Big Data
Curs Master – anul 1

Modalitatea de notare – semestrul 2, 2024-2025

Nota la materia "Big Data" se obține pe baza unui Proiect realizat asupra unui set de date, asupra căruia vor fi aplicate noțiuni studiate la curs / laborator.

Proiectul se va realiza individual. În funcție de dimensiunea problemei tratate poate fi acceptată alcătuirea unor echipe (se va discuta separat posibilitatea unei astfel de opțiuni).

1. Formula de notare

Nota va fi acordată astfel (cerințele menționate sunt detaliate în secțiunea 6 a acestui document):

Nota = 1 pc oficiu + N_1 + N_2 + N_3, unde:

- N_1 = maximum 4 puncte, obținute din rezolvarea cerințelor 1-4, indiferent de complexitate, însă cu condiția ca proiectul să fie funcțional (codul să fie implementat și să ruleze).
- N_2 = maximum 2 puncte, obținute din rezolvarea cerințelor 5-7 indiferent de complexitate.
- N_3 = maximum 3 puncte, obținute din complexitatea proiectului (aceasta include tratare complexă a fiecărui punct, grad de dificultate, exemple originale și coerente, metode ce nu au fost prezentate sau lucrate la laborator etc.) și / sau întocmirea unui referat.

2. Date importante și parcurs

- Toate datele limită și link-urile pentru încărcarea referatelor și proiectelor se află în fișierul ***date_importante.txt***. Sunteți rugați să țineți seama de datele respective în stabilirea altor examene sau activități.

- Referatul este opțional, contând doar pentru obținerea de puncte pentru partea de complexitate. Opțiunea de realizare a unui referat se va anunța prin formular/assignment până pe data de **30 aprilie 2025, ora 23:59**. Încărcarea referatului se va face până pe **18 mai 2025, ora 23:59**, iar prezentarea referatului va avea loc în cadrul cursului, în ultimele 2 săptămâni ale semestrului.

- Proiectul va fi încărcat până la data de **14 iunie 2025, ora 23:59**.

- **Prezentarea proiectului** este obligatorie și va avea loc la data examenului planificat în **sesiune**. Încărcarea proiectului și neprezentarea la examen implică **restanță**.

Observație: La momentul începerii semestrului, datele specificate pentru examene și restanțe sunt estimative, putând fi modificate în funcție de unele constrângeri ce nu pot fi prevăzute din timp. Termenele limită sunt fixe și nu vor mai fi modificate.

3. Condiții pentru promovare

- Obținerea notei 5, conform formulei de notare, cu condiția $N_1 = 4$.
- **Respectarea condițiilor de etică și profesionalism ale facultății pentru întreaga activitate, inclusiv proiectul și referatul.** Acestea trebuie să fie **originale**, reprezentând rezultatul muncii studentului.

4. Restanță / reexaminare/ mărire de notă

- Se aplică aceeași modalitate de notare, cu alte termene pentru încărcarea proiectului. Aceste termene se află în fișierul ***date_importante.txt***.

5. Sistem

- Proiectul va utiliza limbajele/tehnologiile prezentate la curs / laborator:
  - Limbajul Python
  - Librării Python pentru calcul științific: numpy, pandas (opțional, pentru vizualizarea datelor: matplotlib, seaborn)
  - Spark (componentele SQL, Streaming și MLlib) – folosind interfața în Python pentru Apache Spark (pySpark)
  - TensorFlow

6. Cerințe

6.1 Cerințe proiect

- Proiectul va porni de la un set de date (ales de către student). În cadrul laboratoarelor vor fi enumerate câteva site-uri de unde pot fi descărcate seturi de date, însă alegerea nu este limitată la acestea.
- Structura proiectului este următoarea:

1. Introducere
   a. Prezentarea succintă a setului de date
   b. Enunțarea obiectivelor

   Observație: Se va indica obligatoriu link-ul către setul de date.

2. Implementarea de script-uri Spark pentru procesarea datelor, pregătirea, curățarea, transformarea acestora etc. Vor fi prezente grupări și agregări de date. Se vor utiliza Dataframes și Spark SQL (ambele).
    - Fiecare exemplu va fi precedat de descrierea sa în limbaj natural (aceasta poate fi parte a unei celule de text).
    - Codul sursă va fi inserat în document ca text (nu doar ca imagine).
3. Aplicarea a cel puțin două metode ML:
    - Utilizați Spark MLlib pentru clasificări, regresii etc. (în funcție de problemă)
    - Pentru fiecare metodă se va include în document :
        a. Enunțul problemei, justificarea alegerii metodei, explicarea soluției
        b. Aplicarea și evaluarea metodei
    - Codul sursă va fi inserat în document ca text (nu doar ca imagine).
4. Utilizarea a cel puțin unui Data Pipeline.
5. Utilizarea unei funcții definite de utilizator (UDF), optimizarea hiperparametrilor.
6. Aplicarea a cel puțin unei metode DL:
    - Utilizați Tensorflow
    - Pentru fiecare metodă se va include în document :
        a. Enunțul problemei, justificarea alegerii metodei, explicarea soluției
        b. Aplicarea și evaluarea metodei
    - Codul sursă va fi inserat în document ca text (nu doar ca imagine).
7. Creați un proces de streaming (din orice sursă) folosind Spark Streaming. Aplicați modelele ML antrenate anterior pentru inferența în timp real sau realizați învățarea online a unui model. De exemplu: procesarea în timp real a imaginilor care provin de la diferite camere video într-un sistem distribuit.
    - Pentru un punctaj mai mic, se poate aplica Spark Streaming în afara contextului ML.

6.2 Cerințe referat (opțional)

- Temele de referat vor putea fi alese dintr-o listă disponibilă sau vor putea fi propuse de către studenți. Temele propuse trebuie să fie relevante în raport cu subiectul cursului.
- Referatul va conține obligatoriu o parte de rezumat (abstract) și bibliografie ce va fi referită în textul referatului.
- Este recomandată existența unei părți practice a referatului (implementare / ilustrarea conținutului cu ajutorul unor exemple).

6.3 Complexitate

Pe lângă realizarea unui referat, punctajul pentru complexitatea proiectului va putea fi acordat în funcție de aspecte precum:

- gradul de dificultate al rezolvărilor;
- originalitatea exemplelor (chiar dacă folosesc aceleași tehnici sau metode, exemplele să fie diferite de cele de la laborator sau curs);
- utilizarea unor metode care nu au fost implementate la laborator sau curs, respectând cerințele impuse asupra tehnologiilor utilizate;
- prezentarea unor exemple folosind tehnologii amintite la curs / laborator, dar neimplementate la laborator;
- calitatea documentului în care vor fi integrate toate cerințele proiectului;
- calitatea fișierelor notebook (celule de text, comentarii etc.).

7. Condiții de redactare

Pentru a fi luat în considerare, proiectul trebuie să conțină **obligatoriu**:

- Un fișier *docx* sau *notebook* (recomandat) salvat ca *pdf* care să integreze toate rezolvările cerințelor și care să îndeplinească următoarele condiții:
  - va fi **structurat** conform cerințelor proiectului, va avea pagină de titlu și cuprins generat;
  - va include *print-screen*-uri prin care să se demonstreze că tot codul inclus în proiect a fost rulat; dacă optați pentru scrierea întregului proiect ca *notebook*, este suficient să păstrați *output*-urile rezultate în urma execuției;
  - va conține obligatoriu **tot codul și sub formă de text** (nu doar ca imagine!);
  - pentru fiecare cerință vor fi incluse **explicații / enunțul în limbaj natural.**
- Fișiere separate corespunzătoare setului de date și codului sursă corespunzător cerințelor 2-4, respectiv 2-7. Fișierele notebook vor conține **rezultatele execuției codului**! Dacă setul de date este prea mare pentru a fi trimis, în locul său se va insera în arhivă un fișier text ce va conține *link*-ul către setul de date respectiv

- Fișierele de mai sus vor fi denumite astfel: <grupa>_<Nume>_<Prenume>-<tip_document>.<extensie>, unde tip_document va avea valorile "Proiect_BigData", "Set_date", respectiv "Cod_sursa_cerinta_<n>" (de exemplu: 405_Popescu_Ana-Proiect_BigData.pdf, 405_Popescu_Ana_Cod_sursa_cerinta_1.ipynb). Fișierele astfel denumite vor fi încărcate până **la termenul limită stabilit**, *link*-ul pentru încărcare fiind anunțat în fișierul *date_importante.txt*.

8. Condiții de eligibilitate

Pentru a fi luat în considerare, proiectul trebuie să fie conform:

- standardelor pentru promovare și
- tuturor cerințelor marcate ca obligatorii în secțiunile anterioare. **Nerespectarea oricărei condiții obligatorii conduce la neeligibilitatea proiectului.**

Reiterăm condiția de respectare a Regulamentului de etică și profesionalism (https://drive.google.com/file/d/1gw2Fy44KnqaBKqLGh70vZPo3T9-ZdrDi/view). Ca urmare a acestuia, sunt interzise acțiuni precum: copierea de la colegi (inclusiv cei din anii anteriori) sau din alte surse, generarea de cod cu ajutorul utilitarelor sau platformelor bazate pe LLM etc. Neîndeplinirea acestui criteriu constituie incident ce trebuie raportat conducerii facultății și dezbătut în cadrul Consiliului. În acest caz, nota obținută va fi 1 (unu).

Indicați sursele pe care le folosiți prin referințe bibliografice. Acestea permit preluarea și prelucrarea unor idei din perspectivă proprie, dar nu copierea și însușirea lor!