

Laborator 8

Modele de clasificare

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
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 panleul 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.
5. (numar de modele *seturi de date* 2 = 40 de puncte) Pentru fiecare set de date aplicati 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. Raportati mediile rezultatelor atat pentru fold-urile de antrenare, cat si pentru cele de testare. Revedeti formele ulterioare ale acestui document pentru precizari despre: continutul rezultatelor raportate, modalitate de notare.
6. (numar modele * 2 puncte = 10 puncte) Documentati in jupyter notebook fiecare din modelele folosite, in limba romana. Puteti face o sectiune o sectiune separata cu documentarea algoritmilor + trimitere la algoritm.
7. (numar modele *numar de seturi de date* 2 = 40 de puncte) Pentru fiecare model: efectuati o cautare a hiperparametrilor optimi folosind grid search si random search.

Se acorda 20 de puncte din oficiu.

Pentru fiecare set de date raportati rezultatele obtinute de fiecare model.

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