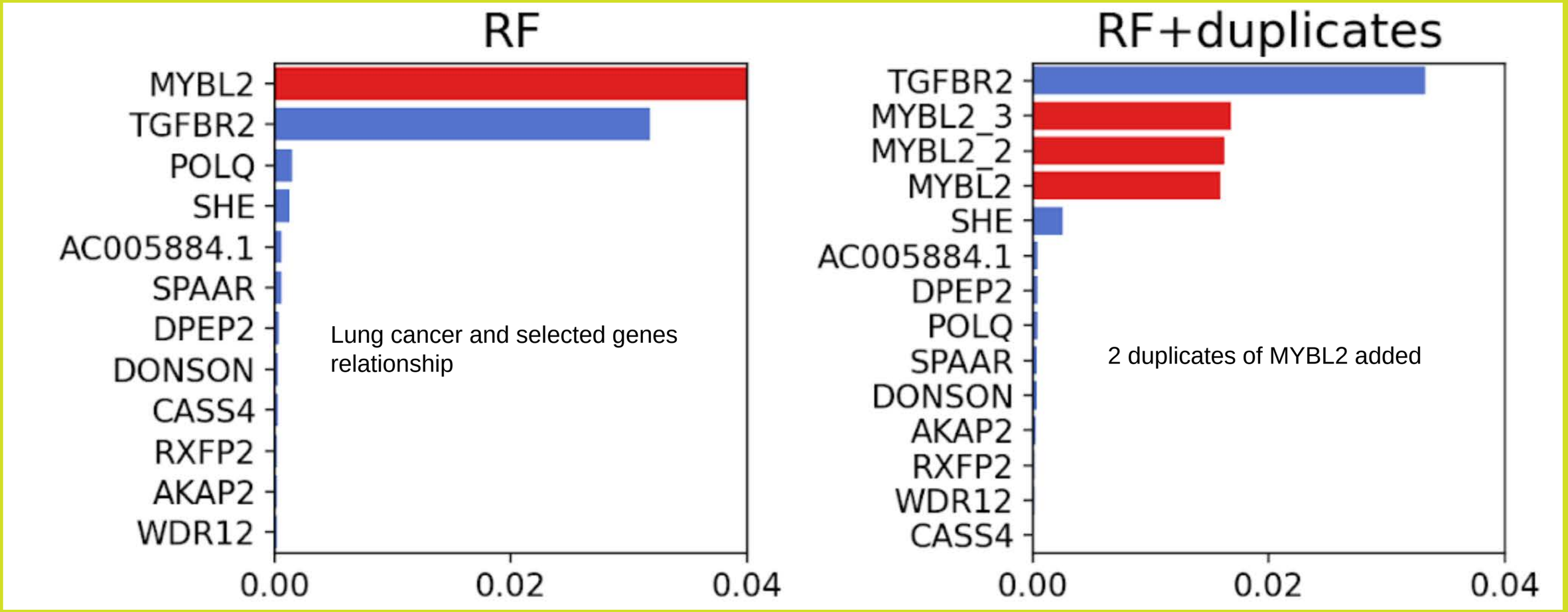
Breast cancer example in python:
https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance_multicollinear



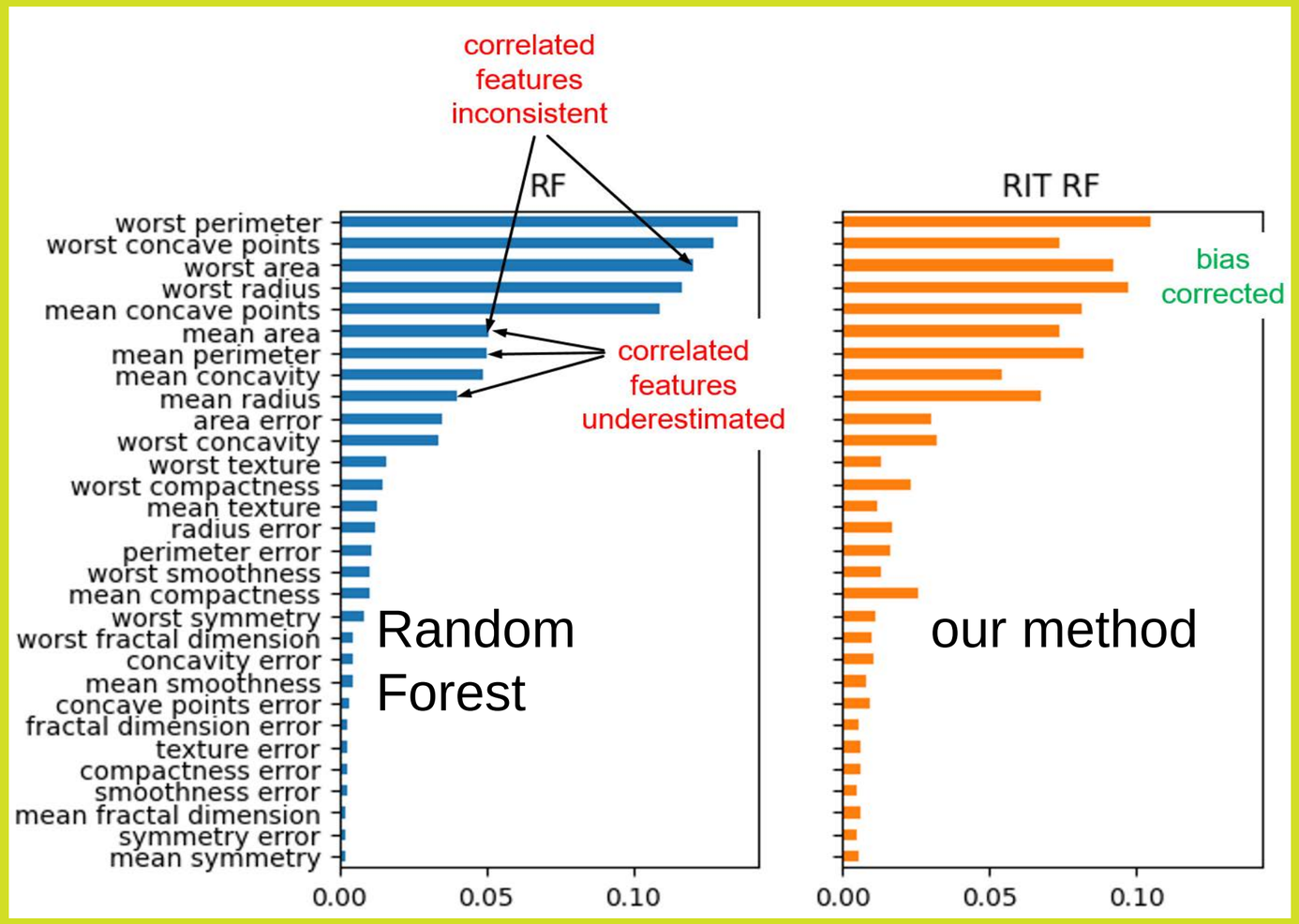


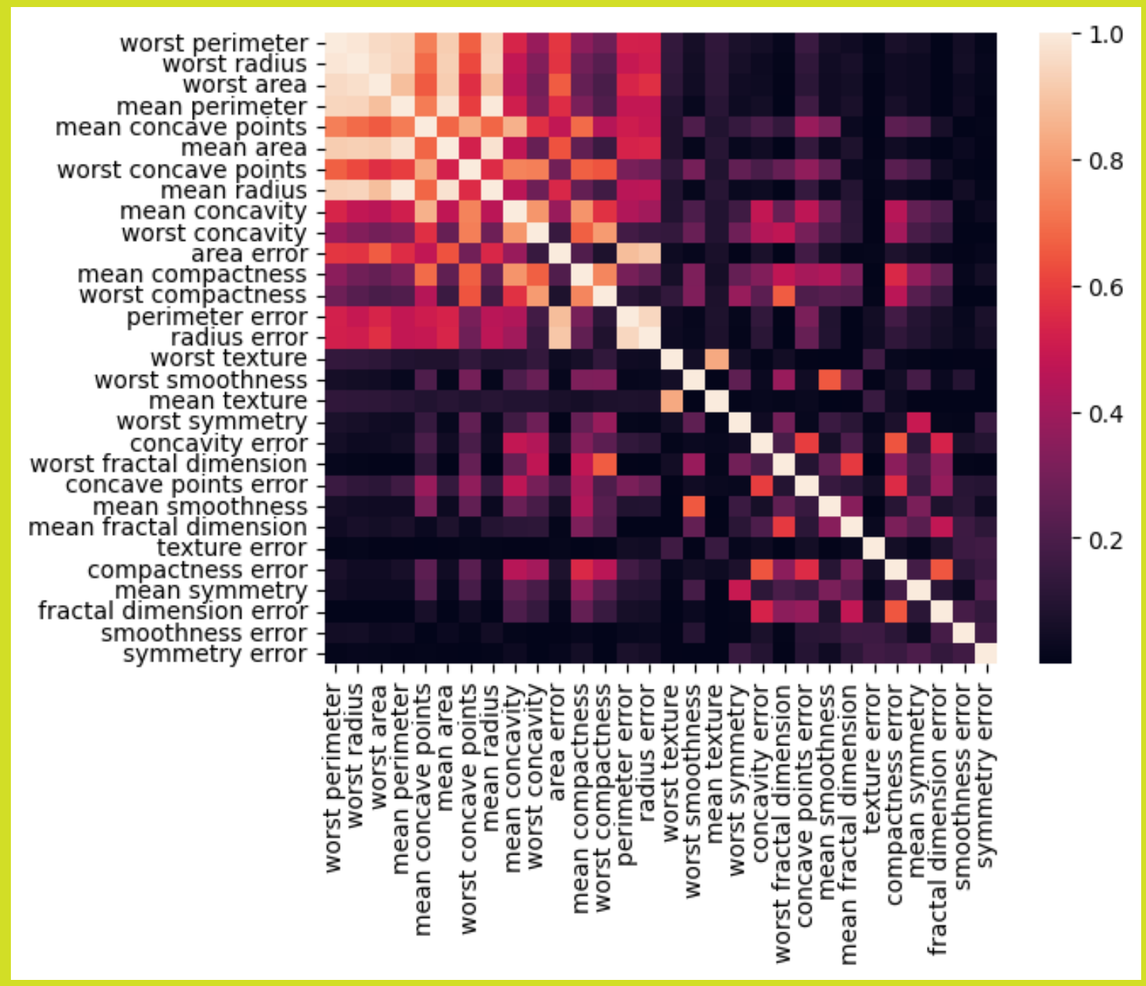
Correlation Bias (Multicollinearity) Problem

In Random Forest Gini Importance:



Correlation Bias Correction In Random Forest



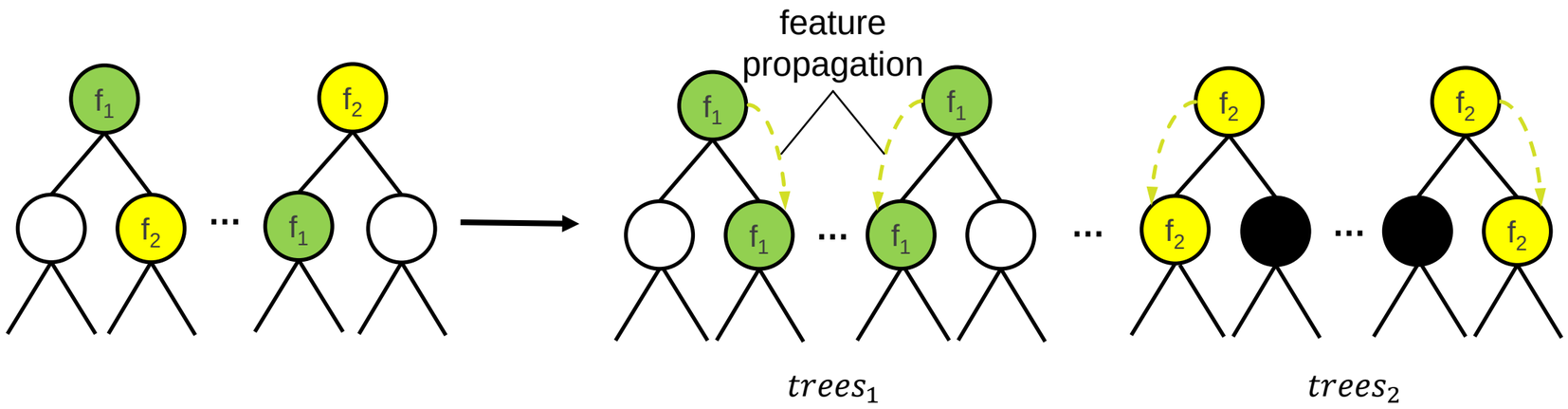


Skoltech

Correlation Bias Correction In Random Forest (method's idea)

Random Forest

Modified Random Forest:
Redundancy Insensitive (Conservative) Trees

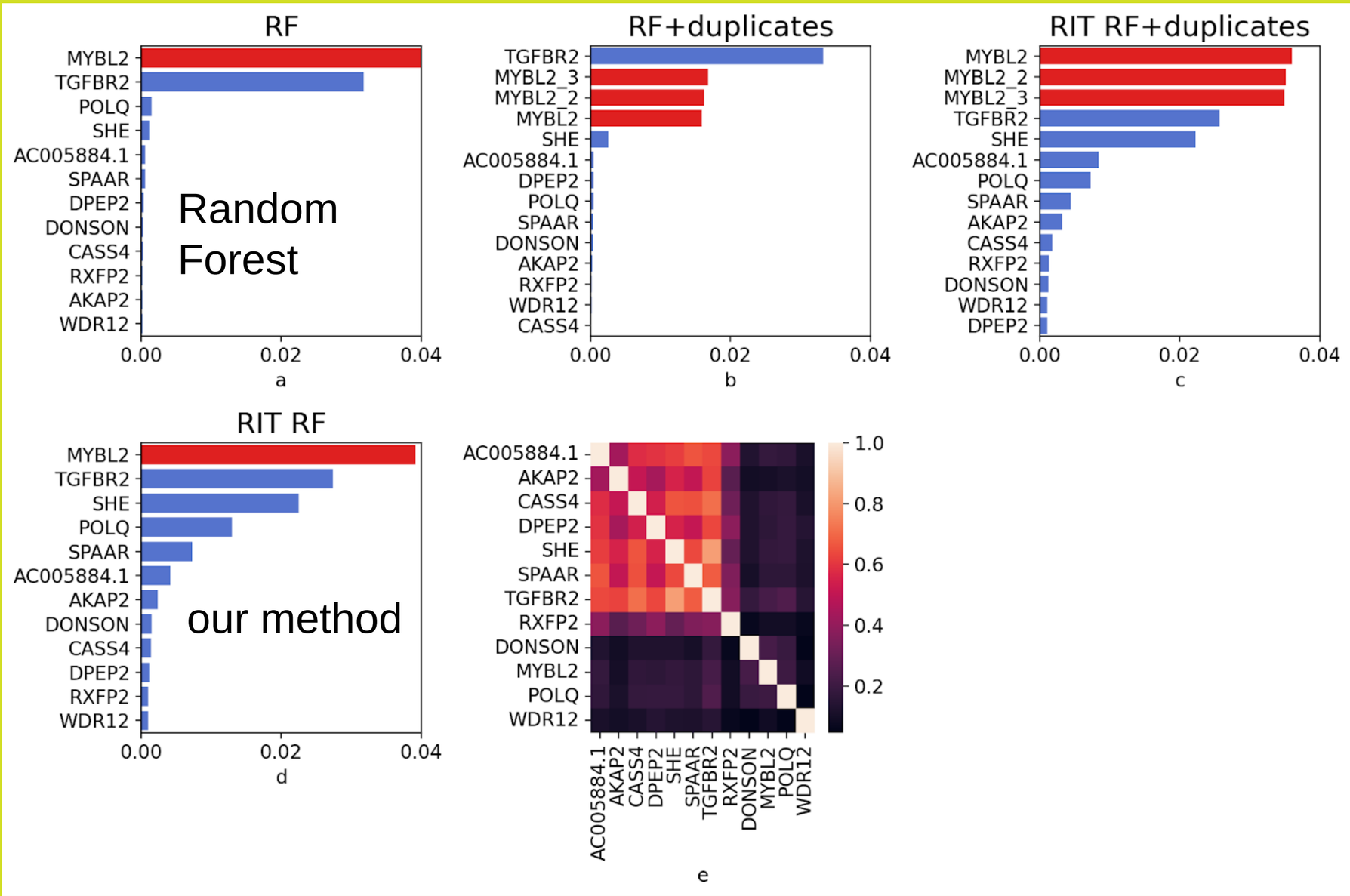


$$\text{Importance}_{f_i} = \frac{1}{\#trees} \sum_{t \in trees} gain_t\,(f_i)$$

$$\text{Importance}_{f_i} = \frac{1}{\#trees_i} \sum_{t \in trees_i} gain_t\,(f_i)$$

Skoltech

Correlation Bias Correction In RF



Skoltech

Skoltech