Laborator 2 - Analiza și Preprocesarea Datelor

1. Identificarea setului de date
   * Setul de date utilizat: **Forest Fires Dataset**
   * Sursă: UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/forest+fires>)
   * Descriere: Acest set de date conține informații despre condițiile meteorologice, umiditate, vânt, precipitații și zona afectată de incendii în pădurile din Parcul Natural Montesinho, Portugalia.
   * Tipul datelor: Variabile numerice (temperatură, umiditate, vânt) și categorice (lună, zi).
2. Analiza exploratorie a datelor (EDA)

a) **Descrierea setului de date**

* + Rezultatele funcției describe():
    - Număr de instanțe: 517
    - Atribute: 13 (11 numerice, 2 categorice)
    - Temperatură medie: 18.8°C
    - Umiditate medie: 44.3%
    - Viteza medie a vântului: 4.0 km/h
    - Suprafața medie afectată de incendii: 12.8 ha

b) **Gestionarea datelor lipsă**

* + Nu au fost detectate valori lipsă.

c) **Gestionarea valorilor extreme**

* + Identificare outlieri utilizând **BoxPlot** și **Z-score**:
    - Suprafața afectată de incendii prezintă valori extreme.
    - Valorile extreme au fost păstrate pentru analiza ulterioară.

d) **Vizualizări și relații între variabile**

* + **Histograma temperaturii:** Distribuție normală.
  + **Heatmap cu matricea de corelație:**
    - Corelație negativă între umiditate și zona afectată (-0.39).
    - Corelație pozitivă între temperatură și zona afectată (0.58).

1. Preprocesarea setului de date

a) **Standardizare și normalizare**

* + Aplicare StandardScaler pentru variabilele continue.
  + Aplicare MinMaxScaler pentru variabilele meteorologice.

b) **Codificare variabile categoriale**

* + Transformare one-hot encoding pentru lună și zi.

c) **Selecția caracteristicilor**

* + Folosirea **corelației** pentru eliminarea caracteristicilor redundante.
  + Aplicarea **PCA** pentru reducerea dimensiunii.

1. Link spre cod sursă
   * Link către notebook-ul utilizat: [GitHub/Google Drive/Arhivă locală]

Concluzie:

* Datele au fost curățate, analizate și preprocesate pentru a fi utilizate într-un model de predicție a incendiilor de pădure.
* Următorul pas este antrenarea și evaluarea modelelor de învățare automată.