## **Laborator 9**

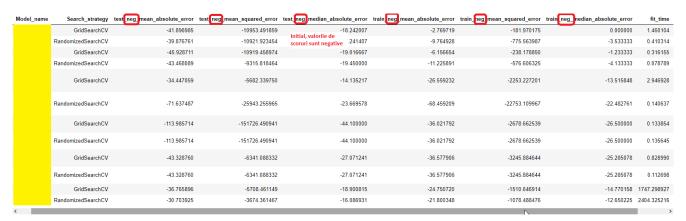
## Modele de regresie

Folositi urmatoarele seturi de date:

- CPU Computer Hardware (https://archive.ics.uci.edu/ml/datasets/Computer+Hardware); excludeti din dataset coloanele: vendor name, model name, estimated relative performance; se va estima coloana "published relative performance".
- 2. Boston Housing (http://archive.ics.uci.edu/ml/machine-learning-databases/housing/)
- 3. <u>Wisconsin Breast Cancer (http://www.dcc.fc.up.pt/~ltorgo/Regression/DataSets.html)</u>; cautati in panelul din stanga Wisconsin Breast Cancer si urmati pasii din "My personal Notes"
- Communities and Crime (http://archive.ics.uci.edu/ml/datasets/communities+and+crime); stergeti primele 5
  dimensiuni si trasaturile cu missing values.

Pentru fiecare set de date aplicati minim 5 modele de regresie din scikit learn. Pentru fiecare raportati: mean absolute error, mean squared error, median absolute error - a se vedea <a href="sklearn.metrics">sklearn.metrics</a> (<a href="http://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics">http://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics</a>) - folosind 5 fold cross validation. Valorile hiperparametrilor trebuie cautate cu grid search (cv=3) si random search (n\_iter dat de voi). Metrica folosita pentru cautarea hiperparametrilor va fi mean squared error. Raportati mediile rezultatelor atat pentru fold-urile de antrenare, cat si pentru cele de testare; indicatie: puteti folosi metoda cross\_validate cu parametrul return\_train\_score=True">https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics</a>) - folosind 5 fold cross validation. Valorile hiperparametrilor trebuie cautate cu grid search (cv=3) si random search (n\_iter dat de voi). Metrica folosita pentru cautarea hiperparametrilor va fi mean squared error. Raportati mediile rezultatelor atat pentru fold-urile de antrenare, cat si pentru cele de testare; indicatie: puteti folosi metoda cross\_validate cu parametrul return\_train\_score=True , iar ca model un obiect de tip GridSearchCV sau RandomizedSearchCV .

Rezultatele vor fi trecute intr-un dataframe. Intr-o stare intermediara, valorile vor fi calculate cu semnul minus: din motive de implementare, biblioteca sklearn transforma scorurile in numere negative; a se vedea imaginea de mai jos:



Valorile vor fi aduse la interval pozitiv, apoi vor fi marcate cele maxime si minime; orientativ, se poate folosi imaginea de mai jos, reprezentand dataframe afisat in notebook; puteti folosi alte variante de styling pe dataframe precum la <a href="https://pandas.pydata.org/pandas-docs/stable/user\_guide/style.html#">https://pandas.pydata.org/pandas.pydata.org/pandas-docs/stable/user\_guide/style.html#</a>).

Se va crea un raport final in format HTML sau PDF - fisier(e) separat(e). Raportul trebuie sa contina minimal: numele setului de date si obiectul dataframe; preferabil sa se pastreze marcajul de culori realizat in notebook.

	Model_name	Search_strategy	test_mean_absolute_error	test_mean_squar	red_error	$test\_median\_absolute\_error$	$train\_mean\_absolute\_error$	train_mean_squared_error	$train\_median\_absolute\_error$	fit_time	score_time
0		GridSearchCV	41.899	Scorurile de	10953.5	18.242	2.76972	181.97	-0	1.4601	0.00239739
1		RandomizedSearchCV	39.8768	eroare sunt	10921.9	17.2414	9.76493	775.564	3.53333	0.410314	0.0027926
2		GridSearchCV	45.9287	aduse in interval pozitiv	10919.5	19.0167	6.15665	238.179	1.23333	0.316155	0.000797749
3		RandomizedSearchCV	43.4681	meer an poerer	9315.82	19.45	11.2259	576.606	4.13333	0.0787893	0.00119648
4		GridSearchCV	34.4471		5682.34	14.1352	26.5592	2253.23	13.5158	2.94693	0.000997543
5		RandomizedSearchCV	71.6375		25943.3	23.6696	68.4592	22753.1	22.4828	0.140637	0.00158916
6		GridSearchCV	113.986		151726	44.1	36.0218	2678.66	26.5	0.133854	0.00139489
7		RandomizedSearchCV	113.986		151726	44.1	36.0218	2678.66	26.5	0.135645	0.0017952
8		GridSearchCV	43.3288		6341.09	27.0712	36.5779	3245.88	25.2851	0.82899	0.000802088
9		RandomizedSearchCV	43.3288		6341.09	27.0712	36.5779	3245.88	25.2851	0.112698	0.000998211
10		GridSearchCV	36.7659		5708.46	18.9008	24.7507	1510.65	14.7702	1747.3	0.00159583
11		RandomizedSearchCV	30.7039		3674.36	16.0869	21.8003	1078.49	12.6502	2404.33	0.00119677

## Notare:

- 1. Se acorda 20 de puncte din oficiu.
- 2. Optimizare si cuantificare de performanta a modelelor: 3 puncte pentru fiecare combinatie set de date + model = 60 de puncte
- 3 Documentare modele: numar modele \* 2 nuncte = 10 nuncte. Documentati in iunvter notehook fiecare din

Notare: rezolvarea va fi incarcata pe platforma de elearning in saptamana 11-15 mai.