

1 Laborator 12

Folosind un set de date - de exemplu de la <https://archive.ics.uci.edu/ml/datasets.php?format=&task=&att=&area=&numAtt=&numIns=&type=text&sort=taskDown&view=table> (<https://archive.ics.uci.edu/ml/datasets.php?format=&task=&att=&area=&numAtt=&numIns=&type=text&sort=taskDown&view=table>) - sa se rezolve o problema de clasificare sau regresie, plecand de la intrari de tip text.

Se poate opta pentru codificare BOW, n-grams, word2vec sau altele adecvate. Modelele de predictie pot fi din biblioteca scikit-learn. Puteti folosi pentru preprocesare biblioteca [NLTK](https://www.nltk.org) (<https://www.nltk.org>).

Pentru clasificare se va optimiza scorul F1; se vor raporta scorurile F1 si acuratetea. Pentru regresie se va optimiza scorul mean squared error; se vor raporta scorurile MSE, mean absolute error, r2.

Exemple:

1. [Clasificare de SMS-uri](https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection) (<https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection>)
2. [Sentence Classification Data Set](https://archive.ics.uci.edu/ml/datasets/Sentence+Classification) (<https://archive.ics.uci.edu/ml/datasets/Sentence+Classification>)
3. [Sentiment Labelled Sentences Data Set](https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences) (<https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences>)
4. [Victorian Era Authorship Attribution Data Set](https://archive.ics.uci.edu/ml/datasets/Victorian+Era+Authorship+Attribution) (<https://archive.ics.uci.edu/ml/datasets/Victorian+Era+Authorship+Attribution>)
5. [Amazon Commerce reviews set Data Set](https://archive.ics.uci.edu/ml/datasets/Amazon+Commerce+reviews+set) (<https://archive.ics.uci.edu/ml/datasets/Amazon+Commerce+reviews+set>)
6. [Farm Ads Data Set](https://archive.ics.uci.edu/ml/datasets/Farm+Ads) (<https://archive.ics.uci.edu/ml/datasets/Farm+Ads>)

Se vor investiga minim 2 seturi de date si pentru fiecare din ele minim 4 modele de clasificare sau de regresie. Daca setul de date e deja impartit in train si test, se va folosi ca atare - setul de antrenare se va imparti, suplimentar in train + validation; altfel, se va face 5 fold CV. Valorile optime ale hiperparametrilor vor fi alese prin random search sau grid search.

Pentru fiecare set de date:

1. (2 x 0.5 p) Se descrie setul de date, in limba romana (continut, provenienta, problema etc.)
2. (2 x 1 p) Se face analiza exploratorie, folosind cod Python: distributia claselor sau a valorilor continue de iesire - numeric si grafic, statistici asupra textelor (de exemplu: lungime minima/medie/maxima; cele mai frecvente k cuvinte; clustering etc.). Se va explica fiecare pas si ce se urmareste prin efectuarea lui. Graficele vor avea axele numite (ce se reprezinta, eventual unitate de masura)
3. (2 x 0.5 p) Se face preprocesare de date; se explica in limba romana care sunt metodele de preprocesare folosite, efectul lor pe datele de intrare, ce forma are iesirea obtinuta; se arata efectele pasilor de preprocesare asupra setului de date (noul numar de documente, dinamica vocabularului, trasaturile rezultate etc.) Se pot aduga grafice si tabele la acest pas.
4. (2 x 4 x 0.5 p) Clasificare sau regresie, dupa caz: se face o descriere a modelelor considerate, in limba romana; se descrie modalitatea de cautare a hiperparametrilor; rezultatele obtinute se

vor prezenta tabelar, similar cu tema precedenta.

Se acorda doua puncte din oficiu.

Descrierea modelelor si a pasilor de preprocesare pot fi in sectiuni separate, cu referinte la acestea unde e necesar.

Exemple:

1. [Working With Text Data \(https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html\)](https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html)
2. [Text Classification with Python and Scikit-Learn \(https://stackabuse.com/text-classification-with-python-and-scikit-learn/\)](https://stackabuse.com/text-classification-with-python-and-scikit-learn/)
3. [How to Prepare Text Data for Machine Learning with scikit-learn \(https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/\)](https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/)

Predarea temei se va face prin platforma de elearning in saptamana 24-28 mai.

Precizari:

1. Depunerea se face in Tema 6 pe elearning, pana cel tarziu 28 mai ora 23. Prezentarea lor se va face la o data ce va fi precizata ulterior.
2. Specificatiile legate de depunere, numele fisierelor etc. sunt aceleasi ca la tema 4.