Data partialului: 17 noiembrie

Ordinea de intrare in examen:

Toti sunteti bineveniti la ora 14 dar examinarea se face pe echipe si astfel trebuie organizata o ordine de intrare in examen.

Am tras la sorti si momentan ordinea de intrare in examen aleasa de numpy este:

[i+1 for i in np.random.permutation(9)]

Out[2]: [4, 5, 8, 7, 2, 9, 1, 3, 6]

|  |  |
| --- | --- |
| Echipa 4 | Ora 14:00 |
| Echipa 5 | Ora 14:15 |
| Echipa 8 | Ora 14:30 |
| Echipa 7 | Ora 14:45 |
| Echipa 2 | Ora 15:00 |
| Echipa 9 | Ora 15:15 |
| Echipa 1 | Ora 15:30 |
| Echipa 3 | Ora 15:45 |
| Echipa 6 | Ora 16:00 |
|  |  |

Va rog negociati intre voi si anuntati-ma de eventualele schimbari

Administrative:

Fisierele pentru partial vor fi **depuse cu o zi inainte de inceperea examinarii** - va rog sa trimiteti un email la adresa teodor.bildea@unitbv in care sa mentionati:

* Numarul si Componenta echipei
* Numele proiectului
* Fisierele atasate ( fie jupyter notebooks fie slide-uri + sursa graficelor si a numerelor prezentate)

Prezentarea va dura 10-15 minute pentru fiecare echipa asa ca ma astept ca in cazul slide-urilor sa fie 10 slide-uri max de continut pt prezentare; daca aveti mai multe le puteti organiza in cele mai importante 10 pt prezentarea orala, restul la sfarsitul prezentarii, intr-un appendix.

In cazul notebookurilor jupyter veti prezenta cat incape in cele 10 minute.

Teoretic sunt 10 echipe deci ne vom incadra in 3 ore.

Va incurajez sa **repetati prezentarea** pentru a va incadra in timp si a evita situatia in care nu ajungeti sa prezentati ceva interesant din lipsa de timp

Cerinte partial

1. Prezentarea cantitativa a datelor -20% din nota:
   1. Cate tabele, cate inregistrari
   2. In cazul tabelelor multiple, cheile de legatura
   3. Analiza datelor lipsa:
      1. Pe coloane, numarul de inregistrari lipsa (proportia de date lipsa la nivel de coloana)
      2. Pe linii, numarul de inregistari cu cel putin o valoare lipsa (proporita de inregistrari incomplete)
      3. Decizii luate - inregsitrari inlaturare/pastrate, atribute inlaturate/pastrate

a+b: 10%

C: 10%

1. Prezentarea calitativa a datelor - 20% din nota
   1. Prezentarea atributelor fiecare in parte:
      1. Tipurile de atribute (4%)
      2. Distributia valorilor - reprezentari grafice (4%)
      3. Statistici simple (4%)
      4. Analiza valori extreme - Daca exista la nivel de atribut: Outliers - valori extreme care ar putea fi eliminate pentru o mai buna intelegere a problemei si a modelarii (4%)
      5. Analiza valori aberante/erori umane - de exemplu preturi negative, facturi negative sau date neconforme (e.g. programare la doctor in evul mediu) (4%)
2. Matricea de corelatie - 20% din nota
   1. Pentru variabile numerice matricea de corelatie Pearson (5%)
   2. Pentru variabile categoriale, dupa codificarea valorilor categoriale, matricea de corelatie Spearman (rank correlation) dupa ce in prealabil toate valorile categoriale au fost convertite in numerice (e.g. LabelEncoder in Python sau manual in excel/Python) (5%)
   3. Reprezentarea grafica a matricei de corelatie (5%)
   4. Analiza componentelor principale pe cea mai mare dintre cele doua matrici
      1. Gasirea componentelor principale (3%)
      2. Determinarea acelor componente principale care impreuna explica 90% din variatia totala din date (2%)
3. Tendinte, trenduri, relatii intre variabile - 40% din nota

Sunt 2 tipuri de proiecte: clasificare (variabila de prognozat este discreta/categoriala), si regresie (variabila pe care o veti prognoza este numerica/continua)

* 1. Pentru regresie si clasificare este/sunt important(e)
     1. grafice cu distributii ale variabilei dependente pe clasele unor atribute discrete - pentru a determina:
        1. Distributia atributelor pe clasele care trebuie prognozate
           1. Care atribute separa datele cel mai bine in clase?
        2. Legaturi intre mai multe atribute si clase (e.g. grafic cu distributia in clase de client fidel/client dezertor grupate pe sex si vechime
     2. daca un atribut discret poate discerne valorile continue
     3. Grafice de tip scatter plot (imprastiere) pentru a observa eventuale corelatii/sabloane (poate pornind de la numerele din matricea de corelatie)

Examenul consta in prezentarea a ceea ce ati inteles din datele primite de la ‘client’. Este ocazia de a lamuri aspecte neclare in date, si de a cadea de acord asupra setului de date cu care se incepe modelarea. Puteti considera ca eu sunt clientul si ma astept sa imi prezentati datele in asa fel incat sa imi dau seama ca am incredintat proiectul pe maini bune.

Toate cele de mai sus se pot face pornind de la cod existent in notebookurile Jupyter vehiculate la orele de curs, dar si folosind alte informatii de pe site-ul kaggle.com, sursa datelor.

Daca va este mai la indemana sa utilizati altceva nu am nimic impotriva atata timp cat puteti sa imi dati codul sau aplicatia care sa poata reproduce graficele si alte elemente de calcul din prezentare.

Cei care vor face prezentarea intr-un Jupyter notebook vor primi puncte bonus.