# Șablon de enunț pentru problema la Olimpiada de Inteligență Artificială

1) Enunt - text introductiv

2) Un set de date de antrenare si unul de evaluare

* preferabil în format cat mai simplu de procesat si vizualizat, d.ex., CSV, imagini.

3) Preferabil doua cerințe, P1 si P2, ambele putând avea mai multe sub-cerințe.

* P1 - mai simpla cu o pondere de 20-40 din punctaj.
* P2 – pentru departajare

4) Candidaților li se va cere sa producă output-urile in diferite formate si fișiere care vor fi comparate automat de evaluator. Spre deosebire de OJI/ONI, codul elevilor nu va fi rulat (decât opțional, la latitudinea comisiei).

# Exemplu de problema

Implementati un model de AI pentru a estima preturile din jocurile de pe Steam avand la dispozitie un set de antrenare – dataset\_train.csv - si unul pe care trebuie sa faceti doar predictia - dataset\_eval.csv

Dataset-ul contine urmatoarele campuri avand semnificatiile:

* Name: nume aplicație
* AppID: identificatorul jocului
* Metacritic score: Nota medie a jocului de pe platforma de recenzii Metacritic
* Genres : O lista separata prin virgule conținând tag-uri referitoare la tipul de joc
* Publishers: Numele producatorilor
* Estimated Owners: de forma min – max, insemnand asteptarile (producatorilor jocului) minime si maxime de jocuri vandute
* Negatives: numarul de dislike-uri
* Positives: numarul de like-uri
* Recommendations: cate persoane recomanda acest joc

Note despre setul de date:

* Câmpul-ținta este „Price”. Date fiind celelalte feature-uri, scopul este de a prezice „Price”. Setul de date de evaluare vizibil pentru candidați nu conține acest câmp, însă el este necesar platformei de evaluare. Metrica de evaluare folosita este MAE (media sumei diferentelor in modul dintre output-ul predictiei voastre si cea corecta).
* Câmpul Genres, fiind o lista, trebuie transformat in date numerice într-un fel sau altul pentru a putea fi folosit in algoritmii de ML.
* Asemănător pentru câmpul Estimated Owners
* Unele câmpuri pot fi inutile in predicția campului-tinta. Incercati sa analizati datele si sa selectați doar ce este nevoie sau ar putea explica predictia.

**P1 (40p**). Incarcati setul de date de antrenare si scrieți un fișier output\_1.csv in care sa aveți header-ul:

"Samples,Average Price,Average Owners,Unique Genres”

* Samples: numarul de linii de input din setul de date de antrenare, folosit in scop de verificare locala
* Average Price: pretul mediul pentru toate produsele din setul de date
* Average Owners: pentru campurile din coloana Estimated Owners de forma min – max, extrageti media tutor valorilor considerand pentru fiecare linie int ((min + max) / 2).
* Unique Genres: numarul unic de tag-uri existet in dataset pentru toata coloana Genres

Fiecare din cele 4 componente valoreaza 10p.

**P2 (60p)** Implementati un model de AI si rulați predicția pentru fiecare rând din dataset\_eval.csv.

Output-ul trebuie sa fie un fișier output\_2.csv cu o singura coloana numita „Price”.

# Recomandări pentru propunătorii de probleme

1. Input si Output (Ex: folder-ul Output\_CandidatX)

Candidații primesc un set de date pentru a face antrenarea – dataset\_train.csv si unul pentru evaluare – dataset\_eval.csv. La cel de-al doilea, campul tinta, “Price”, este ascuns.

Candidații au doua fisiere de salvat in folderul lor:

* Output\_1.csv in care sunt cele 4 valori punctate
* Output\_2.csv, fiecare linie contine predictia campului target corespunzator liniei din dataset\_eval.csv

Candidații trebuie sa știe dinainte metrica de evaluare.

1. Procesarea si crearea unui dataset

Porniti de la un dataset existent, nu foarte cunoscut de preferat dar incercati sa evidentiati capcanele pe cat posibil sau dificultatile ulterioare in incarcare si procesare.

Spre exemplu mi-am dat seama ca ar trebui sa las coloanele din setul original care sunt clare sau care sa le pot procesa astfel incat sa fie usor de vizualizat.

In acest sens am procesat dataset-ul original cu Dataset.ipynb, posibil sa gasiti acest cod util.

1. Implementati solutia si iterati asupra ei astfel incat sa fie realizabila cu cat mai puține librarii aditionale sau API-uri.

Spre exemplu, intentionat in solutia Output\_CandidatX am lasat procesarea datelor manuala, cu cat mai putin interactiuni.