Universitatea „Ovidius” din Constanța

Facultatea de Matematică și Informatică

Specializarea: Medii virtuale multi-modale distribuite

Profesor: Conf. dr. Pelican Elena

Student: Andronache Diana

Proiect Machine Learning

# Problema aleasă

## Dataset

Am utilizat un dataset cu **5856 de imagini** ce conțin radiografii la nivelul pieptului (plămânilor).Data setul este unul dezechilibrat. Din totalul pozelor, 1583 aparțin unor oameni normali (sănătoși), iar restul de 4273 prezintă afecțiuni pulmonare.

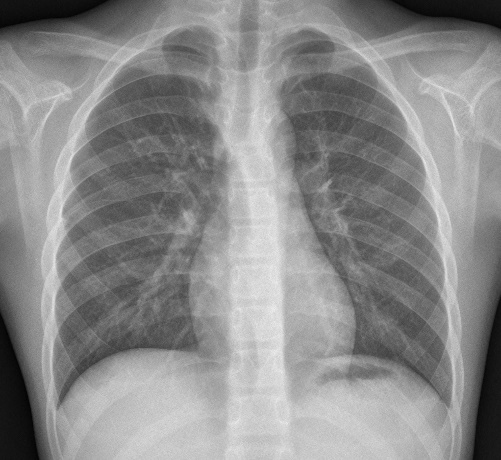
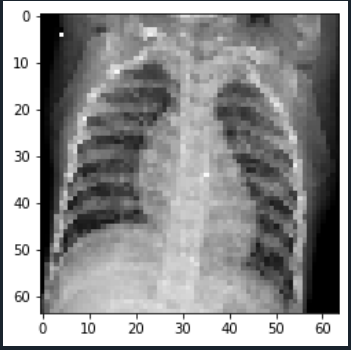
Sursa datasetului: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

## Sopul

Problema aleasă este de **clasificare** (**supervizată**) a pozelor în funcție de prezența sau lipsa afecțiunii pulmonare (pneumonie).

Am împărțit datasetul în raportul de 80% training -20% testing pentru toate metodele aplicate.

Pentru a optimiza timpul de execuție și pentru a scala toate pozele la aceeași dimensiune (pozele inițiale aveau rezoluții care difereau de la o poză la alta), am dat la toate pozele resize la dimensiunile (64, 64). Iar, pentru a normaliza pozele, am făcut trecerea de la valori cuprinse în intervalul [0; 255] la unele cuprinse în intervalul [0; 1].

# Metode implementate & rezultate obținute

## Regresie Logistică

Precizia este mai mare pentru cele care au pneumonie (probabil datorită numărului mai mare de imagini pentru cei care sunt afectați de boala, față de cei care nu sunt).

Am facut media rezultatelor a 13 rulări:

* Accuracy = 0.94
* run time = 50.7 sec
* precision = 0.92
* recall = 0.91
* f1-score = 0.92

## SVM (Support Vector Machine)

### Cu kernel polynomial (de gradul 3)

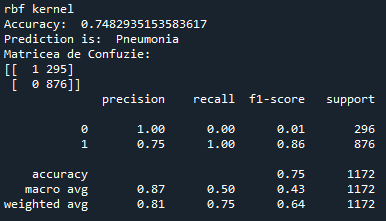
Ca la regresia logistică, precizia este mai bună pentru cele care au pneumonie .

Am facut media rezultatelor pentru 10 rulări:

* Accuracy = 0.94
* run time 60 sec
* precision = 0.92
* recall = 0.92
* f1-score = 0.92

### Cu kernel radial (rbf =radial based function)

Precizia este mai mare pentru cei sănatoși, iar recall-ul și f1-score a rezultat foarte mica (aproape de 0) la cei sănătoși și foarte mare ( aproape de 1) la cei bolnavi. De aici rezultă că rata de FP este mare și că algoritmul tinde să clasifice toate imaginile în categoria de bolnavi de pneumonie.



Am facut media rezultatelor pentru 10 rulări:

* Accuracy = 0.73
* run time 101 sec
* precision = 0.76
* recall = 0.5
* f1-score = 0.42

## Random Forest

Am facut media rezultatelor pentru 10 rulări:

* Accuracy = 0.94
* run time = 80.3 sec
* precision = 0.94
* recall = 0.92
* f1-score = 0.93

## Concluzie

Metoda cu SVM având kernel radial nu se potrivește pentru acest dataset. Iar pentru restul metodelor, având în vedere că acuratețea medie obținută este la fel, cea mai bună pentru datasetul ales este regresia logistică datorită timpului de execuție mai mic.

Menționez că rezultatele medii au fost obținute în urma calculelor făcute în documentul Results.xlsx.

# Problema Inițială

Problema aleasă inițial este tot una de **clasificare supervizată**, dar pe un dataset diferit, care este mai mic (**72 inregistrari**). Scopul ei este de a determina starea de depresie a unor pacienți pe baza unor dimensiuni ale creierului deduse pe baza rmi-urilor. Baza de date contine valori numerice corespunzatoare dimensiunilor anumoitor zone ale creierului precum: pericalosală, occipital superioară, occipital-temporală, circular-insular superioară, poli temporali etc. În total sunt **10 criterii** care sunt sunt folosite pentru a determina starea individului.

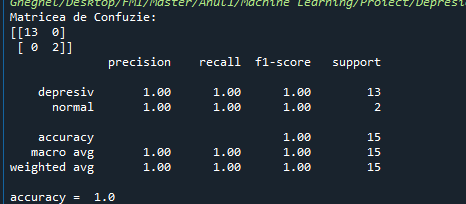
Sursa setului de date: https://brainlife.io/project/5fe56eca57aacdb8482fd198/dataset

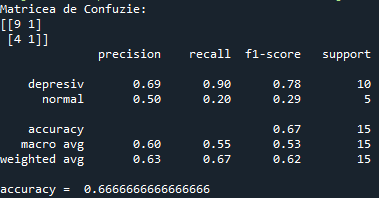
## Metode implementate

### Regresie Logistică

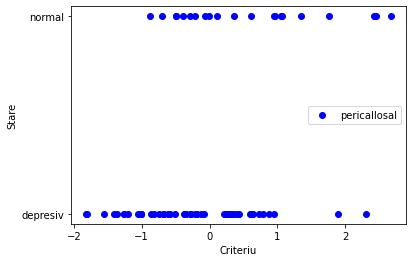
Rezultatele nu sunt concludente datorită numărului mic de date. Totuși, prin comparație, clasificările făcute pentru persoanele care sunt depresive sunt mai precise, deoarece numarul persoanelor depresive (din dataset) este mai mare decât cel al persoanelor normale.

Prin urmare, uneori obținem rezultate bune:

, iar alteori obținem rezultate foarte slabe (deși este diferență doar de maxim 5 clasificări greșite):



Aceste rezultate inconcludente se pot datora și valorilor aberante prezente in dataset, dupa cum reiese din grafic:



Pentru 30 de rulări am obținut rezultatele (făcând media aritmetică):

* acuratețe: 0.85
* precizie: 0.85
* recall: 0.8
* f1-score: 0.81

### SVM cu Cross Validation (folosind motoda k-fold)

Media rezultatelor pentru 30 de rulări este:

* cross validation result: 0.89
* poly kernel – accuracy: 0.77
* rbf kernel – accuracy: 0.84

## Concluzie

Pentru rezultate relevante este nevoie de un dataset mult mai mare.

# Bibliografie

* <https://scikit-learn.org/stable/index.html>
* Cursuri și laboratoare
* <https://stackoverflow.com/> (pentru erori legate de data type si size)

## Pentru articolul științific prezentat

<http://ilpubs.stanford.edu:8090/778/1/2006-13.pdf>

<http://ilpubs.stanford.edu:8090/778/>