# Temă practică – Învăţare automată 2023 Fall

*Deadline: end-of-day 12 ianuarie 2024*

Studiaţi, din punct de vedere teoretic şi experimental, gradul de adaptare/adecvare al algoritmilor studiaţi până acum în raport cu problema de clasificare a email-urilor spam pentru setul de date Ling-Spam:
http://www.aueb.gr/users/ion/data/lingspam_public.tar.gz.

**Cerinţe:**

- Înţelegeţi setul de date (care sunt atributele, etichetele, cum le extrageţi din reprezentarea textuală) — este necesar ca această cerinţă să fie, pe scurt, documentată sub forma procesării datelor înainte de introducerea în algoritm. Mesajele de tip spam conţin acest indiciu în titlul fişierului (sub forma prefixului "*spm*"). Utilizaţi, în cele ce urmează, 9 foldere (de la part1 la part9) pentru antrenare şi păstraţi câte unul pentru testare (cel intitulat part10), din fiecare dintre categoriile *lemn*, *bare*, *stop*, *lemn_stop*.
- Selectaţi şi implementaţi un algoritm, dintre cei învăţaţi, pe care îl consideraţi adecvat rezolvării acestei probleme. (**0.75** puncte)
- Justificaţi într-un raport LaTeX alegerea făcută, din punct de vedere teoretic şi experimental, atât în mod individual, cât şi prin comparaţie cu ceilalţi algoritmi candidaţi la tipul de problemă studiată. (**0.75** puncte)
- Implementaţi strategia de cross-validare Leave-One-Out şi ataşaţi raportului un grafic care să ilustreze statistic rezultatele. (**0.25** puncte)
- Adăugaţi la raport un grafic care să dovedească performanţa algoritmului vostru pe setul de date de testare, din punct de vedere al acurateţii obţinute. Acurateţea obţinută trebuie să fie relevantă, moderat mai bună decât orice strategie trivială (dat cu banul sau ales mereu aceeaşi clasa). Dacă aţi testat mai mulţi algoritmi, includeţi grafice comparative, atât prin raportare la această cerinţă, cât şi la cea precedentă. (**0.25** puncte)
- Explicaţi şi alte detalii ale experimentului care vi se par relevante, în cuvinte sau în mod grafic. Puteţi cerceta şi variante îmbunătăţite faţă de varianta clasică a algoritmului, studiată la seminar, spre a le implementa şi a spori acurateţea.

***Bonus:*** Presupunând că unele date nu ar fi etichetate cu titlul de spam/non-spam, ci ar fi disponibile doar reprezentările textuale fără a fi clasificate în prealabil, implementaţi şi descrieţi un mod prin care aţi putea preprocesa datele, tot prin strategii cunoscute, astfel încât să le introduceţi apoi în cadrul aceluiaşi algoritm implementat, nemodificat. Scopul acestui exerciţiu este să găsiţi un mod de a folosi eşantioane ("sample"-uri) fără etichetă pentru a îmbunătăţi performanţa algoritmului. Pe lângă descrierea efectivă a pasului de preprocesare, trebuie să includeţi în raport grafice pentru eroarea/acurateţea la CVLOO şi acurateţea la testare şi pentru această metodă. Consideraţi ca fiind neetichetate datele din folderele part1 şi part2 ale fiecărei categorii (ignoraţi dacă fişierul are sau nu prefixul "*spm*"). (**0.5** puncte)

---

Lucru în echipe de maxim 2 persoane.
Punctaj total: **2** puncte (temă) + **0.5** puncte (bonus)