Curs 3 - Tehnologii NoSQL

Procesul de Machine Learning. Probleme de clasificare

Gheorghe Cosmin Silaghi

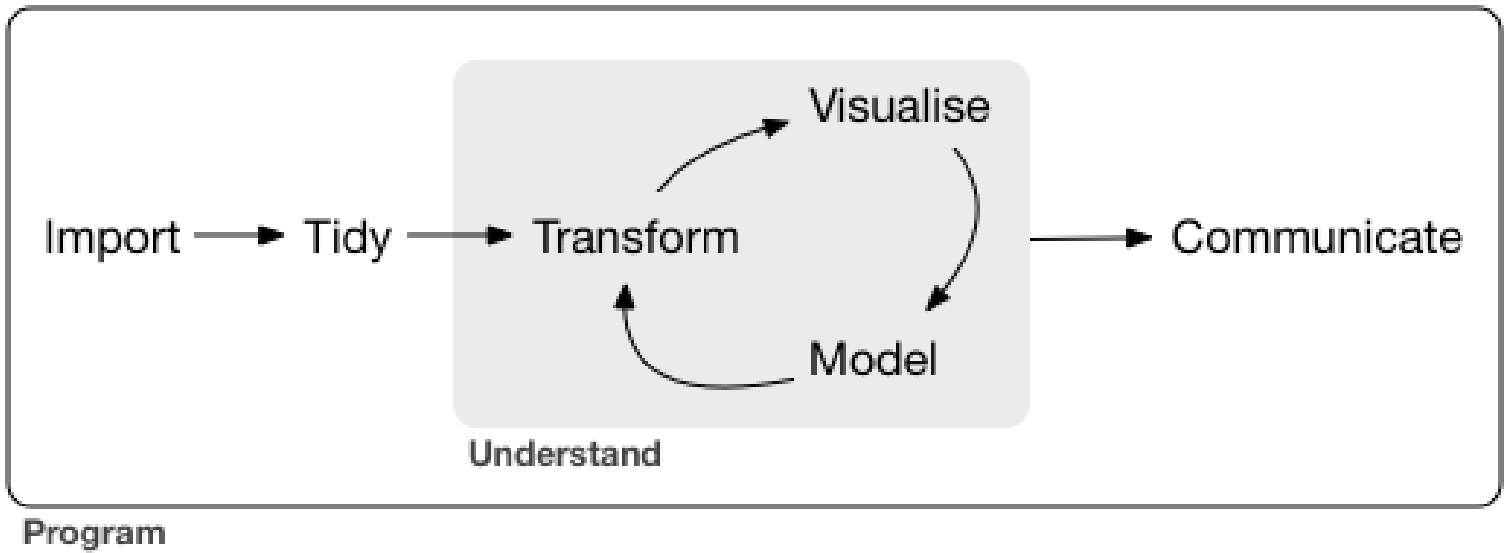
Universitatea Babeș-Bolyai

March 31, 2023

Procesul de Machine Learning (sau *analytics*) (I)

II. Taskuri de baza

- preprocesarea datelor
- invatarea modelului
- analiza si validarea modelului



Obiectivul tinta: **generalizare**

- selectarea unui model capabil de predictie pe date noi, deci capabil de *generalizare*

Predictie - formalizare matematica

Y: variabila de iesire (dependenta) care trebuie prezisa
$X_1, X_2, \ldots, X_p$: p variabilele independente (predictori)

Problema: trebuie sa descoperim o relatie f intre Y si $X = (X_1, X_2, \ldots, X_p)$:

$$Y = f(X) + \epsilon \tag{1}$$

f: o functie necunoscuta, dar fixata, reprezinta *informatia sistematica* pe care X o furnizeaza pentru a prezice pe Y
ϵ: este *termenul de eroare*, aleator, de medie zero

Procesul de Machine Learning (sau *analytics*) (I)

- in conditiile unui set de date disponibil, se vor selecta *metode* potrivite pentru rezolvarea problemei de predictie, adica pentru obtinerea unui model $\hat{f}$ cat mai aproape de functia f reala
- pentru fiecare metoda selectata se vor construi mai multe modele din care se va selecta modelul cel mai bun
- dintre modelele cele mai bune astfel obtinute se va selecta modelul de top, care indica metoda cea mai potrivita pentru rezolvarea problemei studiate

Validare

- *evaluarea modelului*: evaluarea performantei modelului pe date care nu au fost folosite in procesul de antrenare
- *selectarea modelului*: se selecteaza un model cu un nivel acceptat de performanta

Pasi pentru a realiza predictie (numerica sau clasificare)

1. colectarea setului de date si identificarea atributului tinta a predictiei
2. *adnotarea* (labeling) manuala a setului de instante existente
   - fiecare instanta din setul de date va avea o valoare cunoscuta pentru variabila dependenta (tinta)
3. impartirea setului de date in subset de antrenament si subset pentru testare
4. se selecteaza metricile de performanta urmarite si metoda de validare a rezultatelor
5. selectarea mai multor metode de predictie
6. pentru fiecare metoda de predictie selectata la pasul 5
   - se aplica metoda cu un set de parametri si se genereaza un model pentru fiecare set de parametri selectati
   - se evalueaza fiecare model obtinut folosind metoda de validare selectata la pasul 4
   - se obtine cel mai bun model pentru metoda curenta
7. se interpreteaza rezultatele obtinute, se selecteaza metoda care a produs cel mai bun model
8. cu aceasta metoda se creaza un model final aplicand metoda pe intreg setul de antrenament
9. se obtine performanta asteptata a modelului prin aplicarea lui pe setul de testare

Compromisul (trade-off) dintre antrenare si validare

Abordarea cu set de date pentru validare (pasii 4-7

se imparte *setul cu datele disponibile pentru antrenare* in doua parti:

- setul de antrenare: folosit pentru a invata un model
- setul de validare: folosit pentru a calcula valorile indicatorilor de performanta ai modelului

Compromisul (trade-off) dintre antrenare si validare

Abordarea cu set de date pentru validare (pasii 4-7

se imparte *setul cu datele disponibile pentru antrenare* in doua parti:

- setul de antrenare: folosit pentru a invata un model
- setul de validare: folosit pentru a calcula valorile indicatorilor de performanta ai modelului

Obiective ale procesului de invatare a unui model

1. **sa obtinem modele de calidate**: cu cat este mai mare setul de anternare, cu atat mai buna va fi calitatea modelului obtinut
2. **sa avem incredere in modelele obtinute**: cu cat este mai mare setul de validare, cu atat sunt mai de incredere estimarile pentru metricile de performanta

Compromisul (trade-off) dintre antrenare si validare

Abordarea cu set de date pentru validare (pasii 4-7

se imparte *setul cu datele disponibile pentru antrenare* in doua parti:

- setul de antrenare: folosit pentru a invata un model
- setul de validare: folosit pentru a calcula valorile indicatorilor de performanta ai modelului

Obiective ale procesului de invatare a unui model

1. **sa obtinem modele de calidate**: cu cat este mai mare setul de anternare, cu atat mai buna va fi calitatea modelului obtinut
2. **sa avem incredere in modelele obtinute**: cu cat este mai mare setul de validare, cu atat sunt mai de incredere estimarile pentru metricile de performanta

Intrebarea de compromis (trade-off):

- cum vom imparti datele de antrenare in doua seturi (de antrenare si validare), ca sa raspundem intr-un mod cat mai potrivit intrebarilor anterioare?

Obtinerea seturilor de antrenare / validare - metode de esantionare

- sa extragem in mod repetat instante pentru setul de antrenare si sa invatam modelul de interes pentru fiecare esantion obtinut in acest fel, cu scopul de a obtine informatie suplimentara despre modelul rezultat
- in acest fel, obtinem informatii despre variabilitatea modelelor obtinute prin metoda aleasa
- esantionarea este costisitoare dpdv computational, deoarece necesita construirea multor modele, pentru fiecare esantion individual
- acest proces trebuie aplicat doar atunci cand nu sunt destule date disponibile, sau daca dorim sa maximizam utilizarea datelor disponibile in procesul de invatare

Obtinerea seturilor de antrenare / validare - metode de esantionare

- sa extragem in mod repetat instante pentru setul de antrenare si sa invatam modelul de interes pentru fiecare esantion obtinut in acest fel, cu scopul de a obtine informatie suplimentara despre modelul rezultat
- in acest fel, obtinem informatii despre variabilitatea modelelor obtinute prin metoda aleasa
- esantionarea este costisitoare dpdv computational, deoarece necesita construirea multor modele, pentru fiecare esantion individual
- acest proces trebuie aplicat doar atunci cand nu sunt destule date disponibile, sau daca dorim sa maximizam utilizarea datelor disponibile in procesul de invatare
- folosirea esantionarii repetate presupune randomizare

Metode de validare a rezultatelor

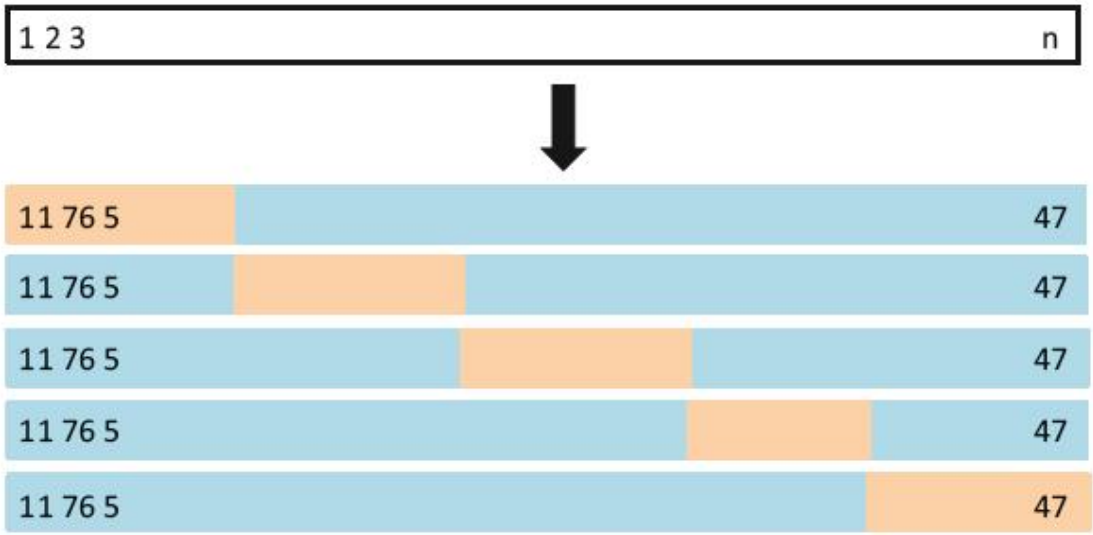
- *hold-out*:
  - se imparte setul de date in 2 parti: una pentru antrenare si cealalta pentru validare
  - proportia instantelor este de obicei $80\% - 20\%$ sau $70\% - 30\%$
  - se aplica daca avem suficiente date disponibile, astfel incat ambele seturi rezultate sa fie suficient de mari
- cross-validation
- bootstrap

Cross validation

- impartim setul de antrenament intr-un numar k de subseturi (de obicei fiecare cu numar egal de instante)
- pastram unul dintre cele k subseturi in afara procesului de antrenare, antrenam modelul pe instantele din celelalte $k-1$ subseturi si estimam rata de eroare pe subsetul pastrat deoparte
- repetam procesul de antrenare de k ori, pana cand toate instantele din setul de antrenament vor fi tinute deoparte

k-folds cross validation

- impartim aleator setul de date in k parti de aproximativ aceasi dimensiune
- primul subset este folosit pentru validare si restul $k-1$ subseturi sunt folosite pentru invatarea modelului
- Err_i este eroarea de test calculata pe instantele din setul lasat in afara invatarii
- se repeta procedura de mai sus de k ori si se calculeaza estimarea erorii de cross validare $CV_k = \frac{1}{k}\sum_{i=1}^{k} Err_i$
- in mod obisnuit folosim cross-validare cu $k = 5$ sau $k = 10$



Leave-one-out cross validation

Leave-one-out cross validation

- setul de validare este format dintr-o singura instanta (x_1, y_1)
- celelalte $n-1$ instante $\{(x_2, y_2), \ldots, (x_n, y_n)\}$ sunt folosite pentru antrenare
- calculam indicatorii de calitate a potrivirii pe instanta lasata pentru validare (de ex. rata erorii de test pentru clasificare: $Err_1 = I(y_1 \neq \hat{y}_1)$)
- repetam procedura de mai sus de n ori, de fiecare data cu alta instanta lasata afara pentru validare
- calculam valoarea estimarii de potrivire a modelului $CV(n) = \frac{1}{n} \sum_{i=1}^{n} Err_i$

Leave-one-out cross validation

- setul de validare este format dintr-o singura instanta (x_1, y_1)
- celelalte $n-1$ instante $\{(x_2, y_2), \ldots, (x_n, y_n)\}$ sunt folosite pentru antrenare
- calculam indicatorii de calitate a potrivirii pe instanta lasata pentru validare (de ex. rata erorii de test pentru clasificare: $Err_1 = I(y_1 \neq \hat{y}_1)$)
- repetam procedura de mai sus de n ori, de fiecare data cu alta instanta lasata afara pentru validare
- calculam valoarea estimarii de potrivire a modelului $CV(n) = \frac{1}{n} \sum_{i=1}^{n} Err_i$

Avantaje

- nu avem randomizare
- tinde sa nu supraestimeze rata de eroare, in comparatie cu abordarile care folosesc seturi de validare

Bootstrap

- reprezinta o procedura care ne permite sa emulam procesul de obtinere de noi seturi de date cu instante
- in mod repetat, esantionam cu inlocuire observatiile din setul de date disponibile pentru antrenament
- daca avem un set de date cu n instante, si extragem cu inlocuire de n ori, setul de date obtinut pentru antrenare va contine aproximativ 63% dintre instantele setului initial (*0.632 bootstrap*)
  - instantele ramase afara (celelalte aprox. 37% instante) sunt utilizate ca si set de validare
  - deoarece eroarea de testare este supraestimata datorita marimii setului de validare, aceasta se poate inlocui cu

$$e = 0.632 \times err_{\textit{test instances}} + 0.368 \times err_{\textit{training instances}} \tag{2}$$

- se poate transmite ca si parametru numarul de instante dorite in setul de antrenare (80%, 70% etc), si atunci numarul de extrageri poate fi diferit de n

Definitia clasificarii

Definitia clasificarii

- O problema de predictie unde variabila dependenta este calitativa

$\rightarrow$ implica asignarea fiecarei observatii intr-o clasa anume

Definitia clasificarii

- O problema de predictie unde variabila dependenta este calitativa

$\rightarrow$ implica asignarea fiecarei observatii intr-o clasa anume

Date de antrenament

Instante care au fost evaluare manual (a-priori), si au fost asignate in clasele considerate. Aceste instante sunt folosite pentru obtinerea modelului de clasificare

Definitia clasificarii

- O problema de predictie unde variabila dependenta este calitativa

$\rightarrow$ implica asignarea fiecarei observatii intr-o clasa anume

Date de antrenament

Instante care au fost evaluare manual (a-priori), si au fost asignate in clasele considerate. Aceste instante sunt folosite pentru obtinerea modelului de clasificare

Date de test

Un set de date folosit pentru a estima puterea modelului de clasificare invatat pe setul de antrenament

Definitia clasificarii

- O problema de predictie unde variabila dependenta este calitativa

$\rightarrow$ implica asignarea fiecarei observatii intr-o clasa anume

Date de antrenament

Instante care au fost evaluare manual (a-priori), si au fost asignate in clasele considerate. Aceste instante sunt folosite pentru obtinerea modelului de clasificare

Date de test

Un set de date folosit pentru a estima puterea modelului de clasificare invatat pe setul de antrenament

Obiectivele procesului de invatare

Dorim sa obtinem un model pentru clasificare (clasificator) care sa functioneze bine nu doar pe datele de antrenament, cat mai ales pe datele de test, care nu au fost utilizate in procesul de invatare a modelului

Clasificare - formalizare matematica

Y: variabila de iesire (dependenta) *calitativa* care trebuie prezisa
$X_1, X_2, \ldots, X_p$: p variabilele independente (predictori)

fiind date instantele de *antrenament* $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
trebuie sa invatam o estimare $\hat{f}$ a relatiei f dintre Y si $X = (X_1, X_2, \ldots, X_p)$ astfel incat:

$$Average(I(y_0 \neq \hat{y}_0)) \tag{3}$$

este cea mai mica, in conditiile in care (x_0, y_0) sunt instantele din *setul de testare* si $I(y_i \neq \hat{y}_i)$ este un indicator care ia valoarea 1 daca $y_i \neq \hat{y}_i$ si ia valoarea 0 daca $y_i = \hat{y}_i$
valoarea din ec. (1) se numeste *rata de eroare de testare*

Metrici pentru evaluarea calitatii modelelor de clasificare

Metrici pentru evaluarea calitatii modelelor de clasificare

Acuratetea / Rata erorii

- S: numarul instantelor cu clasa tinta prezisa corect
- E: numarul instantelor cu clasa tinta prezisa eronat
- Rata erorii: $E/(S+E)$
- Acuratetea: 1 - rata erorii $= S/(S+E)$

Metrici pentru evaluarea calitatii modelelor de clasificare

Acuratetea / Rata erorii

- S: numarul instantelor cu clasa tinta prezisa corect
- E: numarul instantelor cu clasa tinta prezisa eronat
- Rata erorii: $E/(S+E)$
- Acuratetea: 1 - rata erorii $= S/(S+E)$

- Daca seturile de antrenament si testare sunt reprezentative, acuratetea reprezinta un estimator fiabil pentru intreaga populatie
- Putem sa calculam estimari fiabile doar pe seturi de instante care nu au fost utilizate in pasul de invatare a modelului
- *Rata erorii de inlocuire*: rata de eroare obtinuta pe setul de antrenament
  - reprezinta o estimare optimista a ratei de eroare reale

Intervale de incredee

- acuratetea (sau rata erorii) reprezinta estimari punctuale ale valorilor reale
- ele trebuie insotite de intervalele de incredere
- pentru un prag de semnificatie (ex. 95%), se calculeaza un interval de incredere pentru valoarea punctuala a metricii
- cu cat este mai inalt pragul de semnificatie (confidence threshold), cu atat intervalul este mai larg
- pentru o incredere de 95%, intervalul de incredere pentru o estimare $\hat{\theta}$ este:

$$CI(\hat{\theta}) = \left[\hat{E}(\hat{\theta}) \pm 2SE(\hat{\theta})\right] \tag{4}$$

Alte metrici pentru calculul calitatii potrivirii

Alte metrici pentru calculul calitatii potrivirii

Matricea de confuzie

Compara predictiile realizate de model cu valorile reale
Elementele de pe diagonala principala reprezinta indivizii prezisi in mod corect
Elementele care nu sunt pe diagonala reprezinta indivizii clasificati gresit.

		Valorile reale	
		Yes	No
Valorile prezise	Yes	true positive	false positive
	No	false negative	true negative

Alte metrici pentru calculul calitatii potrivirii

Matricea de confuzie

Compara predictiile realizate de model cu valorile reale
Elementele de pe diagonala principala reprezinta indivizii prezisi in mod corect
Elementele care nu sunt pe diagonala reprezinta indivizii clasificati gresit.

		Valorile reale	
		Yes	No
Valorile prezise	Yes	true positive	false positive
	No	false negative	true negative

- *Precizia* clasa Yes: $\frac{TP}{TP+FP}$,
- *Precizia* clasa No: $\frac{TN}{TN+FN}$,

Alte metrici pentru calculul calitatii potrivirii

Matricea de confuzie

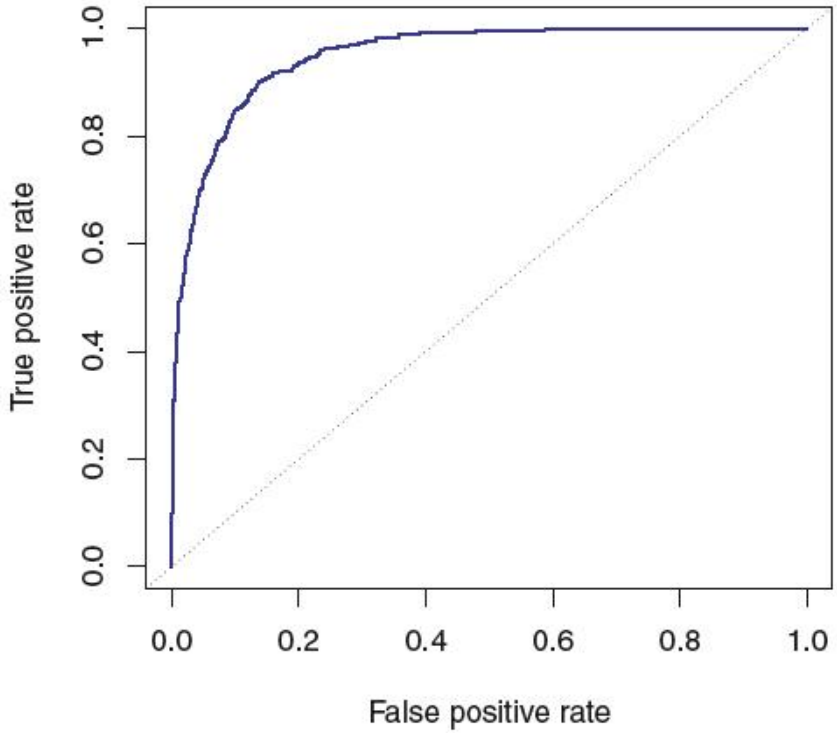
Compara predictiile realizate de model cu valorile reale
Elementele de pe diagonala principala reprezinta indivizii prezisi in mod corect
Elementele care nu sunt pe diagonala reprezinta indivizii clasificati gresit.

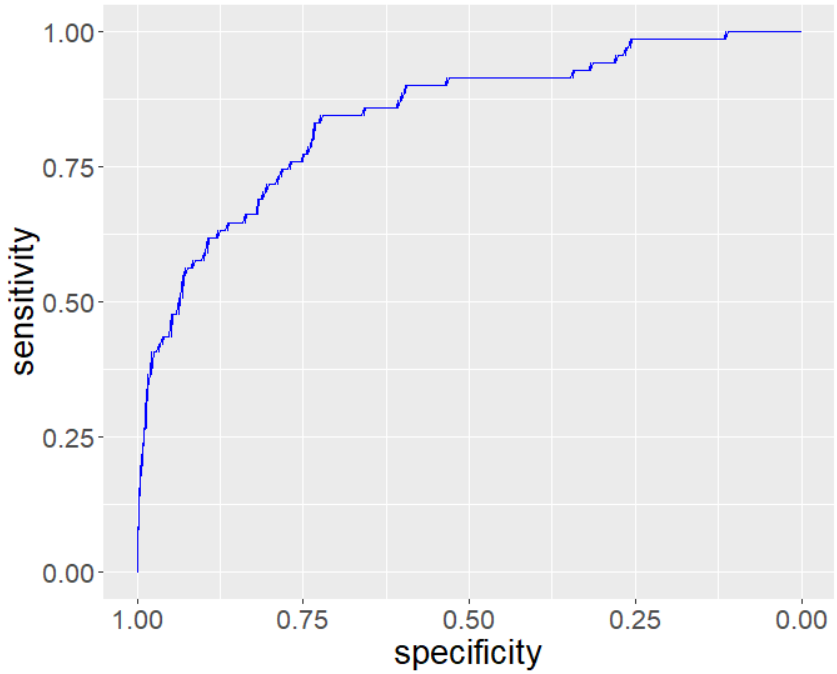
		Valorile reale	
		Yes	No
Valorile prezise	Yes	true positive	false positive
	No	false negative	true negative

- *Precizia* clasa Yes: $\frac{TP}{TP+FP}$,
- *Precizia* clasa No: $\frac{TN}{TN+FN}$,
- *Recall* clasa Yes: $\frac{TP}{TP+FN}$, denumita si *sensitivitate*
- *Recall* clasa No: $\frac{TN}{TN+FP}$, denumita si *specificitate*

Curba ROC

prezinta rata obtinuta pentru *true positives* (*sensitivity*), raportat la rata de *false positives* (*specificity*)





- aria de sub curba ROC (denumita AUC): sumarizeaza performanta clasificatorului pentru toate specificatiile posibile
  - cu cat este mai apropiata de partea de sus a graficului, cu atat mai buna
- daca clasificatorul nu are o performanta mai buna decat sansa: atunci $AUC = 0.5$

Resurse pentru probleme de clasificare in R

- Pachetul Caret `http://topepo.github.io/caret/index.html`

Cod R

```
install.packages("caret")
library(caret)
```

Se va confirma instalarea tuturor pachetelor de care depinde caret.

Metode probabilistice pentru clasificare

- modelul rezultat va calcula probabilitatea de apartenenta a unei instante la o clasa tinta
- suma probabilitatilor de apartenenta a unei instante la toate clasele tinta trebuie sa fie 1
- metode studiate:
  - Laborator 3.1 - Metoda Naive Bayes
  - Laborator 3.2 - Regresia logistica