

Inteligentă Artificială

Bogdan Alexe

bogdan.alexe@fmi.unibuc.ro

Secția Tehnologia Informației, anul III, 2022-2023
Cursul 4

Recapitulare – cursul trecut

1. Modelul celor mai apropiati k vecini (k-nearest neighbors) – continuare
2. Normalizarea caracteristicilor
3. Clasificatorul naïve Bayes

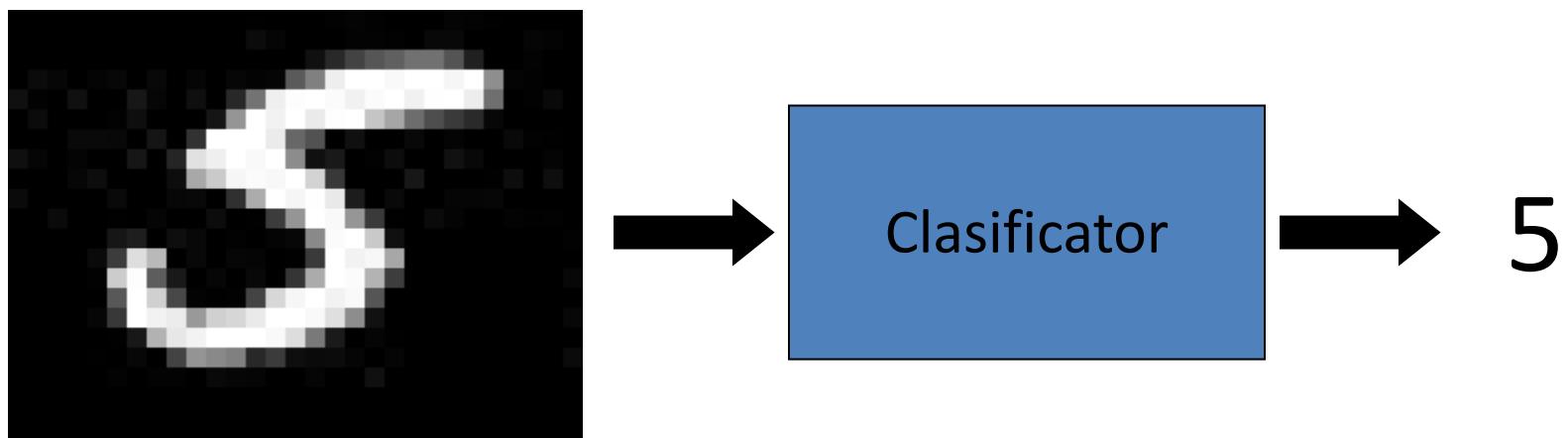
Cuprinsul cursului de azi

1. Clasificatorul naïve Bayes
2. Evaluarea performanței unui model
3. Strategii de împărțire a datelor
4. Proiect
 - concurs pe platforma Kaggle
 - demo
 - laborator săptămâna 5
5. Mașini cu vectori suport (SVMs – Support Vector Machines)

Clasificatorul naïve Bayes

Clasificare - exemplu

- Definirea problemei:
 - date fiind caracteristicile măsurate X_1, X_2, \dots, X_n
 - realizați o predicție a etichetei c
- Clasificarea cifrelor scrise de mână (laborator):
 - X_1, X_2, \dots, X_n sunt intensitățile pixelilor dintr-o imagine grayscale
 - $X_1, X_2, \dots, X_n \in \{0, 1, 2, \dots, 255\}$ – intensități
 - $c \in \{0, 1, 2, \dots, 9\}$ (clasifică ce cifră este imaginea)



Regula lui Bayes

- Clasificarea cifrelor scrise de mâňă:
 - X_1, X_2, \dots, X_n sunt intensităile pixelilor dintr-o imagine grayscale
 - $X_1, X_2, \dots, X_n \in \{0, 1, 2, \dots, 255\}$ – intensităti
 - $c \in \{0, 1, 2, \dots, 9\}$ (clasifică ce cifră este imaginea)
- Considerăm X imaginea cu n pixeli de intensitate X_1, X_2, \dots, X_n : $X = (X_1, X_2, \dots, X_n)$
- Regula lui Bayes: $P(c | X) = \frac{P(X | c) \times P(c)}{P(X)}$

Ω - spaťiu total de evenimente – modelează fenomenul

$\Omega = \{(c, X)\}$ – perechi de forma (clasă, imagine)

Regula lui Bayes

- Considerăm X imaginea cu pixelii de intensitate X_1, X_2, \dots, X_n : $X = (X_1, X_2, \dots, X_n)$

- Regula lui Bayes:

$$P(c | X) = \frac{P(X | c) \times P(c)}{P(X)}$$

Probabilitatea să observăm imaginea X condiționată de faptul că imaginea conține o cifră din clasa c

Probabilitatea a-priori ca o imagine să fie din clasa c (să conțină cifră din clasa c)

Probabilitatea să avem cifra din clasa c dându-se imaginea X

Probabilitatea să observăm imaginea X

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

Regula de clasificare

$$P(c = 0 / X) = \frac{P(X / c = 0) \times P(c = 0)}{P(X)}$$

$$P(c = 1 / X) = \frac{P(X / c = 1) \times P(c = 1)}{P(X)}$$

.....

$$P(c = 9 / X) = \frac{P(X / c = 9) \times P(c = 9)}{P(X)}$$

Presupunem că am calculat cele 10 probabilități a-posteriori:
 $P(c=0|X)$, $P(c=1|X)$, ... $P(c=9|X)$

Regula de clasificare: $c^* = \arg \max_{i=0,1,\dots,9} P(c = i / X)$

alege clasa care maximizează probabilitatea a-posteriori

Regula de clasificare

\propto - direct proporțional

$$P(c = 0 / X) \propto P(X / c = 0) \times P(c = 0)$$

$$P(c = 1 / X) \propto P(X / c = 1) \times P(c = 1)$$

.....

$$P(c = 9 / X) \propto P(X / c = 9) \times P(c = 9)$$

Presupunem că am calculat cele 10 probabilități:

$$\text{Regula de clasificare: } c^* = \operatorname{argmax}_{i=0,1,\dots,9} P(X / c = i) \times P(c = i)$$

alege clasa care maximizează numărătorul

Calculul probabilității likelihood folosind clasificatorul naïve Bayes

$$P(c = i | X) \propto P(X | c = i) \times P(c = i)$$

↑
Probabilitatea
likelihood

- **clasificatorul naïve Bayes consideră caracteristicile independente (nu e întotdeauna adevărat acest lucru)**

$$P(X | c = i) = \prod_{j=1}^{n=784} P(X_j = x_j | c = i)$$

- $P(X|c = i) = P(X_1=x_1, X_2=x_2, \dots, X_{784}=x_{784}|c = i) =$ (folosind presupunerea de independentă între caracteristici) =
 $= P(X_1=x_1|c=i) \times P(X_2=x_2|c=i) \times \dots P(X_{784}=x_{784}|c = i)$

Calculul probabilității individuale $P(X_j = x_j | c=i)$

Nu avem date să estimăm corect asemenea probabilități, avem în jur de 100 de puncte X_j în mulțimea de antrenare.

O posibilă soluție:

Împart intervalul de valori posibile [0,255] în p părți egale și estimez care este probabilitatea ca o valoare să apară într-un astfel de interval.

Exemplu: consider $p = 4$, obțin 4 intervale:

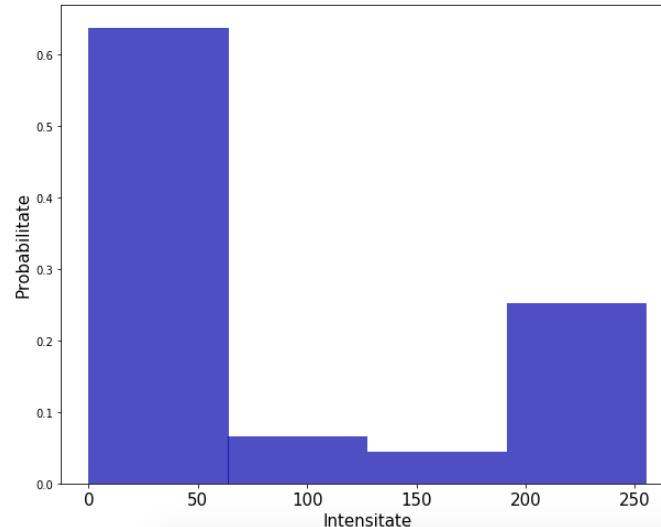
[0, 63], [64, 127], [128, 191], [192,255]

Aproximez:

$P(X_j = 0|c)$ cu $P(X_j \in [0,63]|c)$

$P(X_j = 100|c)$ cu $P(X_j \in [64,127]|c)$

Probabilitatea va arăta astfel:



Stabilitate numerică

Regula de clasificare pentru naïve Bayes:

$$c^* = \operatorname{argmax}_{i=0,1,\dots,9} \left(\prod_{j=1}^{n=784} P(X_j = x_j \mid c = i) \right) \times P(c = i)$$


înmulțesc $n=784$ numere subunitare (sunt probabilități), uneori foarte aproape de 0.

Aplic funcția logaritm (e monoton crescătoare), păstrează ierarhia:

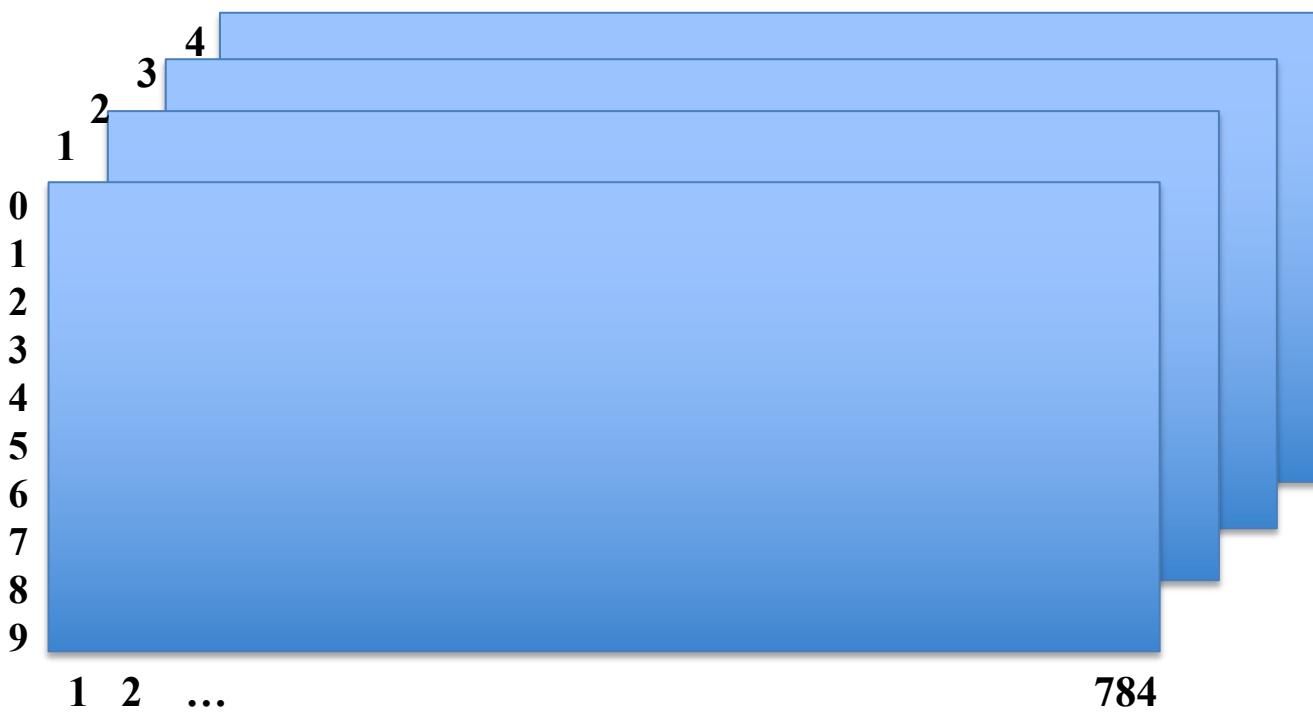
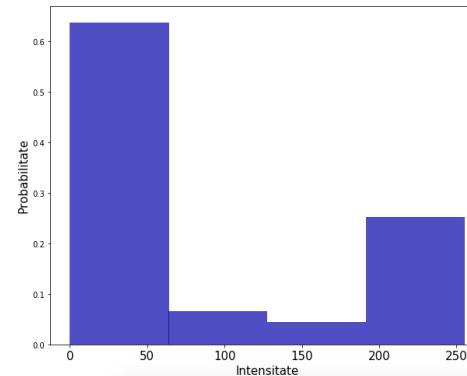
$$c^* = \operatorname{argmax}_{i=0,1,\dots,9} \left(\log \left(\prod_{j=1}^{n=784} P(X_j = x_j \mid c = i) \right) \right) \times P(c = i))$$
$$c^* = \operatorname{argmax}_{i=0,1,\dots,9} \left(\sum_{j=1}^{n=784} \log(P(X_j = x_j \mid c = i)) + \log(P(c = i)) \right)$$

Model parametric

Care sunt parametri învățați de către model?

Modelul învăță din datele de antrenare pentru fiecare clasă i (linie) și componentă j (coloană)

$P(X_j = x_j | c=i)$ sub forma unui vector de probabilități de dimensiune 4



Numărul de parametri:
 $784 \times 10 \times 4 = 31360$

De fapt întrucât fiecare vector de dimensiune 4 are suma 1:
 $784 \times 10 \times 3 = 24520$

Hiperparametrul p = numărul de părți, p = 4

Clasificatorul Bayes pe cazul general

- X vector de n caracteristici $X = (X_1, X_2, \dots, X_n)$
- c clase (admis/respins, bărbat/femeie,litere, cifre, etc.)
- Regula lui Bayes:

$$P(c | X) = \frac{P(X | c) \times P(c)}{P(X)}$$

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

$$P(c = i | X) \propto P(X | c = i) \times P(c = i)$$

- **clasificatorul naïve Bayes consideră caracteristicile independente (nu e întotdeauna adevărat acest lucru)**

$$P(X | c = i) = \prod_{j=1}^n P(X_j = x_j | c = i)$$

Clasificatorul Bayes pe cazul general

- X vector de n caracteristici $X = (X_1, X_2, \dots, X_n)$
- c clase (admis/respins, litere, cifre, etc.)
- Regula de clasificare pentru Naïve Bayes:

$$c^* = \operatorname{argmax}_i \left(\prod_{j=1}^n P(X_j = x_j \mid c = i) \right) \times P(c = i)$$

trebuie estimată din date

- stabilitate numerică:

$$c^* = \operatorname{argmax}_i \left(\log \left(\prod_{j=1}^n P(X_j = x_j \mid c = i) \right) \times P(c = i) \right)$$

$$c^* = \operatorname{argmax}_i \left(\sum_{j=1}^n \log(P(X_j = x_j \mid c = i)) + \log(P(c = i)) \right)$$

Evaluarea performanței unui model

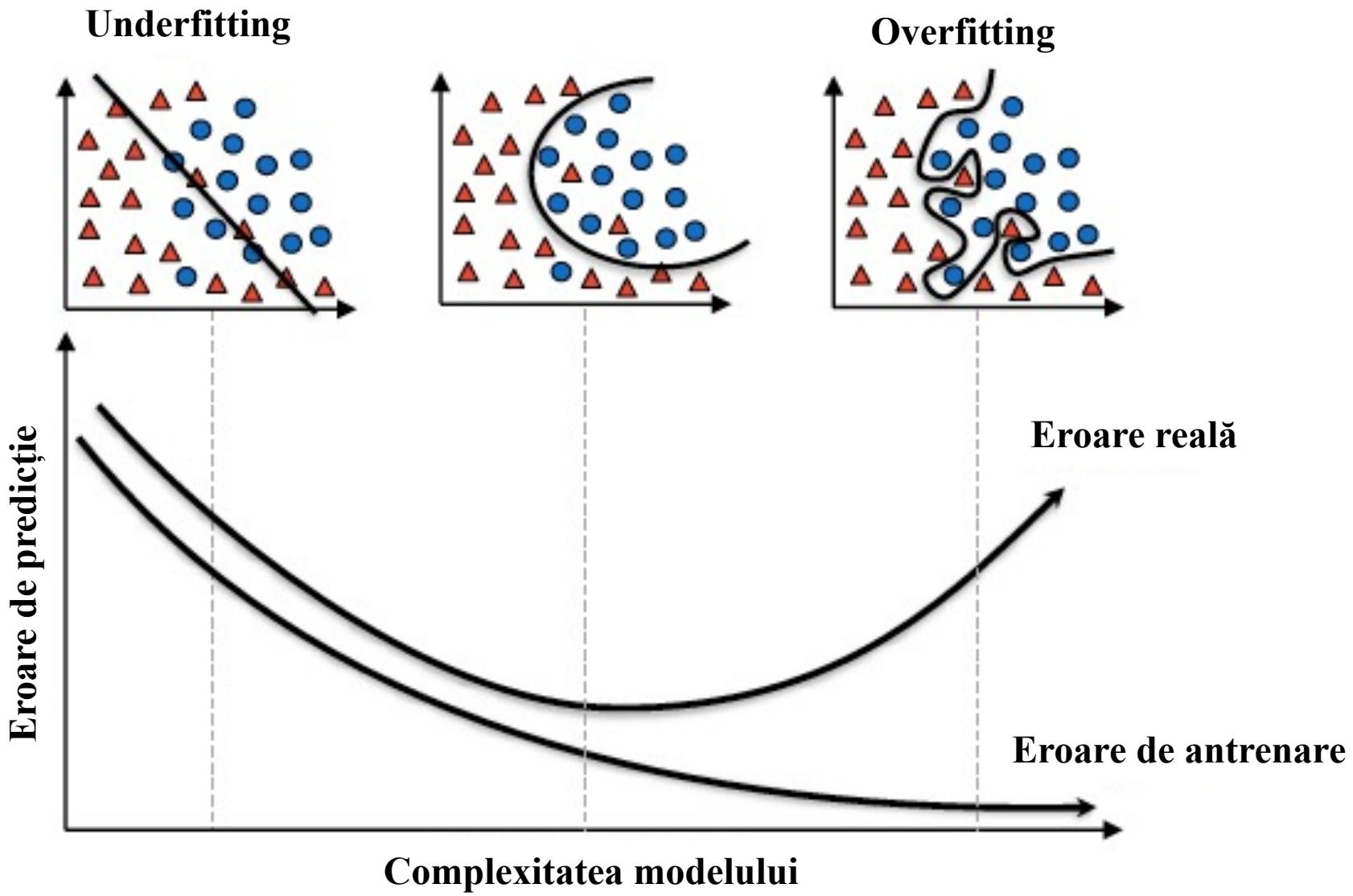
Evaluarea unui model

- Un model este bun dacă performează bine pe date (de test) pe care nu le-a văzut înainte (la antrenare).
 - nu ține minte datele de antrenare ci are capacitatea de generalizare
 - **overfitting (supra-învățare)** – modelul performează bine pe datele de antrenare dar nu poate generaliza
 - **underfitting (sub-învățare)** – modelul performează slab și pe datele de antrenare și pe cele de test.
- Eroarea reală = eroarea pe toate datele posibile
- Eroare empirică = eroarea pe o mulțime finită de puncte
 - se mai numește și **eroare de testare, eroare de generalizare**
- Vrem să minimizăm eroarea totală, imposibil de măsurat
- Vrem să ne asigurăm că eroarea empirică aproximează eroare reală. Cum realizăm acest lucru?

Evitarea fenomenului de overfitting

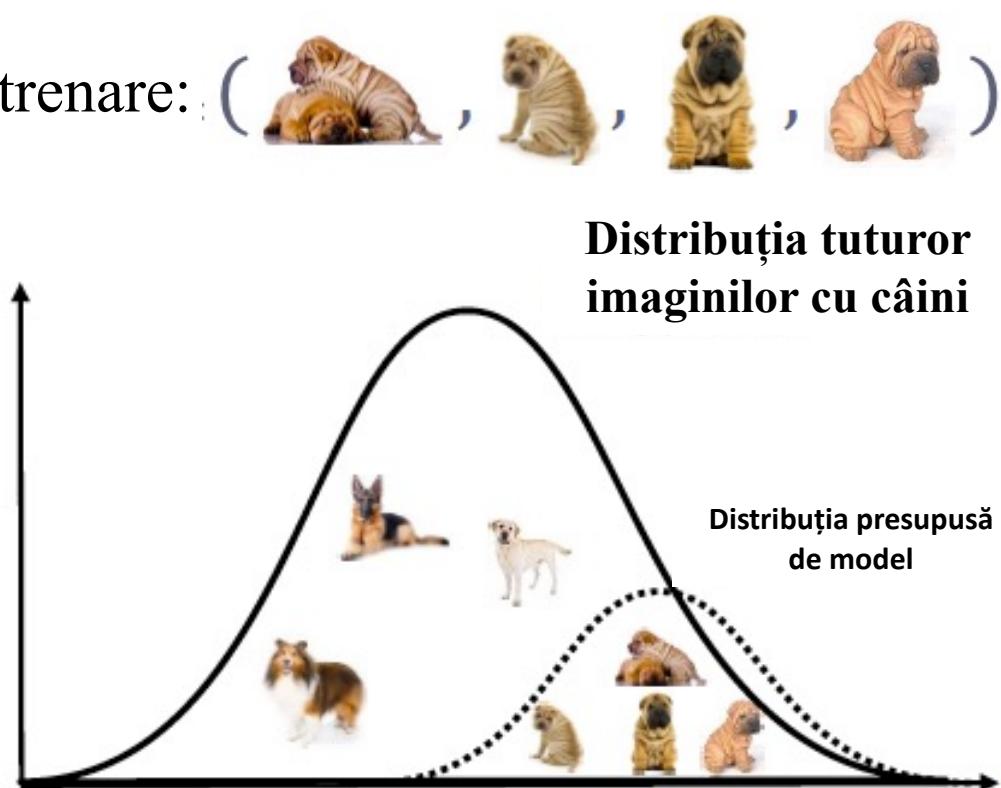
- Nu calculăm eroarea empirică pe datele folosite la antrenare
 - un set de date separat pentru testare
- Nu alegem hiperparametri modelului care să conducă la o performanță bună pe datele de test
 - un set de date separat pentru validare
 - folosim validare încrușitată (Cross Validation)
- Preferăm modele mai simple în locul celor mai complexe
 - penalizarea complexității unui model se numește **regularizare**
- Vrem să ne asigurăm că datele sunt i.i.d. (independente și identic distribuite)
 - fiecare exemplu (de antrenare sau testare) este selectat independent din aceeași distribuție
- Mai multe date!!!

Underfitting vs. overfitting



Date i.i.d.

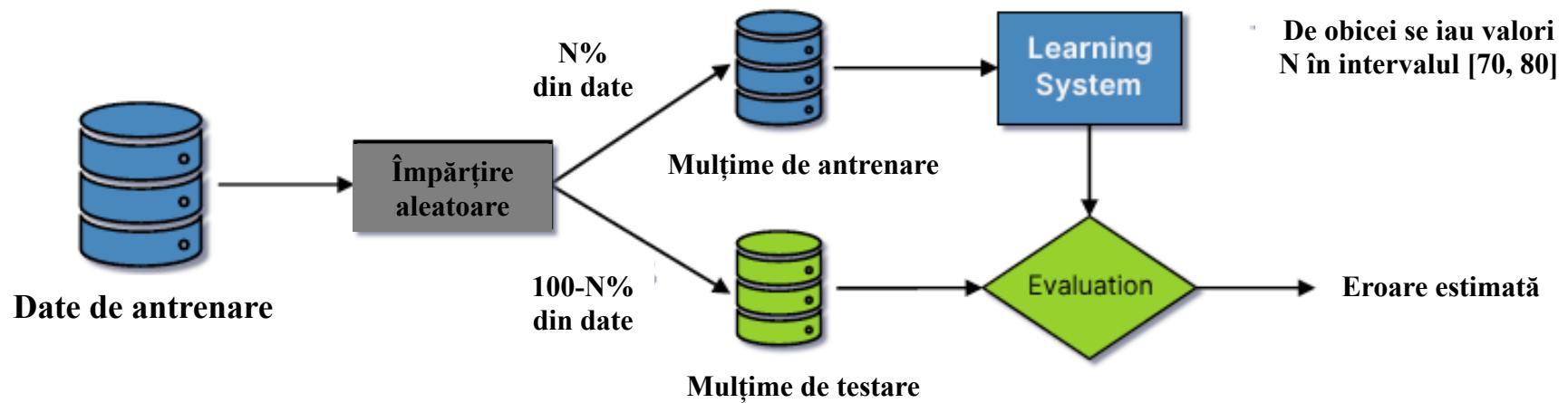
- Cei mai mulți algoritmi prespun că datele sunt independente și identic distribuite
- Presupunem că vrem să antrenăm un model care să recunoască poze cu câini
 - considerăm multimea de antrenare: (, , , )
- Exemplele nu sunt i.i.d
 - cel mai probabil modelul va face greșeli pe imagini care provin din alte părți ale distribuției



Strategii de împărțire a datelor

Mulțime de antrenare și testare

- În practică, nu putem vedea distribuția reală a datelor și uneori nu putem avea acces la foarte multe date din această distribuție
- Așa cum am văzut, eroarea de antrenare este o estimare optimistă a erorii reale:
 - întotdeauna eroarea de antrenare va fi mai mică decât eroarea reală
 - favorizează modele complexe, care pot face overfitting pe datele de antrenare
- O posibilitate de a îmbunătăți estimarea erorii reale este împărțirea datelor în două mulțimi: antrenare și testare



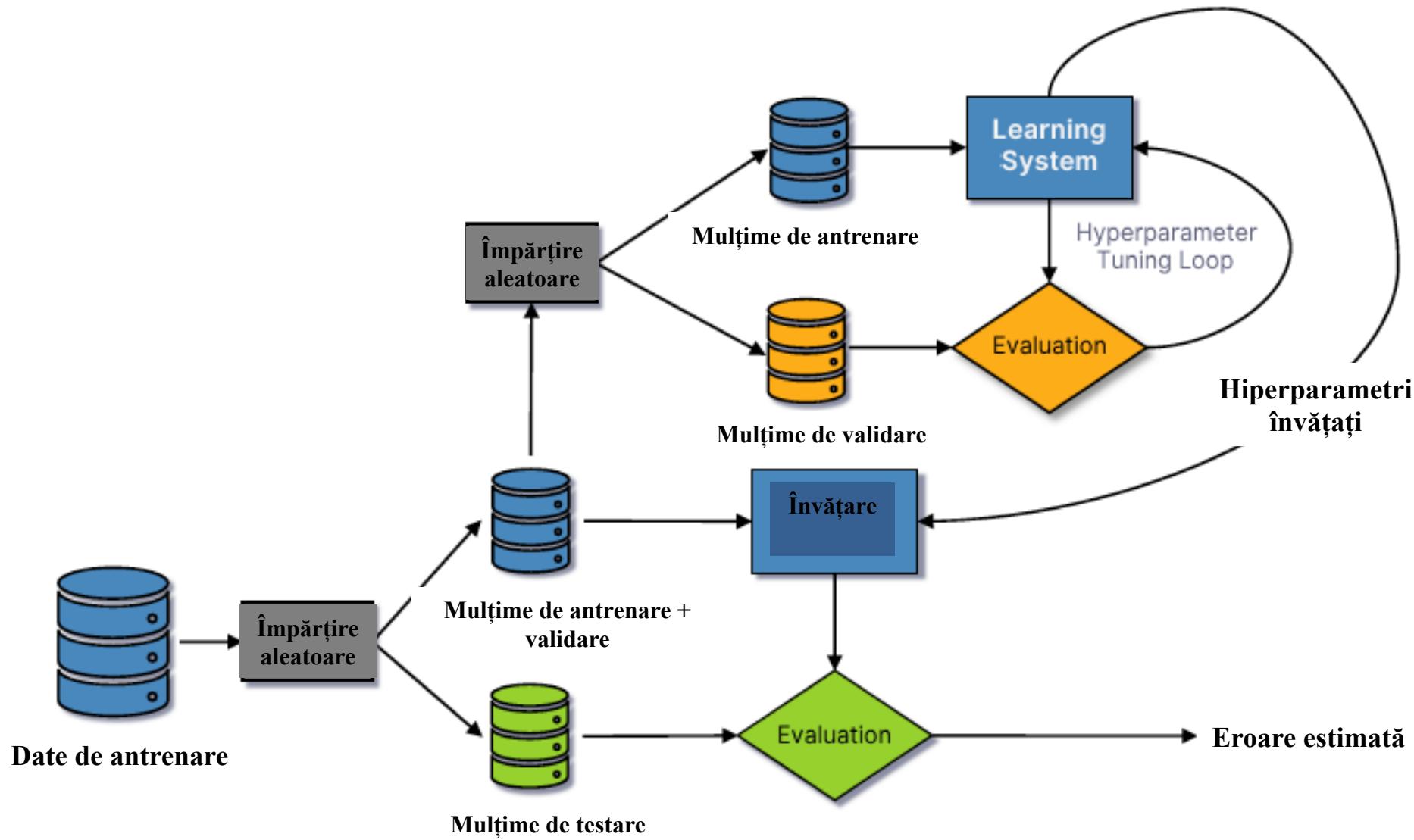
Hiperparametri

- Unele modele au hiperparametri ce controlează procesul de antrenare

Model	Parametri	Hiperparametri
K-NN	Nu are	$K =$ numărul de vecini
Naïve Bayes	$P(X_j = x_j c=i)$ – probabilitatea individuală de likelihood	p – numărul de intervale în care împart domeniul de valori media, deviația standard unei distributii
SVM (va urma)	w, b – definesc hiperplanul de separare	C – controlează trade-off-ul dintre margine și acuratețe
Rețele neuronale de perceptri (va urma)	Ponderile rețelei	Arhitectura rețelei, funcții de activare, etc.

- Vreau să aleg valori pentru hiperparametri care conduc la o eroare reală mică
 - nu e bine să optimizez alegera hiperparametrilor pe mulțimea de testare (adică să fixez diverse valori pentru un hiperparametru, antrenez parametri pe mulțimea de antrenare și să aleg valoarea hiperparametrului care minimizează eroarea pe mulțimea de testare)
 - e posibil să fac overfitting în spațiul hiperparametrilor pe mulțimea de testare
 - aş vrea să nu mă ating de mulțimea de testare
- Folosesc o mulțime separată de *validare*

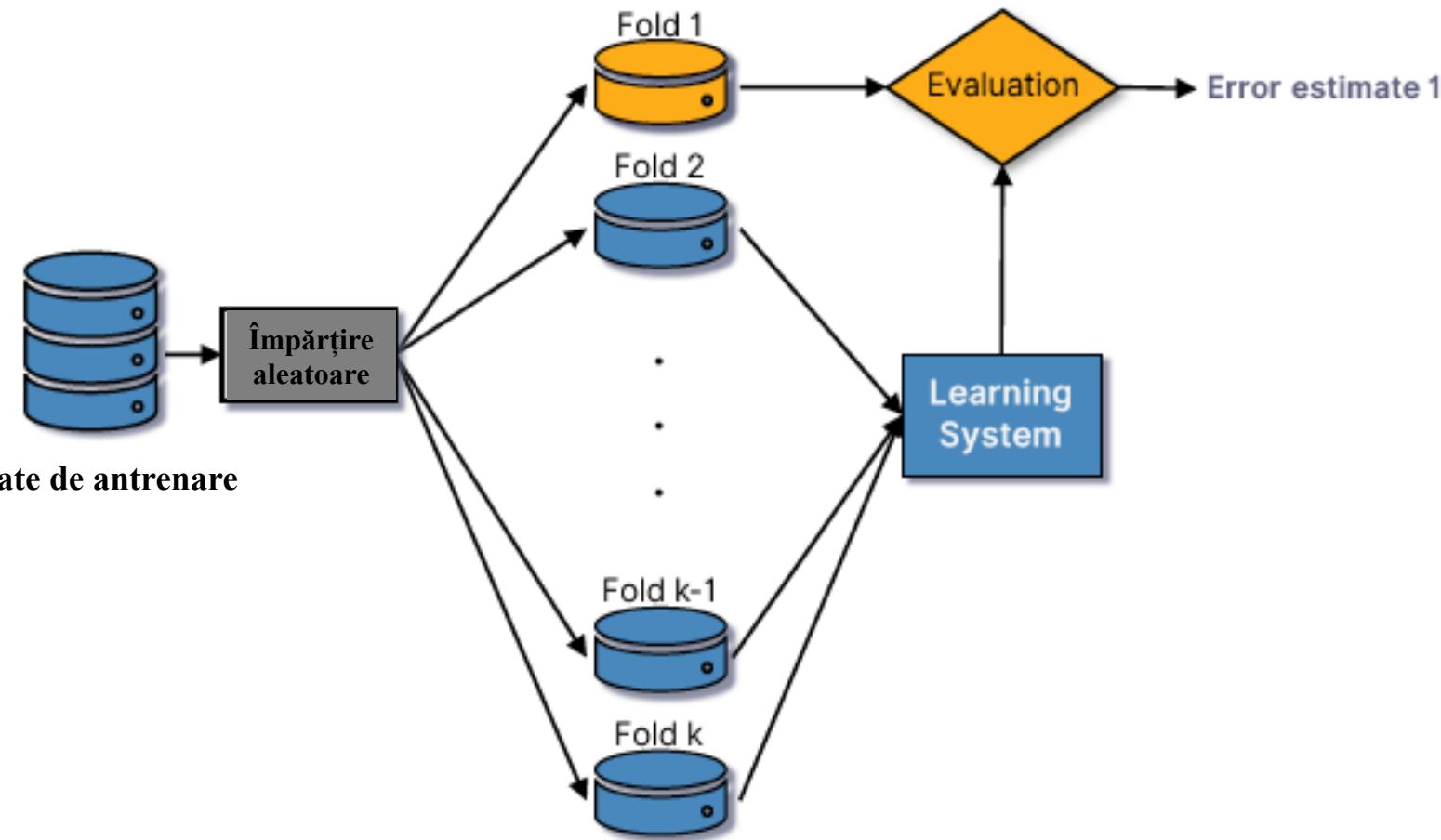
Mulțime de antrenare, validare și testare



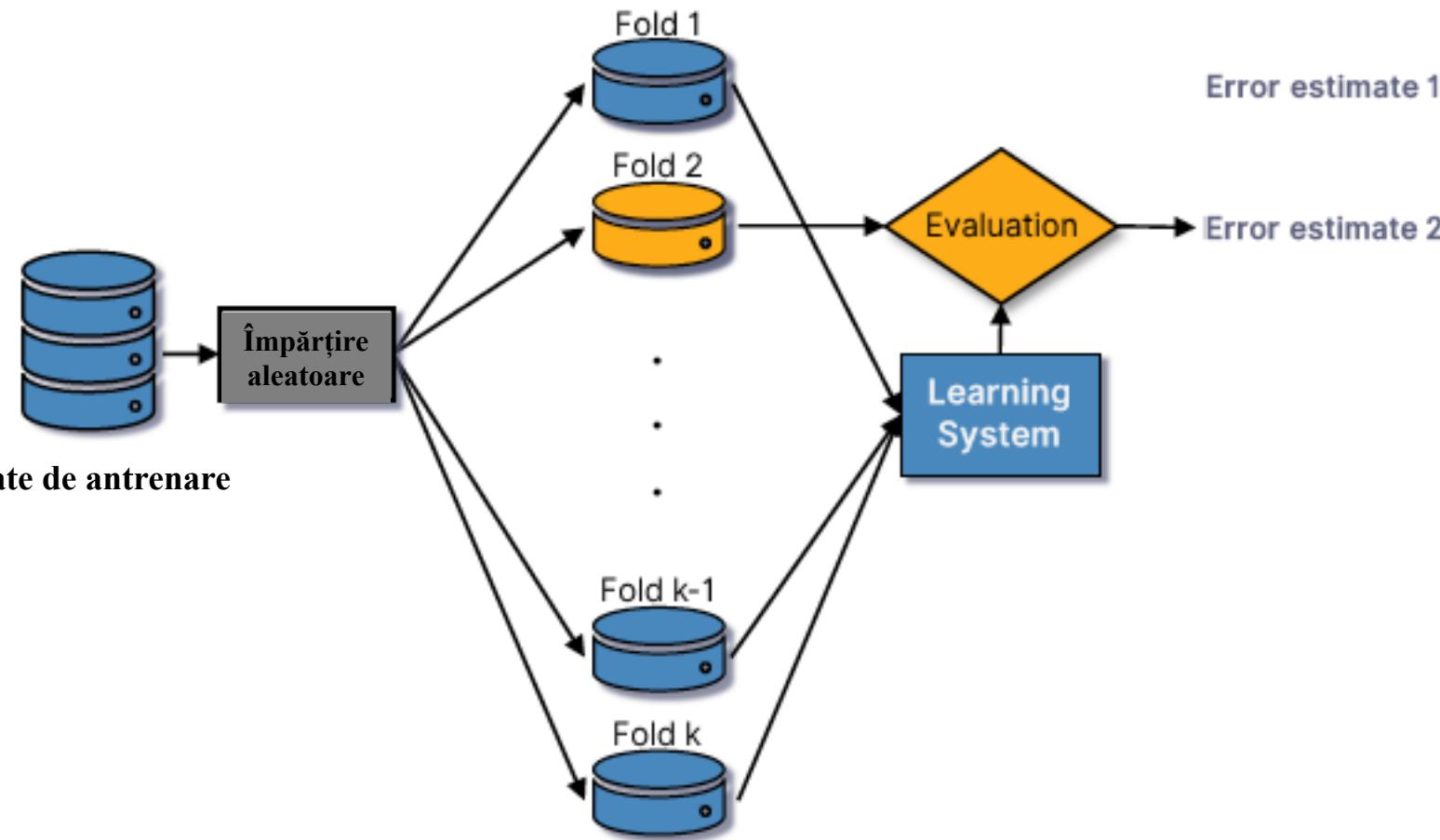
K-fold Cross-Validation

- În practică, e posibil să nu avem atât de multe date încât să ne “permitem” o mulțime de validare
 - dacă am împărți datele în mulțime de antrenare, validare, testare am putea obține mulțimi mult prea mici
- K-fold Cross-Validation (validare încrucișată pe k părți)
 - împarte datele în k părți egale (parte = fold)
 - repetă de k-ori procesul în care antrenezi pe k-1 părți și testezi pe o parte (rămasă)
 - estimează eroarea reală ca media celor k erori de testare pe cele k părți
- Chiar dacă ne “permitem” o mulțime de validare, prin K-fold Cross Validation obținem o estimare mai bună a erorii reale
 - e nevoie de putere computațională mare

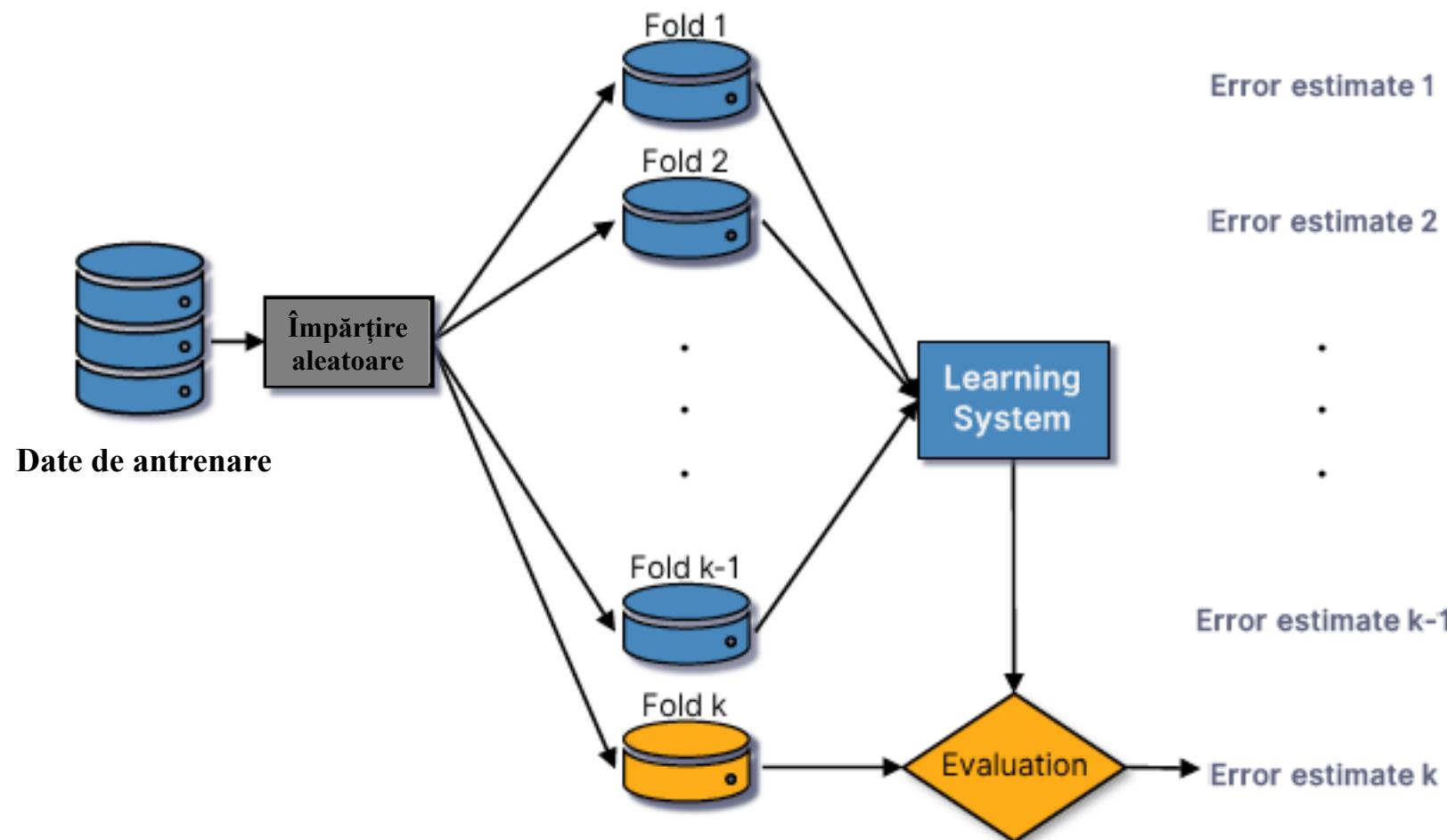
K-fold Cross-Validation



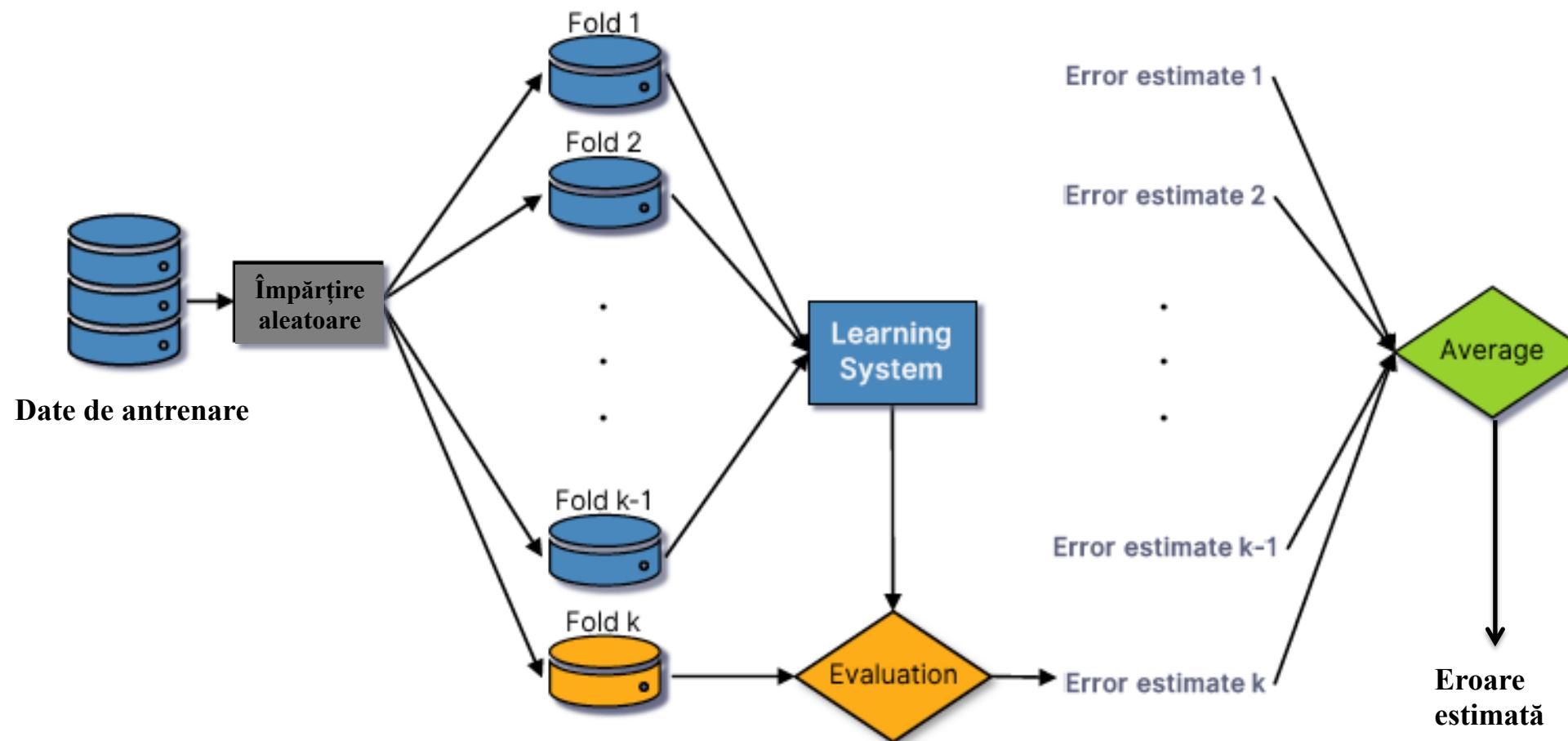
K-fold Cross-Validation



K-fold Cross-Validation



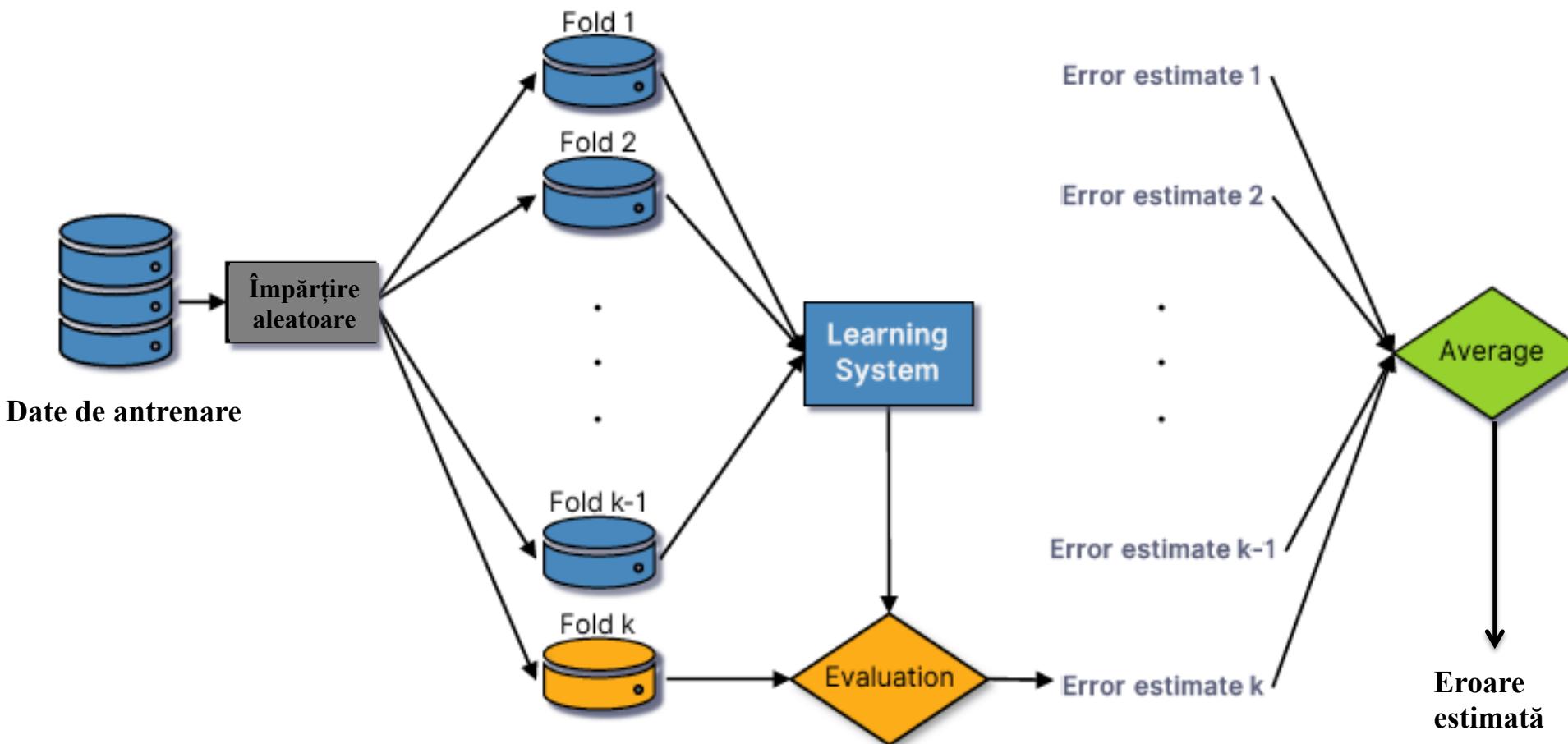
K-fold Cross-Validation



Leave One-Out Cross-Validation

Caz extrem: K = numărul de exemple de antrenare

Antrenăm pe K-1 exemple, testăm pe un exemplu



Strategii de împărțire a datelor - summar

- Folosirea erorii de antrenare pentru evaluarea performanței unui model nu este o idee bună:
 - eroare de antrenare este o estimare optimistă a erorii reale
 - favorizează modele complexe care pot face overfitting
- Folosirea unui mulțimi separate de testare oferă o estimare mai bună a erorii reale
- Dacă folosim o singură mulțime test pentru alegerea hiperparametrilor putem ajunge din nou la o estimare incorectă a erorii reale
 - overfitting pe mulțimea de testare în alegerea valorilor optime a hiperparametrilor
 - soluția este să folosim o mulțime de validare
- Dacă nu avem suficiente date de antrenare sau dacă ne “permitem” din punct de vedere computațional K-Fold Cross Validation oferă o estimare mai bună a erorii reale:
 - Leave-One Out Cross Validation: $K = \text{numărul de exemple de antrenare}$

Proiect – concurs pe platforma Kaggle

Corpus de traduceri

- Discursuri în Parlamentul European ținute de către europarlamentari din Anglia, Scoția, Irlanda – vorbitori de limbă engleză dar fiecare cu dialectul nativ – trei clase = {England, Scotland, Ireland}
- Traduceri în cinci limbi (daneză, germană, spaniolă, italiană, olandeză) a discursurilor eurparlamentarilor din Anglia, Scoția, Irlanda
- Scopul proiectului este de a identifica dialectul sursă (englez, scoțian, irlandez) din care a fost tradus textul într-o anumită limbă
- Traducerile sunt realizate de experti vorbitori nativi ai limbii în care se face traducerea și evident vorbitori de limbă engleză

Corpus de traduceri

		train_data	label
1	language	text	
2	dansk	<p>Dette er et fremragende initiativ, og jeg støtter fuldt ud målet om at fremhæve idrættens pædagogiske rolle. Der kan opnås store fordele ved at etablere partnerskaber mellem undervisningssektoren og idrættens organisationer.</p> <p>Som irsk medlem er jeg specielt glad for, at forslaget går ud på at dedikere 2004 til Idrættens Pædagogiske Dimension. Det falder sammen med det irske formandskab i første halvdel af 2004, og jeg kan forsikre mine kolleger, at vi alle klar over, at en række store sportsbegivenheder løber af stabelen i 2004. Der afholdes EM i fodbold og olympiske og paraolympiske lege i Athen. I forslaget anses 2003 for at være et forberedende år med visse aktiviteter.</p> <p>Ireland er vært for Special Olympics i 2003, og EU's bidrag til organiseringen af disse lege er værd at rose. For atterne fra de 160 internationale delegationer verden over er dette en enestående sportslig og kulturel begivenhed.</p>	Ireland
3	dansk	<p>Hr. formand, jeg er sikker på, at alle her er klar over, at De Grønne har grundlæggende problemer med Den Europæiske Unions behandling af fiskeriaftaler. Vi er principielt ikke imod fiskeriaftaler. Men vi protesterer stærkt mod, at jeg udarbejdede Madagaskar-betænkningen, foretog en detaljeret undersøgelse af fiskeriaftalen med Madagaskar og fandt den særdeles mangelfuld. Siden da har vi stemt imod aftalerne. Mauritius-aftalen ligner meget denne.</p> <p>Dette er grunden til, at vi sammen med EDN-gruppen og hr. Macarthy har fremsat et ændringsforslag om en midtvejsrevision af aftalen. Ved denne midtvejsrevision ser vi gerne en vurdering af EU-fiskeriets indflydelse på både landbrug og miljø.</p> <p>Jeg har et ganske enkelt spørgsmål til Kommissionen og vil meget gerne have et svar her i aften. Spørgsmålet lyder således: Hvis Parlamentet vedtager vort ændringsforslag om en uafhængig midtvejsrevision, vil Kommissionen tage det i betragtning?</p>	Ireland
4	dansk	<p>Hr. formand, folk på den nordlige halvkugle tror, at aids er et overstået kapitel. Det er naturligvis ikke tilfældet, men en skiftende livsstil og fremkomsten af effektive lægemiddelordninger har i det mindste hjulpet os med at kontrollere aids. Men på den sydlige halvkugle er billede meget anderledes. 90% af de 5 millioner nye tilfælde hvert år forekommer i de lavtlønnede lande i den sydlige verdensdel. Der findes 25 millioner mennesker med aids i Afrika og 6 millioner i Asia.</p> <p>Hvad angår aids står vi over for en ond cirkel af infektioner, arbejdssygtighed og fattigdom ud over de 16 millioner dødsfald, som vi allerede har oplevet. Der har i Afrika været 2 millioner dødsfald om året på grund af aids, hvilket er næsten dobbelt så mange som i Europa.</p> <p>Da The Economist tog til Zambia og besøgte et bestemt hospital, så de, at totredjedele af patienterne var ved at dø af aids. De rapporterede, at folks lemmer lignede knækkede kosteskifter. Når de blev spurgt om deres største bekymring, svarede de, at de ikke kunne få et bedre spisevalg. I Zambia vurderede sundhedsministeriet dengang, at halvdelen af befolkningen ville dø af aids. Det er en menneskelig katastrofe. Nu står vi over for menneskelige katastrofer i USA. Verden ser på overskrifterne, verden skrider ud af sit sædvanlige.</p> <p>Fattigdomslempelse går hånd i hånd med bekæmpelse af sygdom, invaliditet og død. Vi har desperat behov for flere ressourcer til lægermidler og vacciner, men skal også sikre os, at de i tilstrækkelig grad deles ud til folk, så de kan få adgang til dem.</p> <p>Tuberkulose dræber også. Sygdommen forårsager 2 millioner dødsfald om året, hvoraf 95% er i udviklingslandene. Det er den største årsag til hiv-dødsfald. En smittet person kan videregive den til 10 andre. Malaria er den tredje største årsag til dødsfald.</p>	England
5	dansk	<p>Hr. formand, med forbehold af nogle få ændringer støtter min gruppe helt beslutningsforslaget, som er fremlagt og udarbejdet af ordføreren, hr. Tsatsos og hr. Gil-Robles Gil-Delgado.</p> <p>Den store udfordring, som vi står over for nu, er at få forfatningen gennem regeringskonferencen, uden at den skiller ad i stumper og stykker. Derfor bifalder vi, at der i dette beslutningsforslag, som vi vedtager, ikke opfordres til at udskifte forfatningen.</p> <p>Der er skjulte farer. Der er regeringer - f.eks. Spanien - som anfægter systemet med dobbelt flertal i Rådet, og som ønsker at vende tilbage til Nice-systemet, som er kompletst, uforståeligt og ikke et særligt fornuftigt system.</p> <p>Jeg opnorer også imod dem, der ønsker henvisning til religion i forfatningen. Vi er en Union af forskellige religioner, af religiøs pluralisme. Der er mennesker i vores Union, som ikke er religiøse. Vi bør ikke gennemtrumpe et bestemt sprog.</p> <p>Til sidst vil jeg sige noget om spørgsmålet om en folkeafstemning. Det er ikke EU's opgave at fortælle medlemsstaterne, hvilken fremgangsmåde de skal anvende internt til at ratificere en traktat om en forfatning. Det er hver stats opgave.</p>	England
	dansk	<p>- Hr. formand, jeg må protestere mod den lømfærdighed, hvormed Pakistanets præsident Musharraf reagerede på Abdul Qadeer Khans tilstædelse på tv. Europas Parlamentet har tidligere taget afstand fra Pakistan og ønsket af den engelske præsident.</p>	England

Corpus de traduceri

train_data

	language	text	label
1	dansk	Dette er et fremragende initiativ, og jeg støtter fuldt ud målet om at fremhæve idrættens pædagogiske rolle. Der kan opnås store fordele ved at etablere partnerskaber mellem undervisningssektorens og idrættens organisation Irland Som irsk medlem er jeg specielt glad for, at forslaget går ud på at dedikere 2004 til Idrættens Pædagogiske Dimension. Det falder sammen med det irske formandskab i første halvdel af 2004, og jeg kan forsikre mine kolleger Vi er alle klar over, at en række store sportsbegivenheder løber af stablen i 2004. Der afholdes EM i fodbold og olympiske og paralympiske lege i Athen. I forslaget anses 2003 for at være et forberedende år med visse aktiviteter. Irland er vært for Special Olympics i 2003, og EU's bidrag til organiseringen af disse lege er værd at rose. For atterne fra de 160 internationale delegationer verden over er dette en enestående sportslig og kulturel begivenhed.	Irland
2	Deutsch	. (EN) Das ist eine ausgezeichnete Initiative, und ich begrüße das Anliegen, den erzieherischen Wert des Sports hervorzuheben, von ganzem Herzen. Die Herstellung von Partnerschaften zwischen Sportorganisationen und Bildungsministerien ist sehr wichtig. Ich freue mich über die Unterstützung Irlands für die Olympischen Spiele 2004 in Athen. Ich möchte auch auf die geplanten Fußball-Europameisterschaften im Jahr 2004 hinweisen. Es ist eine großartige Gelegenheit für Irland, seine Leistung im Bereich des Sports zu zeigen und gleichzeitig die Beziehungen zu anderen europäischen Ländern zu verstetigen. Ich hoffe, dass Irland in Zukunft noch mehr an der Entwicklung des Sports in Europa beteiligt werden wird.	Irland
3	Deutsch	8316 . (EN) Das ist eine ausgezeichnete Initiative, und ich begrüße das Anliegen, den erzieherischen Wert des Sports hervorzuheben, von ganzem Herzen. Die Herstellung von Partnerschaften zwischen Sportorganisationen und Bildungsministerien ist sehr wichtig. Ich freue mich über die Unterstützung Irlands für die Olympischen Spiele 2004 in Athen. Ich möchte auch auf die geplanten Fußball-Europameisterschaften im Jahr 2004 hinweisen. Es ist eine großartige Gelegenheit für Irland, seine Leistung im Bereich des Sports zu zeigen und gleichzeitig die Beziehungen zu anderen europäischen Ländern zu verstetigen. Ich hoffe, dass Irland in Zukunft noch mehr an der Entwicklung des Sports in Europa beteiligt werden wird. Als irischen Abgeordneten freut es mich ganz besonders, dass das Jahr 2004 zum Jahr der Erziehung durch Sport ausgerufen werden soll. In der ersten Hälfte des Jahres 2004 wird Irland den Ratsvorsitz übernehmen, und ich freue mich sehr darüber. Wir alle wissen, dass für 2004 eine Reihe bedeutender Sportereignisse geplant sind. So werden die Fußball-Europameisterschaften sowie die Olympischen Spiele und die Paralympics in Athen stattfinden. Laut Vorschlag soll das Land Irland Gastgeber der Special Olympics im Jahr 2003 sein, und der Beitrag der EU zur Organisation dieses Ereignisses verdient es ebenfalls, hervorgehoben zu werden. Für die Sportler der 160 internationalen Delegationen wird dies eine großartige Gelegenheit sein, um ihre Leistungen zu präsentieren und gleichzeitig die Freundschaften zwischen den Ländern zu fördern.	Irland
4	Deutsch	8317 Herr Präsident, ich bin sicher, daß jedem hier bekannt ist, daß die Grünen ein grundsätzliches Problem mit der Haltung der Europäischen Union gegenüber Fischereiabkommen haben. Wir sind nicht prinzipiell gegen Fischereiabkommen eingestellt, aber wir fordern eine faire und nachhaltige Gestaltung dieser Abkommen. Als ich den Madagaskar-Bericht erstellt habe, haben wir das Fischereiabkommen mit Madagaskar gründlich geprüft und festgestellt, daß es erhebliche Mängel aufwies. Seither haben wir gegen solche Abkommen gestimmt. Ich kann Ihnen versichern, daß wir dies weiter tun werden. Das ist der Grund, warum wir - zusammen mit der EDN-Faktion und Herrn Macartney - einen Änderungsantrag eingebracht haben, der eine Halbzeitbilanz des Abkommens fordert. Wir möchten, daß in dieser Halbzeitbilanz die Mängel aufgedeckt werden. Ich möchte der Kommission eine einfache Frage stellen und hätte gerne noch heute abend eine Antwort darauf. Ich möchte gerne folgendes wissen: Wenn das Parlament heute unseren Änderungsantrag im Hinblick auf eine faire und nachhaltige Gestaltung der Fischereiabkommen annehmen sollte, was würde dann passieren?	Irland
5	Deutsch	8318 Herr Präsident, im Norden denken die Menschen, das Aids-Problem sei überstanden. Das ist es natürlich nicht, aber veränderte Lebensgewohnheiten und neue, wirksame medikamentöse Behandlungen haben uns zumindest in den Norden gebracht. Aber im Süden zeigt sich ein völlig anderes Bild. Von den fünf Millionen neuen Fällen in jedem Jahr treten 95 % in den südlichen Ländern mit niedrigen Einkommen auf. 25 Millionen Aids-Kranke leben in Afrika und sechs Millionen in Asien. Was Aids betrifft, befinden wir uns bei mittlerweile 16 Millionen Toten in einem Teufelskreis von Infektion, Unvermögen und Armut. Jährlich zwei Millionen Tote in Afrika, das ist ein Viertel aller dortigen Todesfälle. Zehn Prozent der Menschen sterben an Aids. Als Mitarbeiter des Economist in Sambia ein spezielles Krankenhaus besuchten, mussten sie sehen, dass zwei Drittel der Patienten an Aids sterben. Die Gliedmaßen der Menschen glichen den Berichten zufolge zerbrochenen Stäben. Das Gesundheitsministerium in Sambia schätzt seinerzeit ein, dass die Hälfte der Bevölkerung an Aids sterben wird. Das ist eine menschliche Katastrophe. Jetzt stehen wir in Amerika vor einer menschlichen Katastrophe. Die Weltgesundheitsorganisation schätzt, dass die Aids-Pandemie die Weltwirtschaft in die Krise bringt. Armutsförderung geht mit dem Kampf gegen Krankheit, Behinderung und Tod Hand in Hand. Wir brauchen dringend mehr Mittel für Medikamente und Impfstoffe, müssen aber auch für ihre ordnungsgemäße Verteilung an die Bevölkerung sorgen. Todbringend ist natürlich auch die Tuberkulose. Auf ihr Konto gehen jährlich zwei Millionen Todesfälle, davon 95 % in Entwicklungsländern; sie ist die Haupttodesursache bei HIV-Kranken; eine infizierte Person kann die Krankheit an andere übertragen.	England
6	Deutsch	8319 Herr Präsident, vorbehaltlich einiger weniger Änderungen unterstützt meine Fraktion den Entschließungsantrag uneingeschränkt, der uns von den Berichterstattern Herrn Tsatsos und Herrn Gil-Robles Gil-Delgado vorgelegt wurde. Wir stehen nun vor der großen Herausforderung, die Verfassung durch die Regierungskonferenz zu bekommen, ohne dass sie Stück für Stück demonstriert wird. Deshalb begrüßen wir es, dass in diesem Entschließungsantrag die Befreiung von der Verfassung vorgesehen ist. Doch es lauern Gefahren. Es gibt Regierungen, wie zum Beispiel die spanische, die das System der doppelten Mehrheit im Rat ablehnen und zu dem komplizierten, unverständlichen und wenig sinnvollen System von Nizza zurückgreifen wollen. Ich stimme auch nicht mit denjenigen überein, die einen Verweis auf die Religion in die Verfassung aufnehmen wollen. Wir sind eine Union des religiösen Pluralismus, in der verschiedene Religionen vertreten sind. In unserer Union ist die Religionsfreiheit eine wichtige Werte. Abschließend möchte ich noch auf das Thema Volkbefragung eingehen. Es steht der Europäischen Union nicht zu, ihren Mitgliedstaaten vorzuschreiben, welche internen Verfahren sie zur Ratifizierung eines Vertrags anwenden.	England

Corpus de traduceri

train_data

	language	text	label
1	dansk	Dette er et fremragende initiativ, og jeg støtter fuldt ud målet om at fremhæve idrættens pædagogiske rolle. Der kan opnås store fordele ved at etablere partnerskaber mellem undervisningssektoren og idrættens organisationer. Irland Som irsk medlem er jeg specielt glad for, at forslaget går ud på at dedikere 2004 til Idrættens Pædagogiske Dimension. Det falder sammen med det irske formandskab i første halvdel af 2004, og jeg kan forsikre mine kolleger Vi er alle klar over, at en række store sportsbegivenheder løber af stablen i 2004. Der afholdes EM i fodbold og olympiske og paralympiske lege i Athen. I forslaget anses 2003 for at være et forberedende år med visse aktiviteter. Irland er vært for Special Olympics i 2003, og EU's bidrag til organiseringen af disse lege er værd at rose. For atterne fra de 160 internationale delegationer verden over er dette en enestående sportslig og kulturel begivenhed.	Ireland
2	Deutsch	. (EN) Das ist eine ausgezeichnete Initiative, und ich begrüße das Anliegen, den erzieherischen Wert des Sports hervorzuheben, von ganzem Herzen. Die Herstellung von Partnerschaften zwischen Sportorganisationen und Bildungsinstanzen ist ein wichtiger Beitrag zu einer besseren Zusammenarbeit. Irland Als irischen Abgeordneten freut es mich ganz besonders, dass das Jahr 2004 zum Jahr der Erziehung durch Sport ausgerufen werden soll. In der ersten Hälfte des Jahres 2004 wird Irland den Ratsvorsitz übernehmen, und ich freue mich auf die Zusammenarbeit mit anderen Mitgliedsstaaten. Wir alle wissen, dass für 2004 eine Reihe bedeutender Sportereignisse geplant sind. So werden die Fußballeuropameisterschaften sowie die Olympischen Spiele und die Paralympics in Athen stattfinden. Laut Vorschlag soll das Land Irland Gastgeber der Special Olympics im Jahr 2003 sein, und der Beitrag der EU zur Organisation dieses Ereignisses verdient es ebenfalls, hervorgehoben zu werden. Für die Sportler der 160 internationalen Delegationen ist dies eine großartige Gelegenheit.	Ireland
3	español	. (EN) Ésta es una excelente iniciativa y yo suscribo plenamente el objetivo de destacar el valor educativo del deporte. Puede ganarse mucho con el establecimiento de colaboraciones entre las organizaciones deportivas y las instituciones de educación. Irlanda Como diputado procedente de Irlanda, me complace particularmente que esta propuesta dedique el año 2004 a la Educación a través del Deporte. Esto coincidirá con la presidencia irlandesa de la Unión en el primer semestre de 2004. Los países miembros sabemos que para 2004 habrá una serie de eventos deportivos importantes. Se van a celebrar el Campeonato Europeo de Fútbol y los Juegos Olímpicos y Paralímpicos de Atenas. La propuesta prevé que Irlanda será la anfitriona de los Juegos Olímpicos Especiales en 2003, y es digna de alabanza también la aportación de la UE a la organización de este acontecimiento. Será una experiencia cultural y deportiva única para los 160 delegados internacionales.	Ireland
4	español	Señor Presidente, seguro que todos los presentes saben que el enfoque de la Unión Europea por lo que respecta a los Acuerdos pesqueros plantea algunos problemas fundamentales para los Verdes. No somos contrarios a la pesca sostenible. Irlanda Cuando elaboré el informe sobre Madagascar, realizamos un análisis detallado del Acuerdo pesquero con dicho país y detectamos serias insuficiencias en el mismo. Desde entonces hemos votado en contra de dichos Acuerdos pesqueros en la Asamblea. Irlanda Por esto hemos presentado, conjuntamente con el Grupo I-EDN y con el Sr. Macartney, una enmienda por la que se solicita una evaluación intermedia del Acuerdo. Nuestro deseo es que esta revisión examine el impacto de la pesca sostenible en Madagascar. Irlanda Yo quisiera plantear una pregunta muy sencilla a la Comisión y desearía recibir una respuesta esta noche. Mi pregunta es: ¿aceptaría la Comisión una revisión de este tipo si el Parlamento aprueba nuestra enmienda por la que se solicita una evaluación intermedia del Acuerdo?	Ireland
5	español	Señor Presidente, en el Norte la gente piensa que en cuanto al SIDA todo se ha acabado. Evidentemente no es así, pero los cambios en los estilos de vida y la llegada de tratamientos con fármacos eficaces nos han ayudado. Irlanda Pero en el Sur, la imagen es muy distinta. El noventa y cinco por ciento de los cinco millones de nuevos casos anuales se dan en países del Sur con baja renta. Hay 25 millones de personas que padecen el SIDA en África y 600 mil en Asia. Irlanda En lo que respecta al SIDA nos enfrentamos a un círculo vicioso de infección, incapacidad y pobreza que acompaña a los 16 millones de muertes a las que ya hemos asistido: dos millones de muertos al año en África, una cifra similar en Asia. Irlanda Cuando The Economist estuvo en Zambia y analizó un hospital en concreto, comprobó que dos tercios de sus pacientes morían de SIDA. Según contaban, los miembros de estas personas parecían palos de escoba rotos. Irlanda En Zambia el departamento de sanidad estimaba entonces que la mitad de la población moriría de SIDA. Esto es un desastre humano. Estamos asistiendo ahora a desastres humanos en EE.UU. El mundo está pendiente de lo que ocurrirá en Zambia. Irlanda La reducción de la pobreza está estrechamente relacionada con la guerra contra la enfermedad, la discapacidad y la muerte. Necesitamos desesperadamente más recursos para fármacos y vacunas, pero debemos garantizar que estos recursos lleguen a las personas más necesitadas. Irlanda La tuberculosis, por supuesto, también mata. A ella se deben dos millones de muertes anuales, el 95% de éstas en países en desarrollo; es la causa principal de muerte en los casos de VIH; una persona infectada puede contagiar a otras.	England
6	español	Señor Presidente, aunque con unas pocas enmiendas, mi Grupo apoya plenamente la propuesta de resolución que se nos ha presentado, y que han elaborado con tanto acierto los ponentes, el Sr. Tsatsos y el Sr. Gil-Robles. Irlanda El gran reto a que nos enfrentamos ahora es que la Constitución sea aprobada en la CIG sin desmantelarla pieza a pieza. Por este motivo, nos complace que esta propuesta de resolución que adoptaremos no pida cambios drásticos. Irlanda Pero nos acechan algunos peligros. Algunos gobiernos -España, por ejemplo- se oponen al sistema de doble mayoría en el Consejo y quieren volver al sistema de Niza, que es un sistema complejo, difícil de entender y no muy transparente. Irlanda También me opongo a quienes desean hacer referencia a la religión en la Constitución. Somos una Unión de diversas religiones, de pluralismo religioso. En la Unión hay personas que no son religiosas. En la Constitución no debe haber referencias a la religión. Irlanda Finalmente, quiero tratar el tema del referendo. No está en manos de la Unión Europea decir a sus Estados miembros qué procedimiento deberían utilizar internamente para ratificar un tratado constitucional. Es un asunto que	England

Corpus de traduceri

		train_data	label
1	language	text	
2	dansk	Dette er et fremragende initiativ, og jeg støtter fuldt ud målet om at fremhæve idrættens pædagogiske rolle. Der kan opnås store fordele ved at etablere partnerskaber mellem undervisningssektorens og idrættens organisationer. Ireland Som irsk medlem er jeg specielt glad for, at forslaget går ud på at dedikere 2004 til Idrættens Pædagogiske Dimension. Det falder sammen med det irske formandskab i første halvdel af 2004, og jeg kan forsikre mine kolleger, at vi alle klar over, at en række store sportsbegivenheder løber af stablen i 2004. Der afholdes EM i fodbold og olympiske og paralympiske lege i Athen. I forslaget anses 2003 for at være et forberedende år med visse aktiviteter. Irland er vært for Special Olympics i 2003, og EU's bidrag til organiseringen af disse lege er værd at rose. For atterne fra de 160 internationale delegationer verden over er dette en enestående sportslig og kulturel begivenhed.	Ireland
3	Deutsch	. (EN) Das ist eine ausgezeichnete Initiative, und ich begrüße das Anliegen, den erzieherischen Wert des Sports hervorzuheben, von ganzem Herzen. Die Herstellung von Partnerschaften zwischen Sportorganisationen und Bildungsinstanzen ist ein wichtiger Beitrag zur Entwicklung des Sports in Irland. Als irischen Abgeordneten freut es mich ganz besonders, dass das Jahr 2004 zum Jahr der Erziehung durch Sport ausgerufen werden soll. In der ersten Hälfte des Jahres 2004 wird Irland den Ratsvorsitz übernehmen, und ich kann mich auf die Arbeit mit dem irischen Präsidenten freuen. Wir alle wissen, dass für 2004 eine Reihe bedeutender Sportereignisse geplant sind. So werden die Fußballeuropameisterschaften sowie die Olympischen Spiele und die Paralympics in Athen stattfinden. Laut Vorschlag soll das Land Irland Gastgeber der Special Olympics im Jahr 2003 sein, und der Beitrag der EU zur Organisation dieses Ereignisses verdient es ebenfalls, hervorgehoben zu werden. Für die Sportler der 160 internationalen Delegationen ist dies eine großartige Gelegenheit. Ireland	Ireland
4	español	. (EN) Ésta es una excelente iniciativa y yo suscribo plenamente el objetivo de destacar el valor educativo del deporte. Puede ganarse mucho con el establecimiento de colaboraciones entre las organizaciones deportivas y las instituciones educativas. Irlanda es un país que apuesta por el deporte y su desarrollo social. Como diputado procedente de Irlanda, me complace particularmente que esta propuesta dedique el año 2004 a la Educación a través del Deporte. Esto coincidirá con la presidencia irlandesa de la Unión en el primer semestre de 2004. Todos somos conscientes de que en 2004 tendrá lugar una serie de eventos deportivos importantes. Se van a celebrar el Campeonato Europeo de Fútbol y los Juegos Olímpicos y Paralímpicos de Atenas. La propuesta prevé que Irlanda sea la anfitriona de los Juegos Olímpicos Especiales en 2003, y es digna de alabanza también la aportación de la UE a la organización de este acontecimiento. Será una experiencia cultural y deportiva única para los países europeos. Irlanda	Irlanda
5	italiano	Si tratta di un'eccellente iniziativa e condivido totalmente l'obiettivo di mettere in rilievo il valore educativo dello sport. Costruire un partenariato tra organizzazioni sportive e istituti scolastici può dare molti frutti. In qualità di deputato irlandese sono particolarmente lieto della proposta di designare il 2004 l'Anno europeo dell'educazione tramite lo sport perché andrà a coincidere con la Presidenza irlandese dell'Unione nella prima metà dell'anno. Come è noto nel corso del 2004 si svolgeranno importanti manifestazioni sportive: il Campionato mondiale di calcio e i Giochi olimpici e paralimpici ad Atene. La proposta prevede che il 2003 sia un anno preparatorio nel quale Irlanda ospiterà le Olimpiadi speciali del 2003 e il contributo dell'Unione europea all'organizzazione di tale evento è degno di nota. Si tratterà di un'esperienza culturale e sportiva unica per gli atleti partecipanti provenienti da tutti i paesi europei. Irlanda	Irlanda
6	italiano	Signor Presidente, tutti i presenti sanno certamente che i verdi non condividono l'operato dell'Unione europea in materia di accordi di pesca. Non siamo contrari per principio agli accordi di pesca ma avanziamo forti riserve sulla loro efficacia. Quando mi occupavo della relazione sul Madagascar, avevamo fatto uno studio dettagliato sull'accordo di pesca con tale paese e lo avevamo trovato gravemente carente. Da allora abbiamo votato contro tali accordi. L'accordo di pesca con il Madagascar è stato respinto. Per questo motivo, assieme al gruppo EDN e all'onorevole Macartney, abbiamo proposto un emendamento inteso ad ottenere una revisione intermedia dell'accordo. Vorremmo che essa esaminasse l'impatto della pesca dell'Unione europea sulle attività peschistiche del Madagascar. Desidero rivolgere una semplice domanda alla Commissione e mi auguro che mi venga fornita una risposta questa sera. La domanda è la seguente: qualora il Parlamento approvasse il nostro emendamento a favore di una revisione intermedia dell'accordo di pesca con il Madagascar, quale sarebbe il vostro voto? Irlanda	Irlanda
7	italiano	Signor Presidente, nel Nord del mondo si pensa che l'AIDS sia un problema ormai risolto; naturalmente non è così, ma i cambiamenti nelle abitudini di vita e l'introduzione di nuovi farmaci ci hanno almeno consentito di compiere progressi importanti. Nel Sud del mondo però il quadro è completamente differente: il 95 percento dei cinque milioni di nuovi casi che si registrano ogni anno si verificano nei paesi a basso reddito del Sud. In Africa vi sono 25 milioni di persone affette da AIDS. Nel caso dell'AIDS ci troviamo di fronte a un circolo vizioso di infezione, incapacità e povertà che spiega i sedici milioni di morti che già dobbiamo lamentare; in Africa, in particolare, si contano due milioni di morti all'anno, ossia più di quelli causati dalla tubercolosi. Quando gli inviati di The Economist nello Zambia hanno visitato un ospedale, hanno constatato che due terzi dei pazienti stavano morendo di AIDS; secondo la testimonianza dei giornalisti, le membra delle persone ricoverate erano talmente gonfie da essere difficili da riconoscere. Secondo le stime allora avanzate dal Ministero della sanità dello Zambia, è possibile che metà della popolazione del paese muoia di AIDS; si tratta di una terribile tragedia umanitaria. Assistiamo ora ad analoghe tragedie anche in altri paesi. La riduzione della povertà si accompagna alla lotta contro le malattie, l'invalidità e la morte. Abbiamo un disperato bisogno di maggiori risorse per farmaci e vaccini, ma dobbiamo anche assicurarne l'adeguata distribuzione fra i paesi più poveri. Anche la tubercolosi è naturalmente un flagello mortale: miete due milioni di vittime l'anno, il 95 percento delle quali nei paesi in via di sviluppo, ed è la causa principale delle morti da HIV poiché una persona infetta può trasmettere il virus a molti altri. England	England
8	italiano	Signor Presidente, fatti salvi alcuni emendamenti, il mio gruppo sostiene pienamente la proposta di risoluzione presentata ed abilmente elaborata dai relatori, onorevoli Tsatsos e Gil-Robles Gil-Delgado. La grande sfida che si presenta ora è far passare la Costituzione attraverso la CIG senza che venga demolita pezzo per pezzo. Questo è il motivo per cui siamo compiaciuti del fatto che la proposta di risoluzione che adotteremo sarà molto più completa e meno frammentaria. Vi sono pericoli in agguato. Alcuni governi - la Spagna, per esempio - si oppongono al sistema della doppia maggioranza in seno al Consiglio e vogliono tornare al sistema di Nizza, che è complesso, incomprensibile e non molto trasparente. Sono anche in disaccordo con coloro che vogliono includere nella Costituzione un riferimento alla religione. Siamo un'Unione di diverse religioni, di pluralismo religioso. Nell'Unione vi sono persone che non sono religiose. Non crediamo che la Costituzione debba avere nulla a che fare con la religione. Infine, la questione del referendum. Non spetta all'Unione europea dire ai suoi Stati membri quale procedura seguire internamente per la ratifica di un Trattato che istituisce la Costituzione. E' una decisione che devono prendere i singoli Stati membri. England	England

Corpus de traduceri

			train_data	label
1	language	text		
2	dansk	Dette er et fremragende initiativ, og jeg støtter fuldt ud målet om at fremhæve idrættens pædagogiske rolle. Der kan opnås store fordele ved at etablere partnerskaber mellem undervisningssektoren og idrættens organisation Ireland Som irsk medlem er jeg specielt glad for, at forslaget går ud på at dedikere 2004 til Idrættens Pædagogiske Dimension. Det falder sammen med det irske formandskab i første halvdel af 2004, og jeg kan forsikre mine kolleger Vi er alle klar over, at en række store sportsbegivenheder løber af stablen i 2004. Der afholdes EM i fodbold og olympiske og paralympiske lege i Athen. I forslaget anses 2003 for at være et forberedende år med visse aktiviteter. Irland er vært for Special Olympics i 2003, og EU's bidrag til organiseringen af disse lege er værd at rose. For atterne fra de 160 internationale delegationer verden over er dette en enestående sportslig og kulturel begivenhed.	Ireland	
3	Deutsch	8316 . (EN) Das ist eine ausgezeichnete Initiative, und ich begrüße das Anliegen, den erzieherischen Wert des Sports hervorzuheben, von ganzem Herzen. Die Herstellung von Partnerschaften zwischen Sportorganisationen und Bil	Ireland	
		Als irischen Abgeordneten freut es mich ganz besonders, dass das Jahr 2004 zum Jahr der Erziehung durch Sport ausgerufen werden soll. In der ersten Hälfte des Jahres 2004 wird Irland den Ratsvorsitz übernehmen, und ich		
		Wir alle wissen, dass für 2004 eine Reihe bedeutender Sportereignisse geplant sind. So werden die Fußballeuropameisterschaften sowie die Olympischen Spiele und die Paralympics in Athen stattfinden. Laut Vorschlag soll das		
		Irland wird Gastgeber der Special Olympics im Jahr 2003 sein, und der Beitrag der EU zur Organisation dieses Ereignisses verdient es ebenfalls, hervorgehoben zu werden. Für die Sportler der 160 internationalen Delegationen		
	español	8317 . (EN) Ésta es una excelente iniciativa y yo suscribo plenamente el objetivo de destacar el valor educativo del deporte. Puede ganarse mucho con el establecimiento de colaboraciones entre las organizaciones deportivas y la	Ireland	
		Como diputado procedente de Irlanda, me complace particularmente que esta propuesta dedique el año 2004 a la Educación a través del Deporte. Esto coincidirá con la presidencia irlandesa de la Unión en el primer semestre		
		Todos somos conscientes de que en 2004 tendrá lugar una serie de eventos deportivos importantes. Se van a celebrar el Campeonato Europeo de Fútbol y los Juegos Olímpicos y Paralímpicos de Atenas. La propuesta prevé		
		Irlanda será la anfitriona de los Juegos Olímpicos Especiales en 2003, y es digna de alabanza también la aportación de la UE a la organización de este acontecimiento. Será una experiencia cultural y deportiva única para los		
4	italiano	16631 24944 Si tratta di un'eccellente iniziativa e condivido totalmente l'obiettivo di mettere in rilievo il valore educativo dello sport. Costruire un partenariato tra organizzazioni sportive e istituti scolastici può dare molti frutti.	Ireland	
		In qualità di deputato irlandese sono particolarmente lieto della proposta di designare il 2004 l'Anno europeo dell'educazione tramite lo sport perché andrà a coincidere con la Presidenza irlandese dell'Unione nella prima metà		
		Come è noto nel corso del 2004 si svolgeranno importanti manifestazioni sportive: il Campionato mondiale di calcio e i Giochi olimpici e paralimpici ad Atene. La proposta prevede che il 2003 sia un anno preparatorio nel quale		
		L'Irlanda ospiterà le Olimpiadi speciali del 2003 e il contributo dell'Unione europea all'organizzazione di tale evento è dunque da nota. Si tratterà di un'esperienza culturale e sportiva unica per gli atleti partecipanti provenienti da		
	Nederland	8318 33258 Dit is een uitstekend initiatief en ik sta volledig achter de doelstelling om speciaal de aandacht te vestigen op de educatieve waarde van de sport. Er valt veel te winnen door partnerschappen tussen sportorganisaties en onderwijsinstellingen.	Ireland	
		Als parlementslid uit Ierland ben ik bijzonder blij met dit voorstel om in het jaar 2004 speciaal aandacht te schenken aan opvoeding door sport. Dit valt in de eerste helft van het jaar samen met het eerste voorzitterschap van de		
		Wij weten allen dat er in 2004 een aantal belangrijke sportevenementen zullen plaatsvinden, zoals de Europese voetbalkampioenschappen en ook de Olympische Spelen en de Paralympics in Athene. Volgens het voorstel wordt Ierland zal in 2003 gastheer zijn voor de Speciale Olympische Spelen in 2003. De bijdrage van de EU aan de organisatie van dit evenement verdient ook onze lof. Dit belooft voor de deelnemende sportlieden, die uitkomen in		
5	Nederland	16632 33259 Mijnheer de Voorzitter, ik weet zeker dat iedereen hier weet dat de Groenen fundamentele problemen hebben met de manier waarop de Europese Unie visserijovereenkomsten benadert. Wij zijn niet tegen visserijovereenkomsten.	Ireland	
		Toen ik aan het verslag over Madagascar werkte, hebben wij een gedetailleerde studie over de visserijovereenkomst met Madagascar gemaakt, en deze overeenkomst kreeg een dikke onvoldoende. Sedertdien hebben we te maken met de Franse overheid.		
		Daarom hebben wij, samen met de Fractie van onafhankelijk voor het Europa van de Nationale Staten en de heer Macartney, een amendement voorgesteld waarin gevraagd wordt om halverwege de looptijd van de overeenkomst een evaluatieverslag te leveren.		
		Ik heb een zeer simpele vraag voor de Commissie en zou graag vanavond een antwoord willen hebben. Ik zou willen weten of, als het Parlement ons amendement voor een evaluatieverslag halverwege de looptijd van de overeenkomst goedkeurt,		
	Nederland	16633 33260 Mijnheer de Voorzitter, in het Noorden denken mensen dat aids alweer achter de rug is. Dat is natuurlijk niet het geval. Andere leefwijzen en de komst van doeltreffende therapieën hebben er echter in ieder geval toe bijgedragen.	England	
		In het Zuiden is het beeld evenwel heel anders. Van de 5 miljoen nieuwe gevallen die er jaarlijks bij komen, nemen lage-inkomenslanden in het Zuiden 95 procent voor hun rekening. Er leven 25 miljoen mensen met aids in Afrika.		
		Waar het aids betreft, worden wij geconfronteerd met een vicieuze cirkel van infectie, arbeidsongeschiktheid en armoede. Daarnaast zijn er ook nog de 16 miljoen doden die we al gezien hebben: 2 miljoen doden per jaar in Afrika.		
		Toen verslaggevers van The Economist een bepaald ziekenhuis in Zambia bezochten, ontdekten zij dat tweederde van de patiënten aldaar aan aids stierven. De ledematen van mensen, zo meldden zij, zagen eruit als gebrokken stokken.		
		In Zambia schatte het ministerie van Volksgezondheid destijds dat de helft van de bevolking aan aids zou sterven. Dat is een humanitaire ramp. We worden op dit moment met humanitaire rampen in Amerika geconfronteerd.		
		Armoedebestrijding gaat hand in hand met de oorlog tegen ziekte, invaliditeit en dood. We hebben dringend meer geld nodig voor geneesmiddelen en vaccins. We moeten er echter ook voor zorgen dat deze op geschikte wijze kunnen worden gebruikt.		
	Nederland	24947 33261 Tuberculose is uiteraard ook moordend. Deze ziekte is verantwoordelijk voor 2 miljoen doden per jaar, waarvan 95 procent in ontwikkelingslanden. Zij is de belangrijkste doodsoorzaak bij HIV-gedetecteerden. Eén besmette persoon kan honderden anderen besmetten.	England	
		Mijnheer de Voorzitter, behoudens enkele amendementen staat mijn fractie vierkant achter de ontwerpresolutie die de rapporteurs, de heer Tsatsos en de heer Gil-Robles Gil-Delgado, ons hebben voorgelegd en die zij zo voorbereid hebben.		

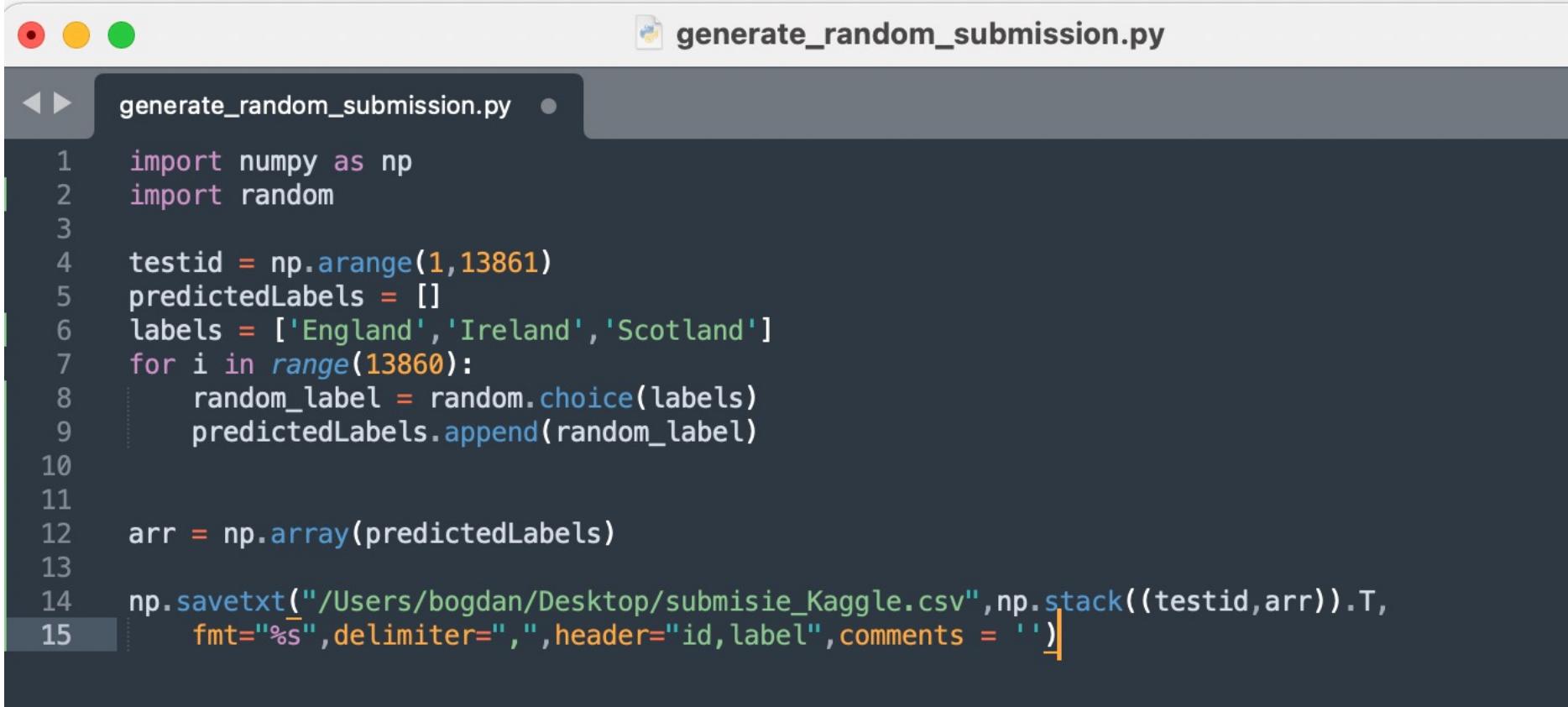
Date proiect

- Antrenare:
 - 41570 exemple de antrenare = texte cu traduceri, ordonate câte 8314 în funcție de limba traducerii (limba traducerii este etichetată)
 - 41570 exemple de antrenare = $8314 \text{ texte traduse} \times 5 \text{ limbi străine}$ (daneză, germană, spaniolă, italiană, olandeză)
 - fiecare exemplu de antrenare este etichetat cu una din clasele {England, Scotland, Ireland}
- Testare:
 - 13860 exemple de testare = $2772 \text{ texte traduse} \times 5 \text{ limbi străine}$ (daneză, germană, spaniolă, italiană, olandeză), dar nu sunt ordonate iar limba traducerii nu este etichetată
 - pentru fiecare exemplu de testare trebuie să preziceți eticheta/clasa
 - public test (40%) vs private test (60%)

Platforma Kaggle

- pentru participare la proiect
 - vă faceți cont pe platforma www.kaggle.com
 - puteți participa numai pe bază de invitație accesând link-ul de mai jos:
<https://www.kaggle.com/t/57190c0bbefe46338fc310f2232edf3f>
 - dați Join Competition și apoi puteți participa încărcând predicțiile voastre.
- **2 submisii/zi** (sub forma unui fișier CSV – demo)
- puteți să vă alegeti la final cele mai bune două submisii pe care le considerați voi
- public test (40%) vs. private test (60%)
- termen limită duminică, 13 noiembrie, 23:59
- prezentare proiecte în săptămâna 14-18 noiembrie

Demo – submisie Kaggle



The screenshot shows a code editor window titled "generate_random_submission.py". The file contains the following Python code:

```
1 import numpy as np
2 import random
3
4 testid = np.arange(1,13861)
5 predictedLabels = []
6 labels = ['England', 'Ireland', 'Scotland']
7 for i in range(13860):
8     random_label = random.choice(labels)
9     predictedLabels.append(random_label)
10
11
12 arr = np.array(predictedLabels)
13
14 np.savetxt("/Users/bogdan/Desktop/submisie_Kaggle.csv", np.stack((testid,arr)).T,
15             fmt="%s", delimiter=",", header="id,label", comments = '')
```

The code uses NumPy and the random module to generate a CSV file named "submisie_Kaggle.csv" for a Kaggle competition. The CSV file contains two columns: "id" and "label". The "id" column ranges from 1 to 13860, and the "label" column is randomly assigned from the list ["England", "Ireland", "Scotland"] for each row.

Sistem de notare proiect

- proiectul valorează 1.5 puncte din nota finală
- partea de concurs Kaggle = 1 punct
 - locul 1 = 1 punct
 - locul 2 = 0.99 puncte
 - ...
 - Locul 50 = 0.51 puncte
 - locul 51+ = 0.5 puncte (cât timp aveți o performanță > baseline)
- partea de documentație + prezentare = 0.5 puncte
- important: pentru fiecare submisie să știți ce ați făcut (cod Python, parametri, etc). La final vă veți alege 2 submisii care credeți voi că sunt cele mai bune (pot fi același model cu parametri diferiți)

Documentație proiect - pdf

- descrieți în detaliu (1-2 pagini) **2 modele diferite** folosite (kNN, Naïve Bayes, SVM, Rețea Neuronală):
 - ce caracteristici folosiți;
 - care sunt parametri, hiperparametri modelului;
 - cum antrenați parametri/hiperparametri;
 - cât durează antrenarea;
 - ce performanță ati obținut pe cele 40% de date din setul de date de test public pe Kaggle;
- **pentru un singur model prezențați rezultatele în urma antrenării în maniera 5 fold cross-validation (cursul de azi) + matricea de confuzie asociată**

Predare proiect

- predarea proiectului înseamnă trimiterea documentației și a codului Python pentru fiecare submisie
- trimiteți la adresa de email: ub.fmi.cti.ia@gmail.com un email până luni, 14 noiembrie, ora 23:59 cu următoarele fișiere:
 - un fișier pdf cu documentația voastră
 - două fișiere python cu codul pentru submisiile voastre
 - respectați formatul de mai jos



361_Alexe_Bogdan_documentatie.pdf



361_Alexe_Bogdan_submisie1_cod.py



361_Alexe_Bogdan_submisie2_cod.py

Prezentare proiect

- este individuală, are loc în săptămâna 7 (14-18 noiembrie)
- constă într-o discuție cu Alexandra/Sergiu/Bogdan de maxim 10 minute
 - prezentarea voastră 3-5 minute (ce modele ați folosit la cele 2 submisii)
 - 3-5 minute întrebări din partea noastră
- vom face o programare pe care o vom afișa din timp
- încercăm să ne încadrăm în timpul orelor de laborator/proiect

Regulament de integritate

- regulament privind activitatea studenților la UB:
http://fmi.unibuc.ro/ro/pdf/2015/consiliu/UB_Regulament_studenti_2015.pdf
- regulament de etică și profesionalism la
FMI:http://fmi.unibuc.ro/ro/pdf/2015/consiliu/Regulament_etica_FMI.pdf

- Se consideră **incident minor** cazul în care un student/ o studentă:
- a. preia codul sursă/ rezolvarea unei teme de la un coleg/ o colegă și pretinde că este rezultatul efortului propriu;
 - nu copiați codul de la colegi – veți primi 0 puncte + referat de incident minor

Project – laborator

Laborator – modelul BOW

1. Modelul Bag-of-Words (BOW)

Modelul Bag-of-words este o reprezentare simplificată a textelor folosită în procesarea limbajului natural și în regăsirea informației. Conform acestui model, reprezentăm un text numărând de câte ori apare un cuvânt dintr-un anumit dicționar în textul respectiv. Prin folosirea unei asemenea reprezentări, informația legată de ordinea cuvintelor, gramatică, topică, sensul cuvintelor se pierde.

Laborator – modelul BOW

Exemplu¹: considerăm textele 1 și 2 de mai jos în limba engleză.

- (1) John likes to watch movies. Mary likes movies too.
- (2) John also likes to watch football games.

Eliminând semnele de punctuație obținem listele de cuvinte pentru cele două texte:

```
"John", "likes", "to", "watch", "movies", "Mary", "likes", "movies", "too"  
"John", "also", "likes", "to", "watch", "football", "games"
```

Considerăm dicționarul D format din reuniunea tuturor cuvintelor din cele 2 texte:

```
D = {"John", "likes", "to", "watch", "movies", "Mary", "too", "also",  
"football", "games"}
```

Laborator – modelul BOW

Considerăm dicționarul D format din reuniunea tuturor cuvintelor din cele 2 texte:

```
D = {"John", "likes", "to", "watch", "movies", "Mary", "too", "also",
      "football", "games"}
```

Reprezentăm fiecare text numărând de câte ori apare fiecare cuvânt din dicționarul D în fiecare text. Obținem reprezentările bag-of-words următoare:

```
BoW1 = {"John":1, "likes":2, "to":1, "watch":1, "movies":2, "Mary":1,
         "too":1, "also":0, "football":0, "games":0};
```

```
BoW2 = {"John":1, "likes":1, "to":1, "watch":1, "movies":0, "Mary":0,
         "too":0, "also":1, "football":1, "games":1};
```

Laborator – modelul BOW

- (1) John likes to watch movies. Mary likes movies too.
- (2) John also likes to watch football games.

Considerăm dicționarul D format din reuniunea tuturor cuvintelor din cele 2 texte:

```
D = {"John", "likes", "to", "watch", "movies", "Mary", "too", "also",
      "football", "games"}
```

Reprezentăm fiecare text numărând de câte ori apare fiecare cuvânt din dicționarul D în fiecare text. Obținem reprezentările bag-of-words următoare:

```
BoW1 = {"John":1, "likes":2, "to":1, "watch":1, "movies":2, "Mary":1,
         "too":1, "also":0, "football":0, "games":0};
```

```
BoW2 = {"John":1, "likes":1, "to":1, "watch":1, "movies":0, "Mary":0,
         "too":0, "also":1, "football":1, "games":1};
```

Pentru dicționarul D fixat, reprezentările se pot scrie sub forma de vectori de frecvențe:

```
v1 = [1, 2, 1, 1, 2, 1, 1, 0, 0, 0];
v2 = [1, 1, 1, 1, 0, 0, 0, 1, 1, 1];
```

Laborator – modelul BOW

Pentru dicționarul D fixat, reprezentările se pot scrie sub forma de vectori de frecvențe:

```
v1 = [1, 2, 1, 1, 2, 1, 1, 0, 0, 0];  
v2 = {1, 1, 1, 1, 0, 0, 0, 1, 1, 1};
```

Suma elementelor fiecarui vector reprezintă numărul de cuvinte din text. De obicei, pentru probleme de clasificare în care încercăm să discriminăm între texte de diferite lungimi se folosesc reprezentări normalize. Spre exemplu, folosind norma L_1 (suma absolută a elementelor unui vector) obținem vectorii normalizați L_1 :

$$v1_{L1} = \frac{v1}{\|v1\|_1} = \left[\frac{1}{9}, \frac{2}{9}, \frac{1}{9}, \frac{1}{9}, \frac{2}{9}, \frac{1}{9}, \frac{1}{9}, 0, 0, 0 \right]$$

$$v2_{L1} = \frac{v2}{\|v2\|_1} = \left[\frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, 0, 0, 0, \frac{1}{7}, \frac{1}{7}, \frac{1}{7} \right]$$

Folosind norma L_2 (norma Euclidiană) obținem vectorii normalizați L_2 :

$$v1_{L2} = \frac{v1}{\|v1\|_2} = \left[\frac{1}{\sqrt{13}}, \frac{2}{\sqrt{13}}, \frac{1}{\sqrt{13}}, \frac{1}{\sqrt{13}}, \frac{2}{\sqrt{13}}, \frac{1}{\sqrt{13}}, \frac{1}{\sqrt{13}}, 0, 0, 0 \right]$$

$$v2_{L2} = \frac{v2}{\|v2\|_2} = \left[\frac{1}{\sqrt{7}}, \frac{1}{\sqrt{7}}, \frac{1}{\sqrt{7}}, \frac{1}{\sqrt{7}}, 0, 0, 0, \frac{1}{\sqrt{7}}, \frac{1}{\sqrt{7}}, \frac{1}{\sqrt{7}} \right]$$