

# Predicția statusului pacientului

Pentru această problemă trebuie să implementați un model de învățare automată pentru a prezice valoarea câmpului **Status** utilizând un set de date disponibil, care conține informații despre pacienți. Setul de date este organizat într-un fișier CSV care include diverse trăsături (*features*), iar evaluarea modelului se va face pe baza preciziei pentru clasa **Dead**.

## Descrierea setului de date

Setul de date conține următoarele câmpuri, fiecare cu semnificația respectivă:

- **Age:** Vârsta pacientului.
- **Race:** Rasa pacientului (ex: White, Other).
- **Marital Status:** Starea civilă a pacientului (Married, Single, Divorced etc.).
- **T Stage:** Stadiul tumorii, conform sistemului TNM (T1, T2, T3, T4).
- **N Stage:** Gradul de afectare a ganglionilor limfatici (N0, N1, etc.).
- **6th Stage:** Stadiul cancerului conform clasificării TNM a celei de-a 6-a ediții.
- **differentiate:** Gradul de diferențiere tumorală (Well, Moderately, Poorly differentiated).
- **Grade:** Gradul histologic al tumorii (1, 2, 3 etc.).
- **A Stage:** Clasificarea stadiului bolii (ex: Regional).
- **Tumor Size:** Dimensiunea tumorii (poate conține valori lipsă).
- **BMI:** Indicele de masă corporală.
- **Heart Rate:** Ritmul cardiac.
- **Serum Creatinine:** Nivelul seric de creatinină.
- **Uric Acid:** Nivelul acidului uric din sânge.
- **Hemoglobin:** Concentrația hemoglobinei.
- **GFR:** Rata de filtrare glomerulară (funcția renală).
- **Serum Sodium:** Concentrația de sodiu seric.
- **Serum Potassium:** Concentrația de potasiu seric.
- **Serum Albumin:** Nivelul albuminei serice.
- **Lactate:** Concentrația lactatului.
- **Status:** Starea pacientului (**Dead** / **Alive**), câmpul țintă.

## Task-uri

Pentru primele subtask-uri, va trebui să încărcați setul de date și să efectuați analize statistice pentru a înțelege mai bine datele.

### Subtask 1 (10p)

Pentru fiecare pacient din setul de testare, clasificați funcția renală în una dintre categoriile:

- **Normal**, dacă  $GFR \geq 90$ .
- **Mildly Decreased**, dacă  $60 \leq GFR < 90$ .

### Subtask 2 (10p)

Folosiți setul de antrenare pentru a calcula cuartilele valorilor din **Serum Creatinine**:

- Q1: 25% dintre valori sunt sub acest prag.
- Q2: Mediana (50%).
- Q3: 75% dintre valori sunt sub acest prag.

Apoi, pentru fiecare pacient din test, atribuiți nivelul de risc:

- Very Low:  $\leq Q1$
- Low:  $Q1 < x \leq Q2$
- High:  $Q2 < x \leq Q3$
- Very High:  $> Q3$

### Subtask 3 (10p)

Din setul de antrenare, determinați media și mediana **BMI**. Pentru fiecare pacient din test:

- scrieți **1**, dacă  $BMI > mediana_{train}$ ;
- altfel, scrieți **0**.

### Subtask 4 (10p)

Pentru fiecare pacient din test, determinați câți pacienți din setul de antrenare se află în același **T Stage**.

### Subtask 5 (60p)

Construiți un model de învățare automată care să prezică **Status** pe baza trăsăturilor disponibile. Evaluarea se face pe baza **preciziei pentru clasa Dead**.

## Note despre setul de date

- Câmpul-țintă este **Status**.
- Câmpuri numerice utile: **Tumor Size**, **BMI**, **Serum Potassium**.
- **Tumor Size** poate avea valori lipsă  $\rightarrow$  trebuie tratate.
- Unele variabile pot fi nerelevante  $\rightarrow$  se recomandă analiză de corelație și eliminare.
- Posibile valori extreme  $\rightarrow$  recomandat tratament (IQR, Z-score, boxplot, histogramă).

## Criterii de evaluare

Metrica: **precizia pentru clasa Dead**. Se evaluează cât de bine este prezis decesul pacienților.

## Notă

Dacă trimiteți și fișierul `sample_output`, primiți 5 puncte suplimentare.