Cerințe laborator Analiza Datelor / Învățare Automată

1 Proiect Învățare automată

Datele se iau de pe unul din site-urile:

- **OpenML** (https://www.openml.org/);
- **Kaggle** (https://www.kaggle.com/datasets);
- **Data.gov** (https://data.gov/);
- **Advaneo** (https://www.advaneo-datamarketplace.de/database/en/);
- Alte site-uri cu baze de date (baza de date trebuie să fie gratuită, disponibilă spre descărcare la data predării proiectului, iar linkul trebuie să fie pus obligatoriu în cadrul documentației).

Proiectul este individual, fiecare este responsabil de munca sa proprie.

Prezența se va face la începutul laboratorului și la finalul acestuia, cine nu este fizic în sală la oricare dintre cele două, va fi considerat absent. Vă rog insistent să nu întârziați. Celor întârziați nu li se pune prezență.

2 Cerințele proiectului, consecințele ne-predării la timp

La predare, emailul va conține:

1. Link Github/Gitlab cu repo-ul proiectului. Este important să se vadă timeline-ul comit-urilor. Proiectele făcute într-un timp scurt se depunctează.
2. În repo va exista proiectul în sine, implementat în limbaj de programare (ex: python + pandas + scikit-learn / tensorflow), sau în medii de dezvoltare (ex: Rapid Miner).
3. În repo va exista documentația de proiect în unul din formatele: Latex, Markdown, HTML, MS Word / LibreOffice / OpenOffice.
4. Prezentarea în orice format de prezentare (ppt/pptx, pdf, latex/beamer, etc).

Predarea finală se va face prin email, la adresa rudolf.erdei@cunbm.utcluj.ro, cu cel puțin o săptămână înainte de prezentare. Dacă se depășește acest termen, prezentarea se va face doar în restanță. Dacă nu se trimite emailul până la examenul fizic al materiei, restanța se va putea da doar anul următor (dacă se parcurg pașii cu noile deadline-uri).

3 Capitolele din documentație

Documentația va trebui să conțină următoarele capitole și să aibă minim 10 pagini:

1. Introducere, motivația alegerii bazei de date respective;
2. Contextul bazei de date și al proiectului, cerințe, ce dorim să obținem;
3. Aspecte teoretice relevante, inclusiv *state-of-the-art* (starea actuală a domeniului) cu minim 10 referințe științifice;
4. Implementarea aspectelor teoretice în cadrul proiectulu;
5. Testare și validare;
6. Rezultate;
7. Concluzii.

Se **punctează suplimentar** concluziile bine scrise, formatarea corectă a paginilor, imagini/grafice relevante, paginile extra (dacă au sens).

Se **depunctează** conținutul "*de umplutură*", cum ar fi descrierile inutile și prea lungi referitorare la algorimi. Textul generat de inteligența artificială se elimină (nu se punctează), iar dacă referatul nu mai îndeplinește cerințele minimale, se refuză.

Notă documentație: între 1 și 10.

4 Cerințele pentru prezentare

- Ținută obligatorie;
- Maxim 5 slide-uri (în orice aplicație de realizare prezentări);
  - 1 slide - introducere, tema aleasă, motivația;
  - 1 slide - prezentarea bazei de date, câmpuri;
  - 1 slide - prelucrarea datelor;
  - 1 slide - algoritmii testați, modelul final;
  - 1 slide - rezultate, concluzii, corelații, curiozități, cunoștințe noi.
- Timpul maxim alocat fiecărui student este de 5 minute. Nu se punctează informațiile prezentate în afara acestui timp.
- Notă prezentare: între 1 și 10.

5 Deadlines

Toate deadline-urile sunt *hard*, nu se punctează contribuțiile făcute după deadline.

1. Alegerea bazei de date, comunicarea la șeful grupei și transmiterea la profesorul coordonator al laboratorului - **săpt 2**;
2. Încărcare, analiză, curățare, eliminare anomalii, înțelegerea și descrierea bazei de date - **săpt 4**;
3. Calcularea, analizarea și explicarea caracteristicilor și indicatorilor diferiți (gini index, information quantity, corelații, etc) - **săpt 6**;
4. Alegerea a 3 algoritmi, realizarea de modele, optimizarea hiperparametrilor, compararea modelelor, analiza rezultatelor antrenării, alegerea modelului final - **săpt 8**;
5. Analiza modelului final în lumina datelor, explicarea rezultatelor pe diferite instanțe de date (model explainability), interpretarea rezultatelor și cunoștințe noi în domeniu - **săpt 10**;
6. Realizarea documentației finale și a prezentării - **săpt 12**;
7. Susținerea proiectului în fața colegilor (5 minute) - **săpt 14**.

6 Structura notei finale

Nota finală va avea următoarea structură:

- **0,1 * Notă activitate** - activitate la laborator/prezență;
- **0,6 * Notă documentație** - documentația și corectitudinea realizării proiectului;
- **0,3 * Notă prezentare**.

Se punctează suplimentar originalitatea/creativitatea dar și efortul suplimentar al fiecărui student. *Este necesară nota minimă 5 la fiecare aspect dintre cele 3 din structura notei.*

Mă aștept să lucrați cam 4-8 ore suplimentar acasă, pe lângă munca de la facultate (4 ore pe săptămână). Vă rog să vă aduceți device-urile personale, pentru a putea lucra mai ușor.

Pentru cine dorește să-și recupereze vreo prezență, fiecare prezență se recuperează cu un referat de 5 pagini despre orice aspect al învățării automate (de la Data Gathering, Data Analysis până la Model Deployment), cu condiția să fie muncă personală (nu copiat, descărcat, ChatGPT) și să demonstreze că înțelegeți conceptul. Referatul trebuie să fie de minim nota 7 ca să fie luat în considerare.