1 Laborator 6

1.1 Modele de clasificare

Folositi 4 seturi de date pentru probleme de clasificare, plecand de la repository-urile specificate in Cursul 6. Cel putin un set de date sa fie cu valori lipsa; pentru un alt set de date care are initial toate valorile, introduceti dvs. in mod artificial valori lipsa, suprascriind un anumit procent din valorile initiale (ex. `p=5%`, `p` parametru) cu `numpy.nan`.

1. (20 puncte) Aplicati o metoda de missing value imputation, unde este cazul; documentati metoda folosita.

Resurse: Pentru missing value imputation, puteti urmari Imputation of missing values (https://scikit-learn.org/stable/modules/impute.html), How to Handle Missing Data with Python (https://machinelearningmastery.com/handle-missing-data-python/), fancyimpute (https://github.com/iskandr/fancyimpute), missingpy (https://github.com/epsilon-machine/missingpy).

Cerinta: In cazul in care folositi un pachet ce trebuie instalat (nu face parte din distributia standard anaconda), includeti intr-o celula o comanda de instalare corespunzatoare folosind semn de exclamare, de exemplu:

```
!pip install missingpy
```

(sursa: https://github.com/epsilon-machine/missingpy (https://github.com/epsilon-machine/missingpy)). La executia celulei in Jupyter Notebook se instaleaza pachetul, iar in celulele ulterioare importurile din noul pachet functioneaza.

2. (numar de modele * numar de seturi de date * 1 punct = 20 de puncte) Pentru fiecare set de date aplicati 5 modele de clasificare din scikit learn. Pentru fiecare raportati: acuratete, precision, recall, scorul F1 - a se vedea sklearn.metrics (http://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics), Precision and recall (https://en.wikipedia.org/wiki/Precision_and_recall) - folosind 5 fold cross validation. Raportati mediile rezultatelor atat pentru fold-urile de antrenare, cat si pentru cele de testare. Rularile se vor face cu valori fixate ale hiperparametrilor.
3. (numar de modele * numar de seturi de date * 1 punct = 20 de puncte) Raportati performanta fiecarui model, folosind 5 fold cross validation. Pentru fiecare din cele 5 rulari, cautati hiperparametrii optimi folosind 4-fold cross validation. Performanta modelului va fi raportata ca medie a celor 5 rulari. *Observatie:* la fiecare din cele 5 rulari, hiperparametrii optimi pot diferi, din cauza datelor utilizate pentru antrenare/validare.
4. (numar modele * 4 puncte = 20 puncte) Documentati in jupyter notebook fiecare din modelele folosite, in limba romana. Daca acelasi algoritm e folosit pentru mai multe seturi de date, puteti face o sectiune separata cu documentarea algoritmilor + trimitere la algoritm.

Se acorda 20 de puncte din oficiu.

Exemple de modele de clasificare:

1. Multi-layer Perceptron classifier (https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html#sklearn.neural_
2. KNN (https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html#sklearn.neighl
3. SVM (https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC)
4. Gaussian processes (https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.GaussianProcessClassifier.html#
5. RBF (https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.kernels.RBF.html#sklearn.gaussi
6. Decision tree (https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.Decisi
7. Random forest (https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ens
8. Gaussian Naive Bayes (https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html#sklearn.naive_baye

Predare:

1. Predarea se face cel tarziu in 24 aprilie 2021 ora 23, in lucrarea de pe elearning (Tema 4) + repo propriu de pe github. Fiecare student va face upload in tema, altfel va fi depunctat cu 40 de puncte.
2. La prezentarea temei coechipierii trebuie sa fie prezenti. Denumirea fisierului predat trebuie sa respecte conventia de la tema 2.
3. Obligatoriu: type annotations pentru variabile, parametri, tip de retur; docstrings.
4. Fisierele de date folosite vor fi descarcate local de studenti si puse intr-un director "data". Se va realiza o arhiva zip care contine minim: fiserul/fisierele ipynb si direcotrul de date. Suplimentar, pot fi folosite imagini incluse in ipynb; acestea vor fi puse in directorul "images" ce se va include in arhiva zip predata.