

Tema 1 – Clasificare mail-uri spam/non-spam

Parte practică – 10p

1. Procesarea datelor

- Verificați structura setului de date.
- Verificați și eliminați duplicatele din dataset.
- Eliminați coloanele inutile.
- Encodați variabila țintă.
- Vizualizați distribuția variabilei țintă.
- Implementați și aplicați metode de curățare a textului peste caracteristicile rămase.
- Transformați textul curățat în reprezentare numerică.

2. Antrenarea și evaluarea modelelor

- Separați caracteristicile (X) de variabila țintă (y).
- Împărțiți datele în seturi de antrenament și testare.
- Antrenați mai multe modele de clasificare cu diferite combinații de hiperparametri.
- Evaluați fiecare tip de model cu cea mai bună combinație de hiperparametri.
- Afișați matricea de confuzie pentru fiecare tip de model cu cea mai bună combinație de hiperparametri sub formă de grafic.

Parte explicativă – 5p

- Explicați raționamentul din spatele fiecărui pas implementat. (de ce ați eliminat anumite coloane, de ce ați folosit vectorizarea textului, etc.)
- **HINT:** Puteți crea caracteristici noi și să analizați aceste caracteristici în raport cu variabila țintă.

Observații

- Puneți comentarii înainte de fiecare cerință în cod.
- Partea explicativă trebuie să fie în format PDF, Word (.docx) sau Markdown (fișier separat sau direct în notebook după implementarea codului).
- Tema se va încarca sub formă de arhivă **zip** care să conțină codul și partea explicativă, sau doar notebook-ul **.ipynb**. Numele fișierului încărcat să fie de forma: **Nume_Prenume_Grupă**.

- Folosiți **doar** setul de date oferit.
- Puteți implementa pași extra asupra metodei de rezolvare propusă.
- Deadline: 30.11.2025 23:59