1 Laborator 8

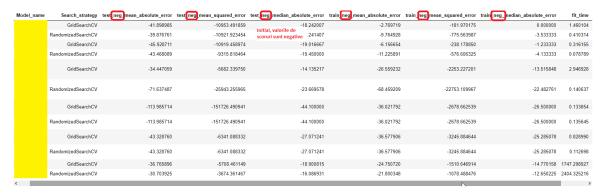
1.1 Modele de regresie

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/)
- Wisconsin Breast Cancer (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. <u>Communities and Crime (http://archive.ics.uci.edu/ml/datasets/communities+and+crime)</u>; 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) - 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 optimizarea 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 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.



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.

Precizari:

- 1. Depunerea se face in Tema 5 pe elearning, pana cel tarziu 29 aprilie ora 23.
- 2. Specificatiile legate de depunere, numele fisierelor etc. sunt aceleasi ca la tema 4.