

# 1 Laborator 8

## 1.1 Modele de regresie

Folositi urmatoarele seturi de date:

1. [CPU Computer Hardware](https://archive.ics.uci.edu/ml/datasets/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/) (<http://archive.ics.uci.edu/ml/machine-learning-databases/housing/>)
3. [Wisconsin Breast Cancer](http://www.dcc.fc.up.pt/~ltorgo/Regression/DataSets.html) (<http://www.dcc.fc.up.pt/~ltorgo/Regression/DataSets.html>); cautati  
in panelul din stanga Wisconsin Breast Cancer si urmati pasii din "My personal Notes"
4. [Communities and Crime](http://archive.ics.uci.edu/ml/datasets/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 [sklearn.metrics](http://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics) (<http://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics>) - 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:

Model_name	Search_strategy	test_neg_mean_absolute_error	test_neg_mean_squared_error	test_neg_median_absolute_error	train_neg_mean_absolute_error	train_neg_mean_squared_error	train_neg_median_absolute_error	fit_time
	GridSearchCV	-41.898985	-10953.491059	-18.242007	-2.769719	-181.970175	0.000000	1.460104
	RandomizedSearchCV	-39.876761	-10921.923454	-241407	-9.764928	-775.563987	-3.533333	0.410314
	GridSearchCV	-45.928711	-10919.458974	-19.016667	-6.156654	-238.178850	-1.233333	0.316155
	RandomizedSearchCV	-43.468089	-9315.818464	-19.450000	-11.225891	-576.606325	-4.133333	0.078789
	GridSearchCV	-34.447059	-5682.339750	-14.135217	-26.559232	-2253.227201	-13.515848	2.946928
	RandomizedSearchCV	-71.637487	-25943.255965	-23.669578	-68.459209	-22753.109967	-22.482761	0.140637
	GridSearchCV	-113.985714	-151726.490941	-44.100000	-36.021792	-2678.662539	-26.500000	0.133854
	RandomizedSearchCV	-113.985714	-151726.490941	-44.100000	-36.021792	-2678.662539	-26.500000	0.135645
	GridSearchCV	-43.328760	-6341.088332	-27.071241	-36.577906	-3245.884644	-25.285878	0.828990
	RandomizedSearchCV	-43.328760	-6341.088332	-27.071241	-36.577906	-3245.884644	-25.285878	0.112698
	GridSearchCV	-36.765896	-5708.461149	-18.900815	-24.750720	-1510.646914	-14.770158	1747.298927
	RandomizedSearchCV	-30.703925	-3674.361467	-16.086931	-21.800348	-1078.488476	-12.650225	2404.325216

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 [https://pandas.pydata.org/pandas-docs/stable/user\\_guide/style.html#](https://pandas.pydata.org/pandas-docs/stable/user_guide/style.html#) ([https://pandas.pydata.org/pandas-docs/stable/user\\_guide/style.html#](https://pandas.pydata.org/pandas-docs/stable/user_guide/style.html#)).

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_squared_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	10953.5	18.242	2.76972	181.97	-0	1.4601	0.00239739
1		RandomizedSearchCV	39.8768	10921.9	17.2414	9.76493	775.564	3.53333	0.410314	0.0027925
2		GridSearchCV	45.9287	10919.5	19.0167	6.15665	238.179	1.23333	0.316155	0.000797749
3		RandomizedSearchCV	43.4681	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	113.988	151726	44.1	36.0218	2678.66	26.5	0.133854	0.00139489
7		RandomizedSearchCV	113.988	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 puncte = 10 puncte. Documentati in jupyter notebook fiecare din modelele folosite, in limba romana. Puteti face o sectiune separata cu documentarea algoritmilor. Fiecare model trebuie sa aiba o descriere de minim 20 de randuri, minim o imagine asociata si minim 2 referinte bibliografice.
4. 10 puncte: export in format HTML sau PDF.

**Notare:** laboratorul va fi salvat in repository-ul de github si prezentat in saptamana 6-10 mai.