Explorarea Datelor (Exploratory Data Analysis)

1. Analiza tipului de atribute și a plajei de valori a acestora

Pentru datele din PIRvision:

	count	mean	std	min	25%	50%	75%	max	missing_values
Day Index	8000.0	2.001125	5.297724e-01	1.0	2.0	2.0	2.0	3.0	0
Temp (F)	8000.0	80.394750	2.286321e+01	0.0	86.0	86.0	88.0	89.0	0
Temp (C)	8000.0	26.701250	1.242631e+01	-17.0	30.0	30.0	31.0	31.0	0
OBS_1	7202.0	298767.516662	4.631447e+06	2613.0	10335.0	10433.0	10563.0	111602625.0	798
OBS_2	8000.0	10959.839000	1.366988e+03	2092.0	10775.0	11000.0	11281.0	16928.0	0
...	...	...	...	...	...	...	...	...	...
OBS_54	8000.0	10585.347125	4.288494e+02	2603.0	10513.0	10596.0	10684.0	16383.0	0
OBS_55	8000.0	10449.253750	4.179482e+02	2602.0	10388.0	10453.0	10523.0	16383.0	0
OBS_56	8000.0	11013.778000	4.554899e+02	2547.0	10928.0	11026.0	11127.0	17146.0	0
OBS_57	8000.0	10027.365500	4.015679e+02	2624.0	9957.0	10033.0	10108.0	15713.0	0
Class	8000.0	0.332750	8.184505e-01	0.0	0.0	0.0	0.0	3.0	0

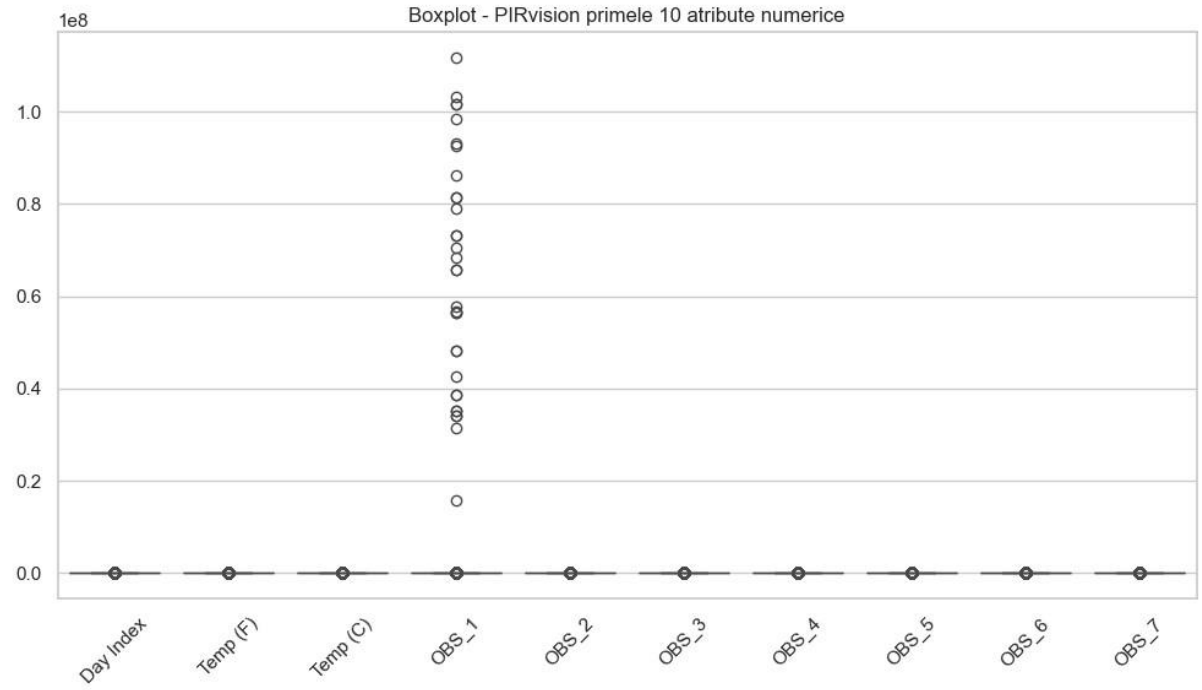
61 rows × 9 columns

Observ ca atributul OBS_1 are valoarea maxima mult mai mare fata de restul atributelor, precum si deviatia standard, adica sunt valori extreme. De asemenea, observ ca sunt multe valori lipsa(aproximativ 10%). Restul valorilor par normale.

Pentru Temp(F) si Temp(C): mediile par normale, insa temperaturile minime sunt cam scazute si poate semnifica o posibila eroare.
Atributul Class este eticheta de clasa, cu valori intre 0 si 3.

Atributul “Day Index” are doar 3 valori posibile, fiind atribut ordinal.

Grafic boxplot realizat pentru primele 10 atribute numerice:



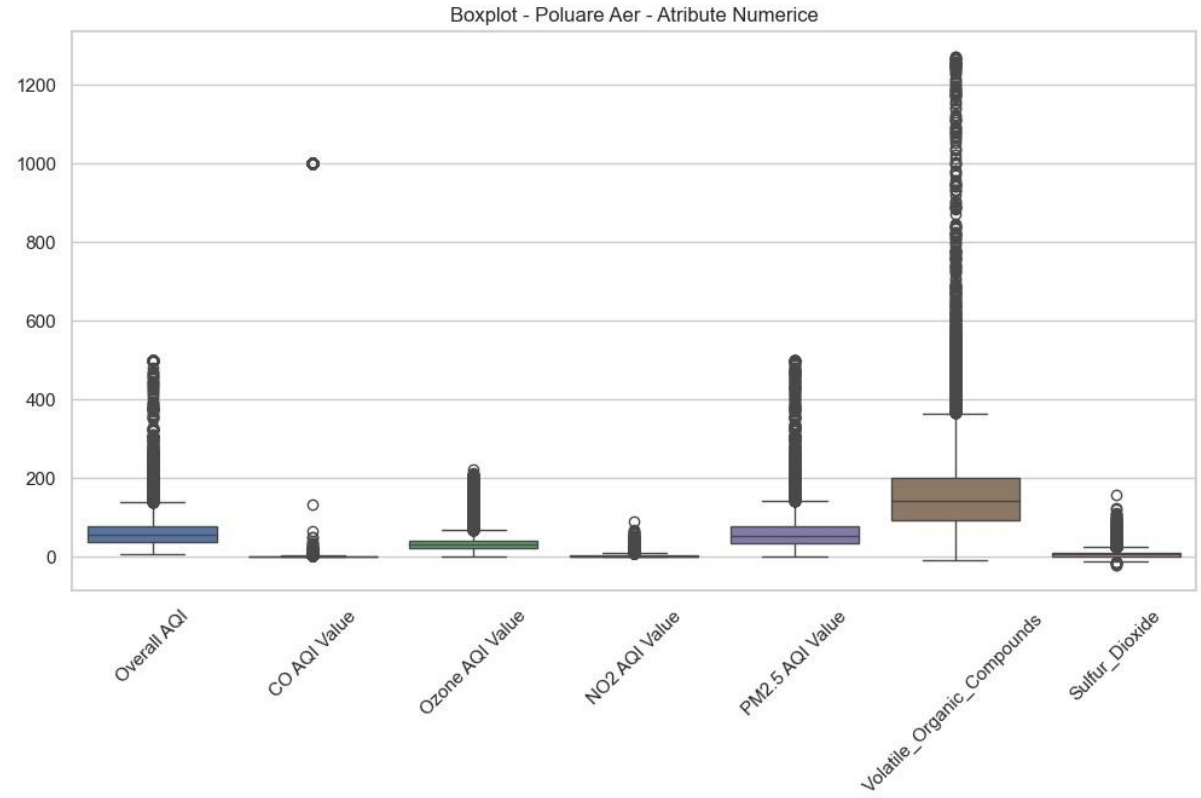
Observ ca doar OBS_1 are valori extreme foarte mari fata de restul atributelor care au valori rezonabile.

Pentru datele din Air pollution:

	count	mean	std	min	25%	50%	75%	max	missing_values
Overall AQI	18770.0	71.981726	56.110722	7.000000	39.000000	55.000000	79.000000	500.000000	0
CO AQI Value	18770.0	41.444700	196.121182	0.000000	1.000000	1.000000	2.000000	1001.368367	0
Ozone AQI Value	16900.0	35.372781	28.422401	0.000000	21.000000	31.000000	40.000000	222.000000	1870
NO2 AQI Value	18770.0	3.068727	5.300815	0.000000	0.000000	1.000000	4.000000	91.000000	0
PM2.5 AQI Value	18770.0	68.490996	54.717105	0.000000	35.000000	54.000000	78.000000	500.000000	0
Volatile_Organic_Compounds	18770.0	176.279602	137.198201	-8.563753	92.398163	141.208711	201.670858	1272.469897	0
Sulfur_Dioxide	18770.0	7.503989	10.745942	-20.328019	1.214184	5.572767	10.804549	159.231403	0

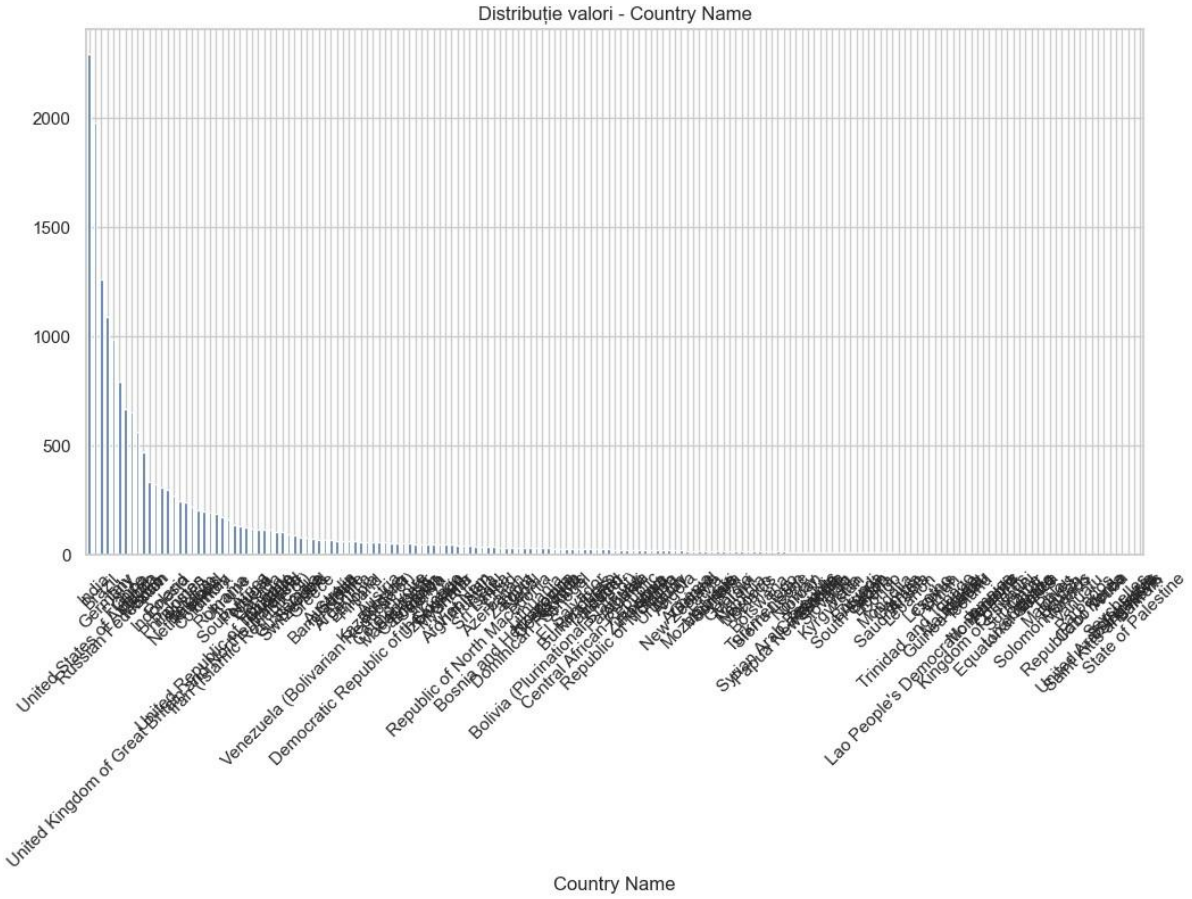
-Atributul “CO AQI Value” are o valoare maxima 1001.37 extrem de ridicata fata de restul valorilor, inclusive fata de 75%(2), fiind valoare extrema. La fel si std este foarte mare fata de medie.
-“Averall AQI” pare sa aiba valori normale.
-“NO2 AQI Value” poate avea outlieri pentru ca are distributie asimetrica, avand majoritatea valorilor mici, dar nu reprezinta o problema mare.
-“PM2.5 AQI Value” este atribut bun.
-“Volatile Oragenic Compounds(VOC)” are minimul mai mic decat 0, cee ace este o eroare de sensor, are valoarea maxima 1272 ceea ce este mult peste 75%(201), adica este outlier. -“Sulfur Dioxide” are valoarea minima negative, ceea ce este imposibil, adica este eroare de masurare.

Graficul Boxplot:

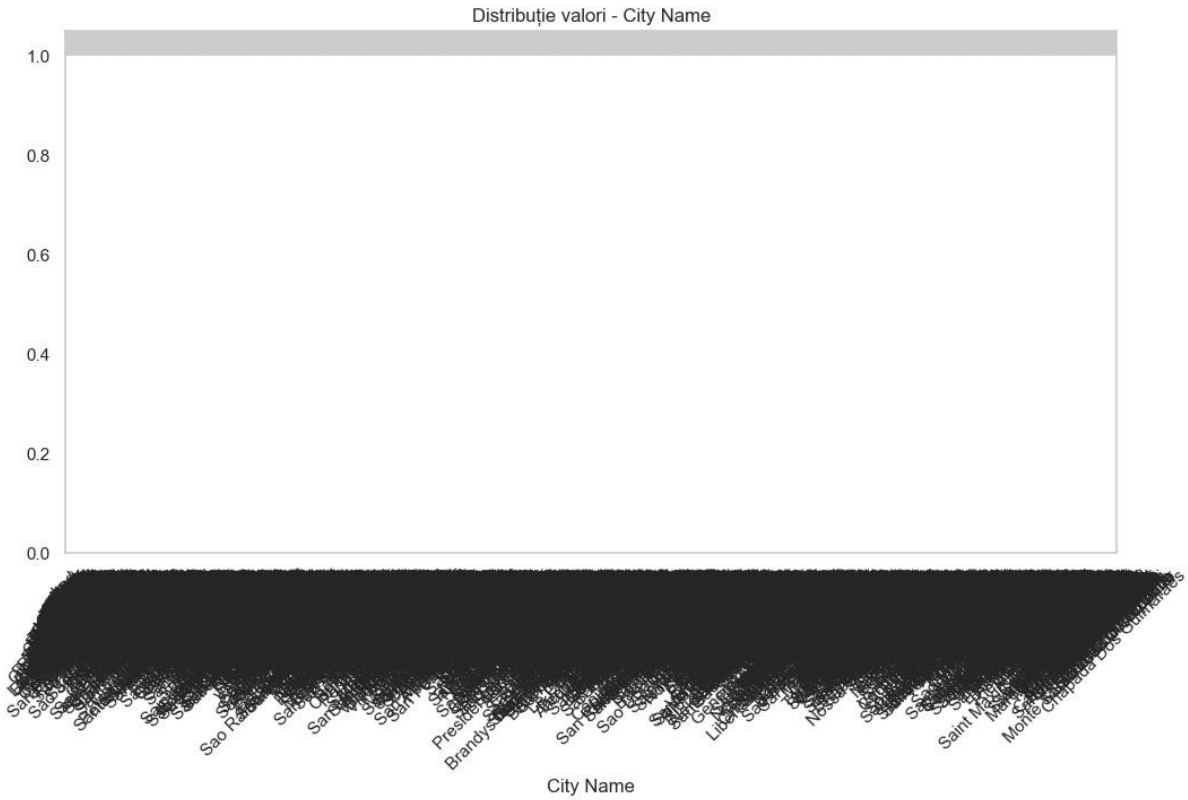


Observ ca “CO AQI Value” are valori foarte mari(> 1000), “PM2.5 AQI Value” are multe valori peste 100 si ajung pana la 400-500, “Volatile Organic Compounds” are extrem de multee valori peste 1000 si “Sulfur Dioxide” are valori negative imposibile. Restul valorilor sunt bune

Pentru atribute discrete sau ordinale:

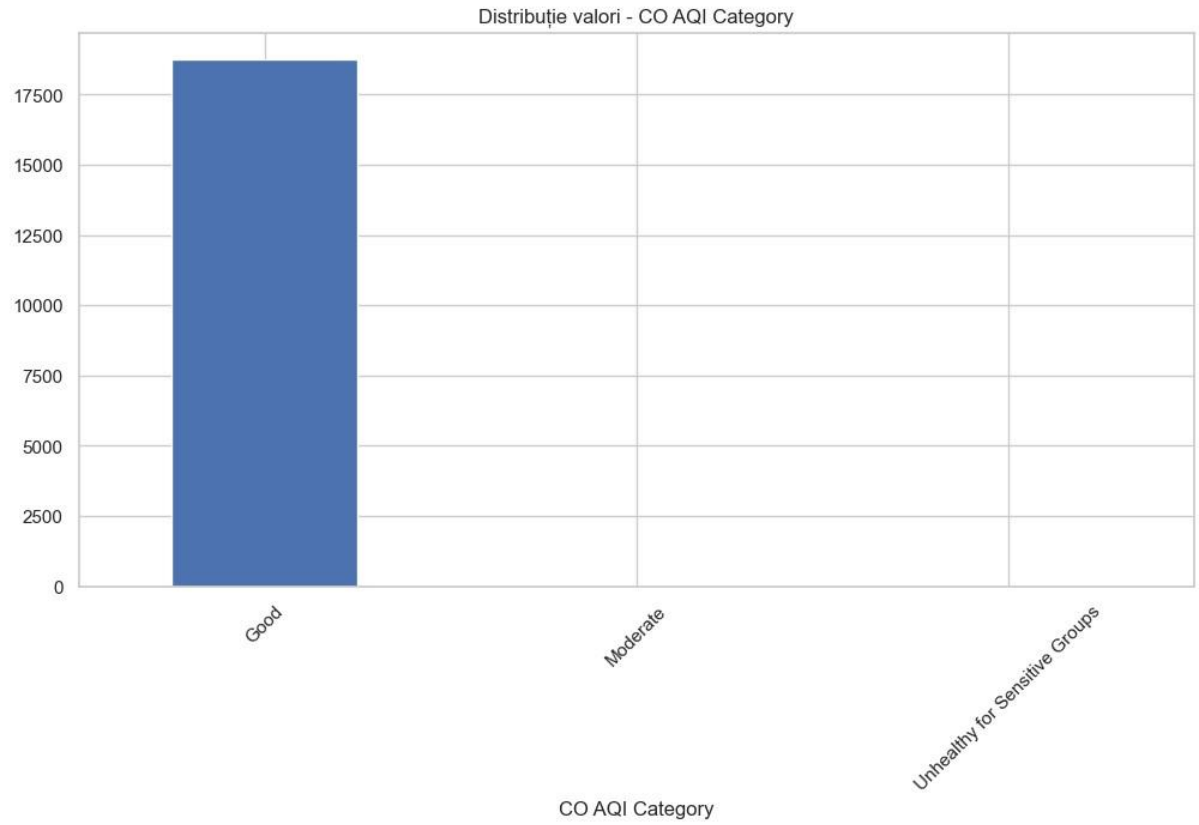


Country Name -> missing: 349, unique: 175
Coloana “Country Name” contine 175 de valori unice, cu o distributie dezechilibrata. Cateva tari dominante au un numar foarte mare de inregistrari(peste 2000), in timp ce majoritatea apar rar.



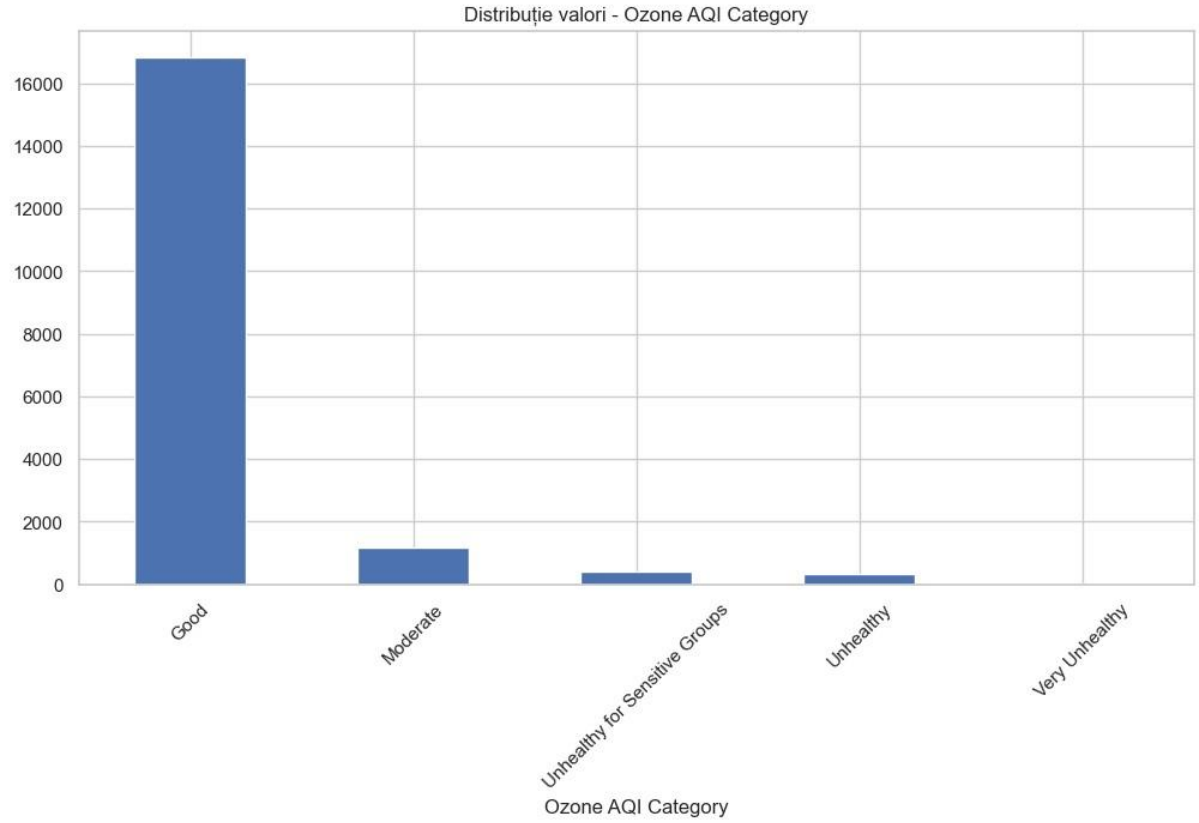
City Name -> missing: 0, unique: 18770

Coloana “Country Name” contine un numar extrem de mare de valori unice, cee ace il face foarte greu de reprezentat visual printr-un grafic de tip barplot.



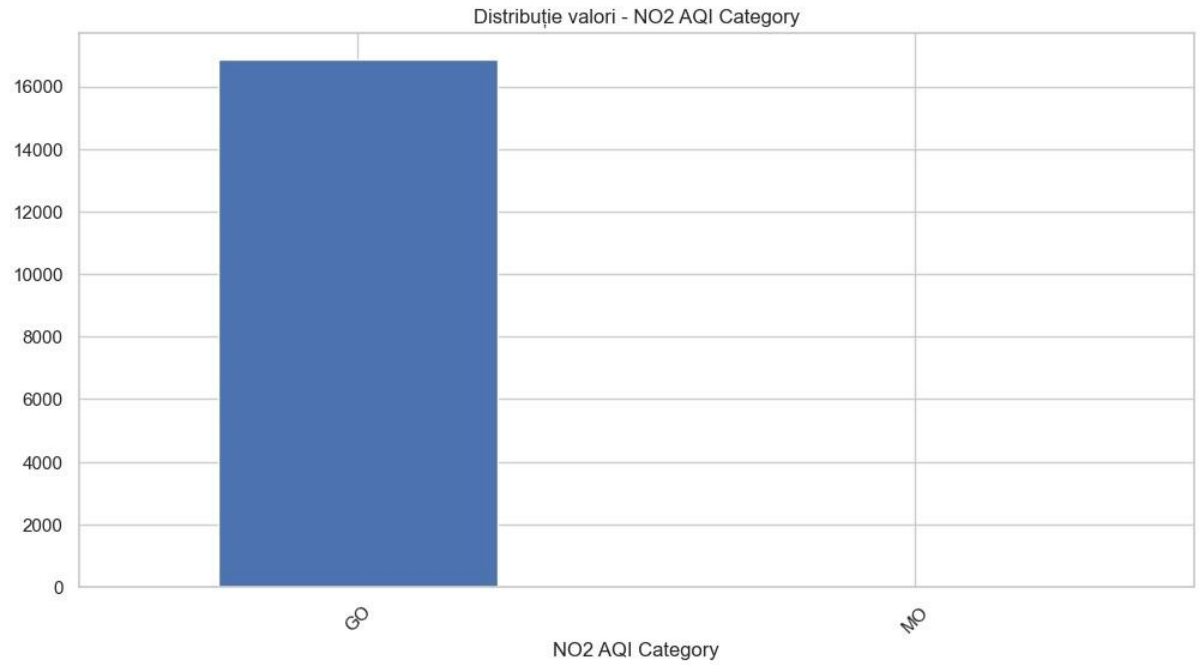
CO AQI Category -> missing: 0, unique: 3

Atributul “CO AQI Category” este puternic dezechilibrat, avand majoritatea valorilor in categoria “Good”, insemnand ca in majoritatea locatiilor monitorizate nivelul de CO este sigur.



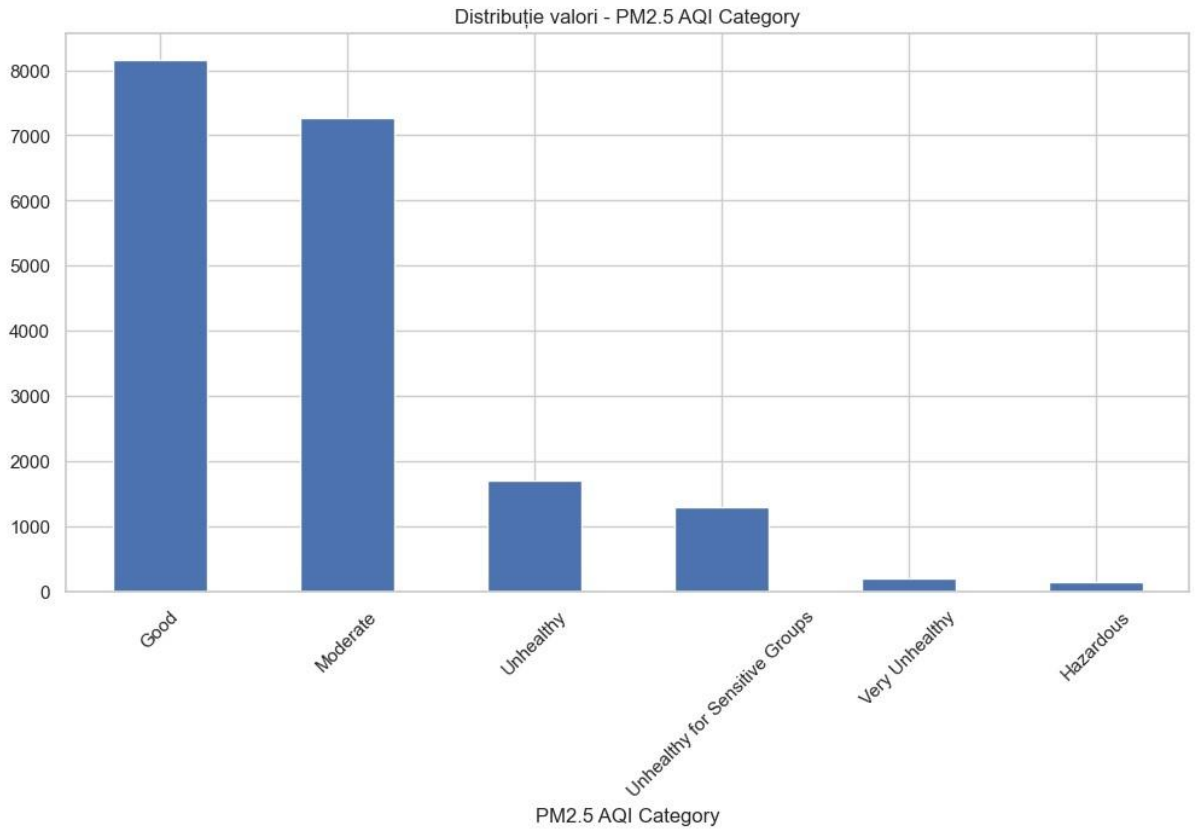
Ozone AQI Category -> missing: 0, unique: 5

Atributul “Ozone AQI Category” este distribuit dezechilibrat, insa mai variat decat cel precedent. “Good” este cea mai frecventa clasa, dar distributia include si valori semnificative din categoriile “Moderate”, “Unhealthy for Sensitive Groups” si “Unhealthy”.



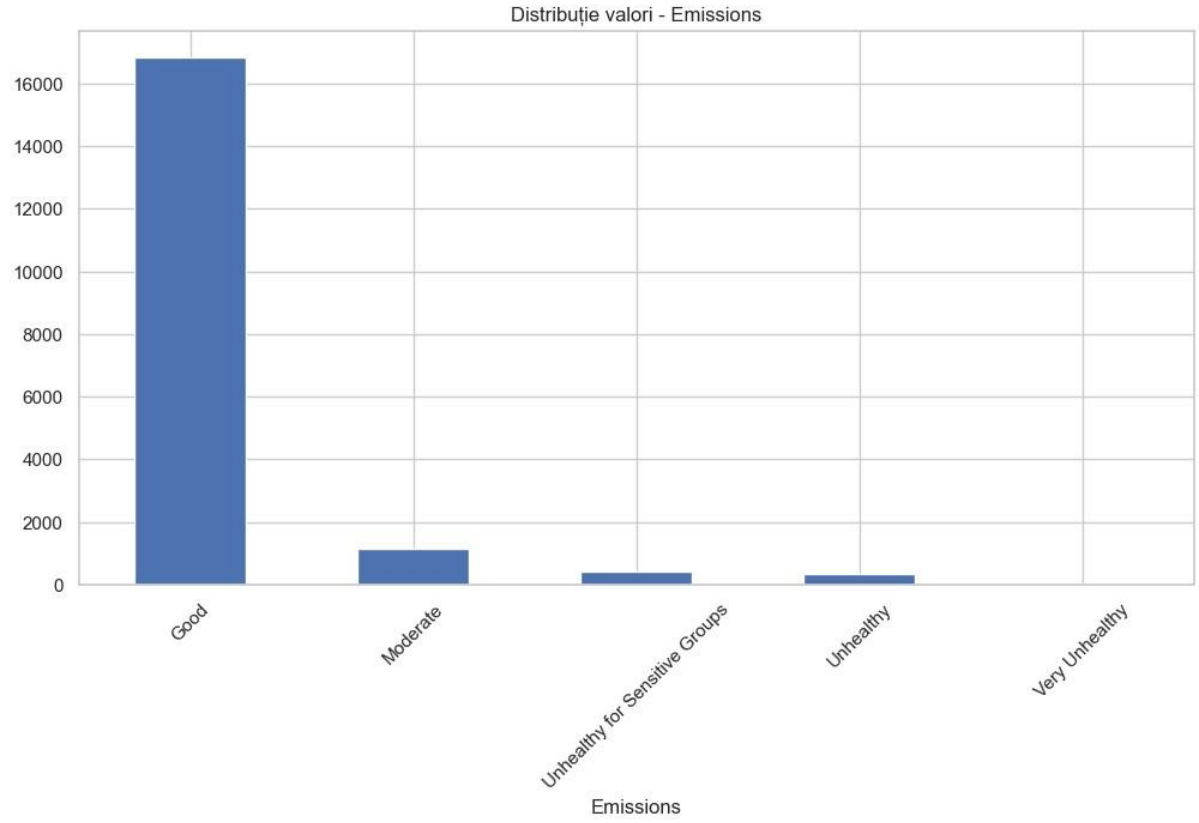
NO2 AQI Category -> missing: 1893, unique: 2

Atributul “NO2 AQI Category” are si el o distributie dezechilibrata, cu majoritatea observatiilor incadrate in clasa “Good”.



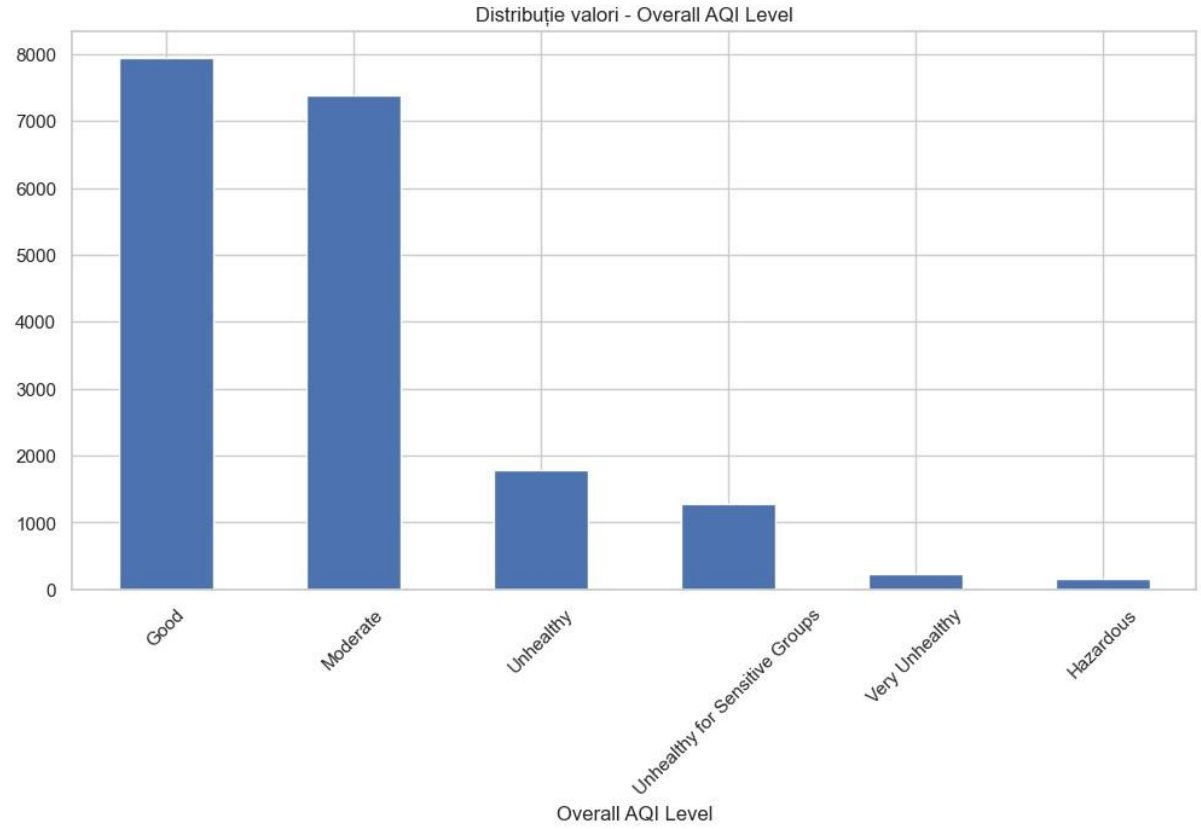
PM2.5 AQI Category -> missing: 0, unique: 6

Atributul “PM2.5 AQI Category” are o mai buna acoperire decat alte categorii. Clasele “Good” si “Moderate” domina distributia, totusi aparitia unor clase semnificative de risc(“Unhealthy”, “Unhealthy for Sensitive Groups”, “Very Unhealthy”, “Hazardous”) ofera o oportunitate buna pentru algoritmii de clasificare sa invete diferentierea intre conditii normale si periculoase.



Emissions -> missing: 0, unique: 5

Atributul “Emissions” are o distributie puternic dezechilibrata, fiind dominate de categoria “Good”. Categoriile mai severe sunt rare.



Overall AQI Level -> missing: 0, unique: 6

Distributia “Overall AQI Level” are predominante starile de aer “Good” si “Moderate”, insa sunt prezente si date despre restul categoriilor.

Tabel cu valori lipsa si unice:

	Atribut	Valori lipsă	Valori unice
0	Country Name	349	175
1	City Name	0	18770
2	CO AQI Category	0	3
3	Ozone AQI Category	0	5
4	NO2 AQI Category	1893	2
5	PM2.5 AQI Category	0	6
6	Emissions	0	5
7	Overall AQI Level	0	6

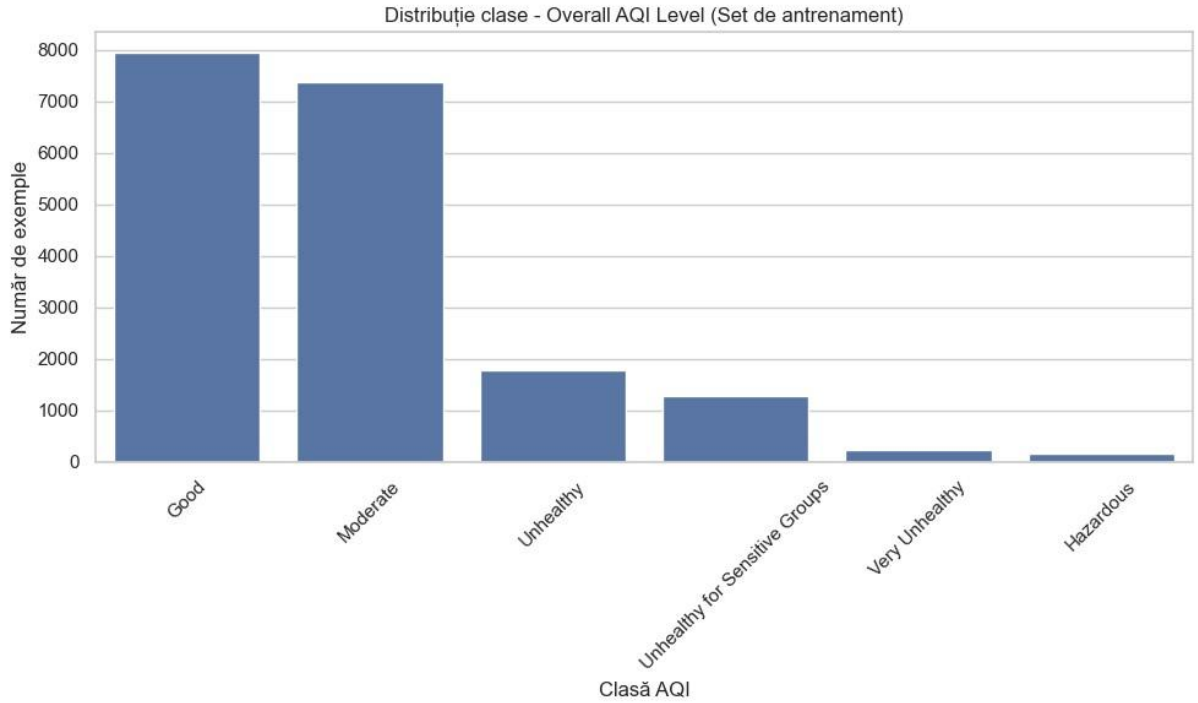
Atributele “Country Name” si “City name” contin multe valori unice.

“NO2 AQI Category” are multe valori lipsa.

“Overall AQI Level” este bine definita, dar dezechilibrata.

Restul atributelor sunt bine structurate.

2.Analiza echilibrului de clase Air Pollution - train:



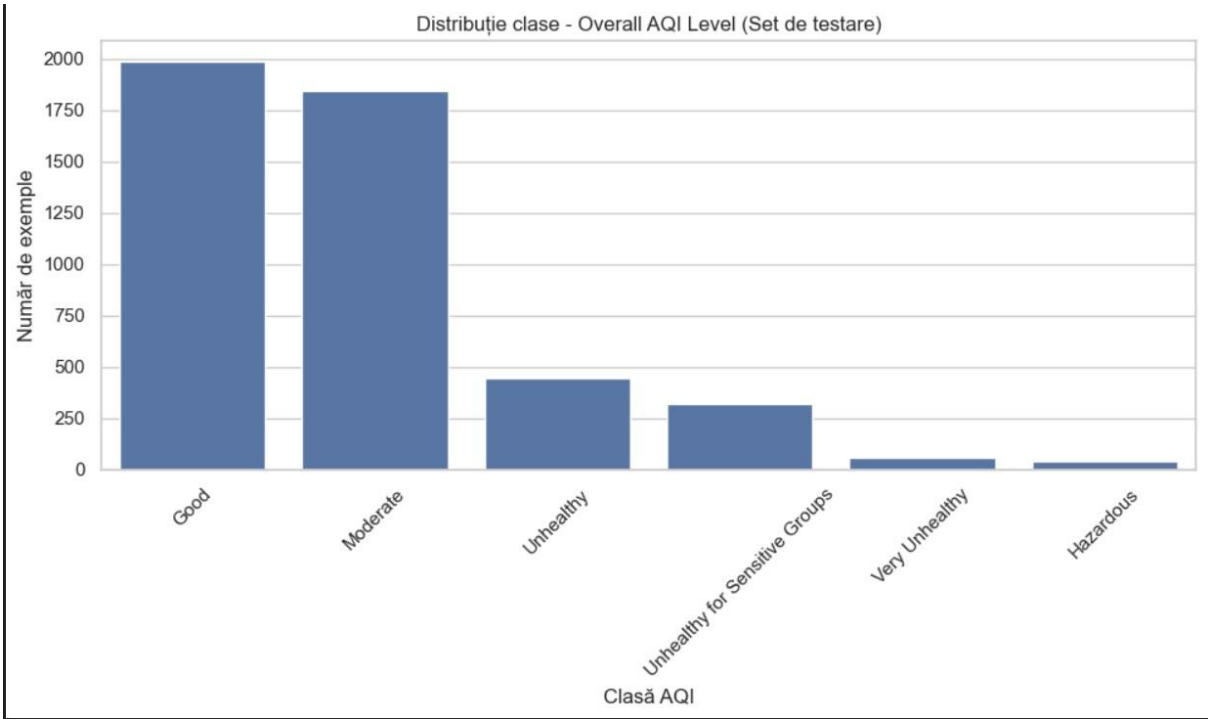
-Clasele “Good” si “Moderate” sunt cele mai frecvente, cu cate aproximativ 8000 exemple fiecare

-“Unhealthy”, “Unhealthy for Sensitive Groups” au semnificativ mai putine exemple(1000-2000)

-“Very Unhealthy” si “Hazardous” au un numar foarte mic de exemple(sub 500).

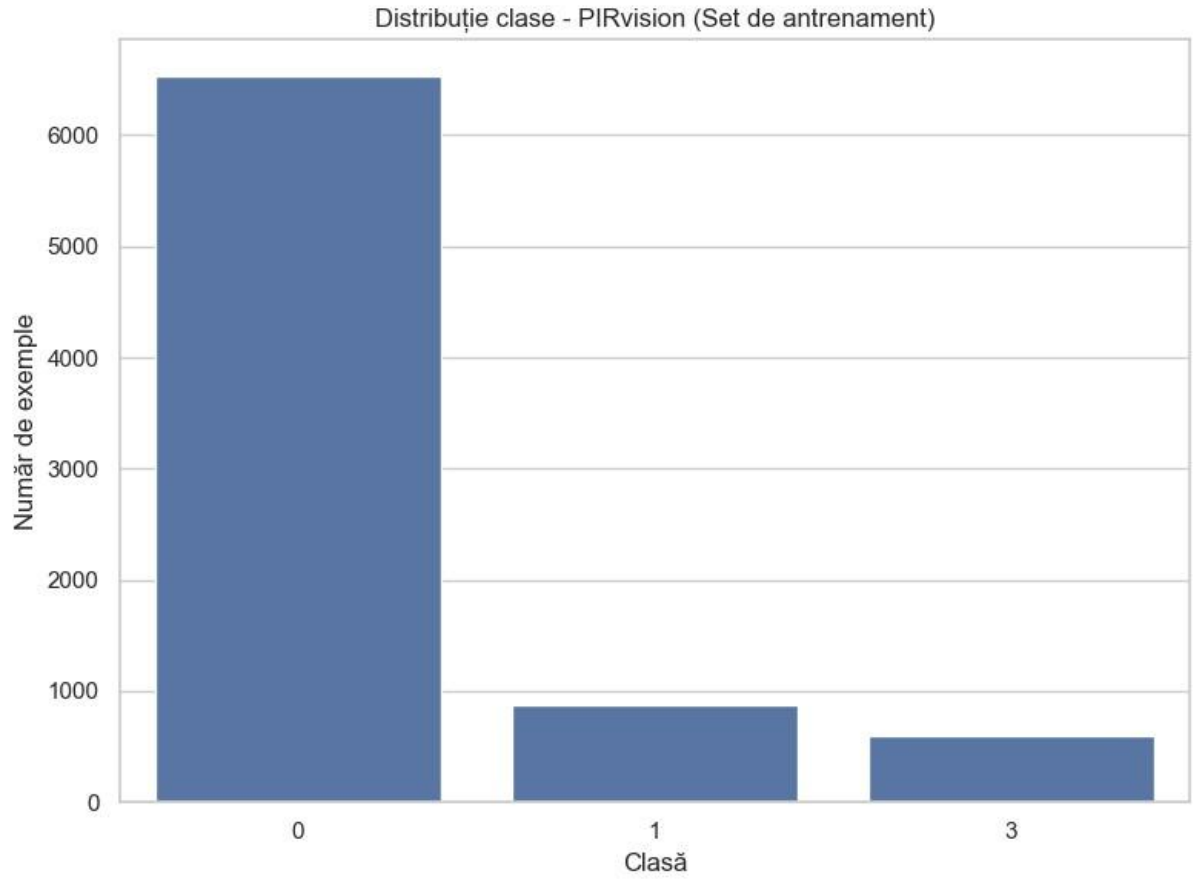
Set de date dezechilibrat

Air pollution – test:



Distributie asemanatoare cu cea din train, insa, fiind un set de date cu raport ¼ din cel din train, asa sunt si valorile din grafic.

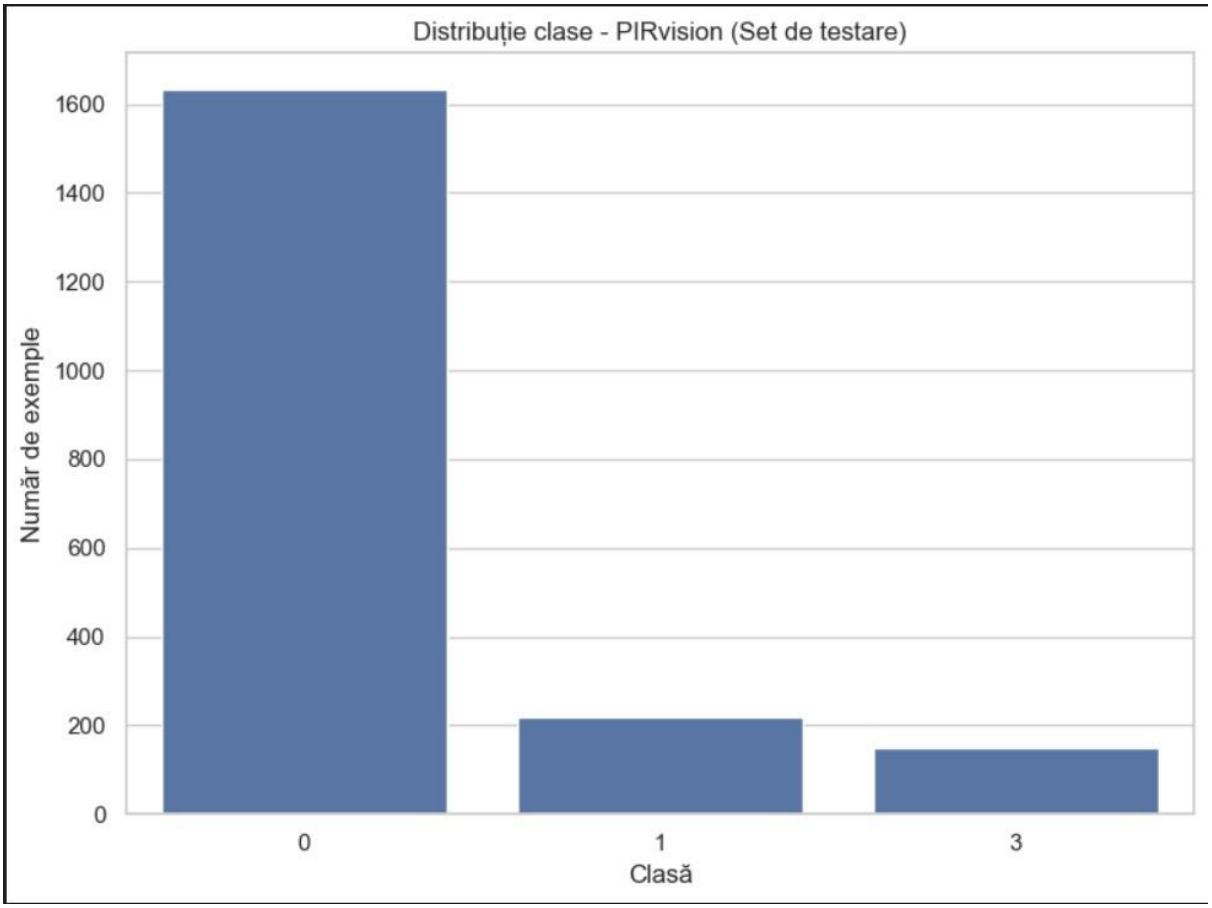
PIRvision – train :



- clasa 0 are peste 6500 exemple, clasa dominanta
- clasa 1 sub 1000 exemple
- clasa 2 sub 700 exemple

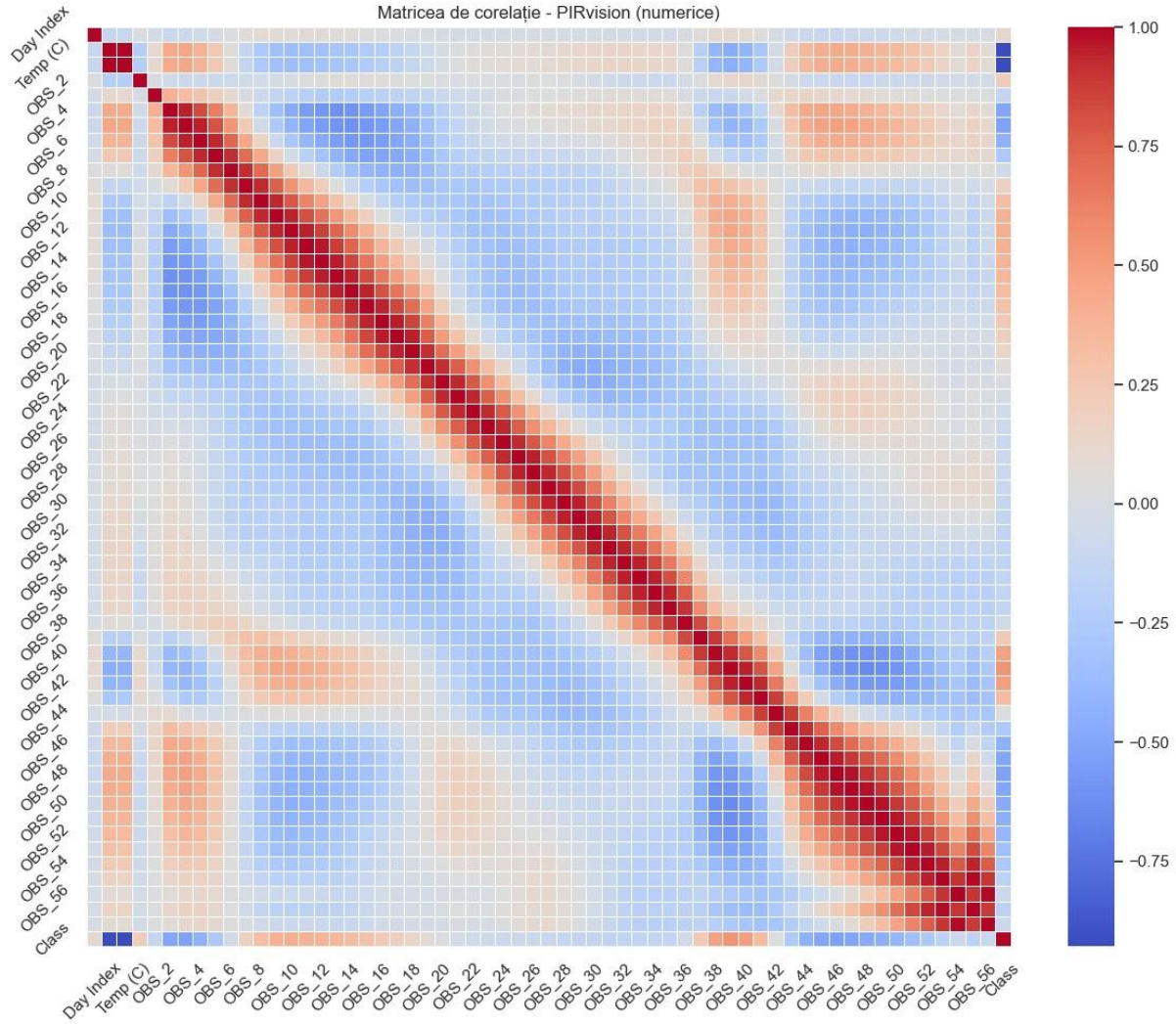
Dezechilibru intre clase. Modelul tinde sa favorizeze clasa 0.

PIRvision – test:



Acelasi lucru ca la Air Pollution. Setul de test are raport ¼ din setul de train si asa sunt si valorile fata de cele din train.

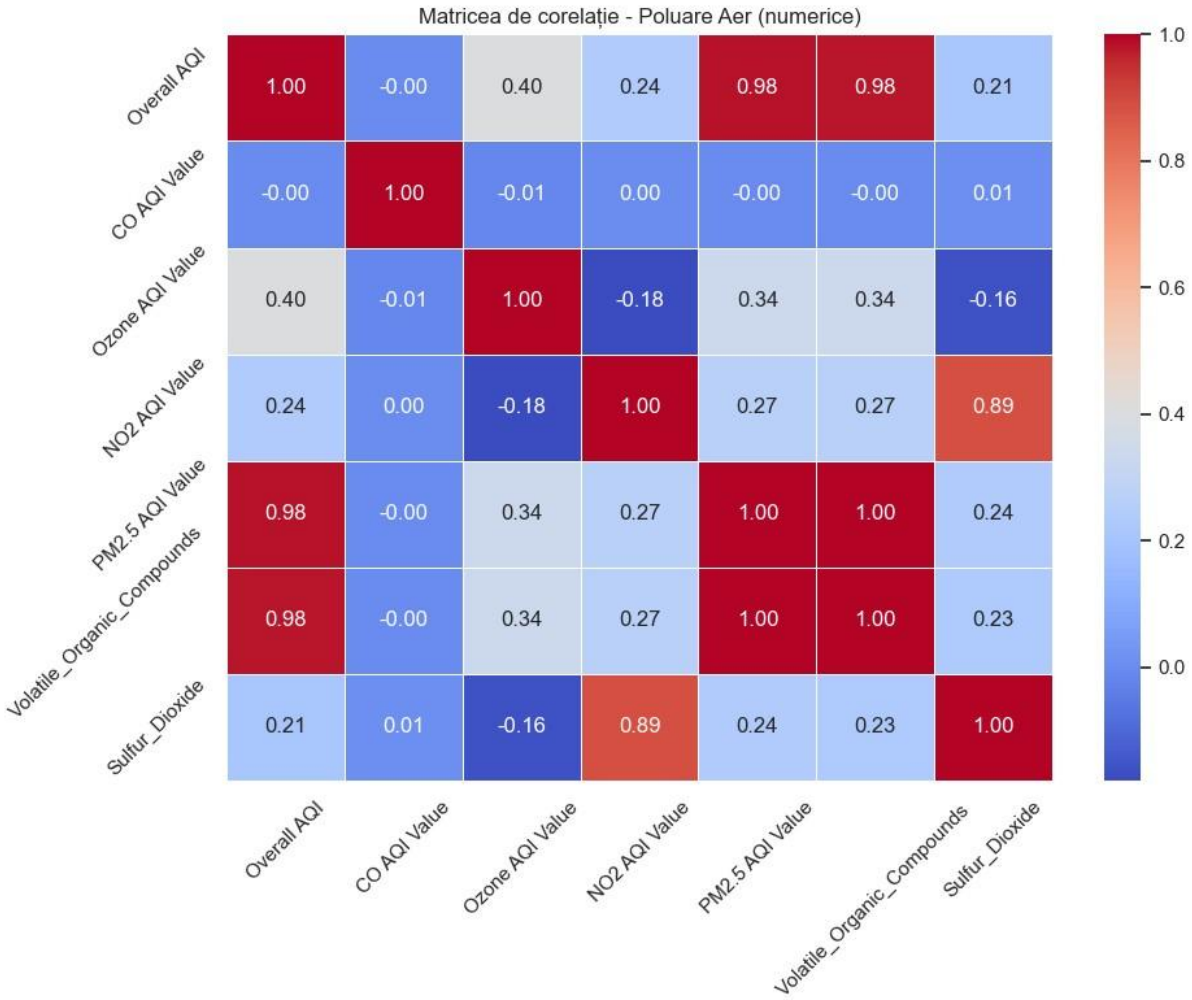
3. Analiza corelației între atribute PIRvision: corelații între atribute numerice



-Observ banda rosie diagonala care confirma ca fiecare atribut este corelat perfect cu el insusi.
Foarte multe dintre atributele OBS1-57 sunt puternic corelate intre ele, ceea ce este suprapunere informationala si un potential de eliminare a unora dintre ele.
-Temp(C) si Temp(F) sunt perfect corelate negative deoarece exprima aceeasi temperature in unitati diferite.
-“Day Index” are corelatii slabe cu celelalte, cee ace indica faptul ca nu este redundant.

-“Class” are o corelatie slaba cu majoritatea atributelor

Poluare aer: corelații între atribute numerice



-“Overall AQI”, “PM2.5 AQI Value” si “Volatile Organic Compounds” sunt puternic corelate intre ele(coef. 0.98-1.00). Ele contin informatie suprapusa

-“Sulfur Dioxide” este puternic corelat cu “NO2 AQI Value”(0.89), cee ace sugereaza ca pot reflecta surse commune de poluare sau acelasi tip de activitate industriala. -Atributele precum “CO AQI Value” sau “Ozone AQI Value” au

corelatii foarte slabe cu celelalte variabile, ceea ce inseamna ca sunt relative independente si pot aduce informatie suplimentara.
Poluare aer: analiză relații între categorice (frecvență combinată)

Preprocesarea datelor

1. Date lipsă pentru un atribut într-un eșantion

```
PIRvision - Valori lipsă:
Timestamp    788
Day          790
OBS_1        798
dtype: int64

PIRvision - Valori lipsă în setul de testare:
Timestamp    212
Day          210
OBS_1        202
dtype: int64

Air Pollution - Valori lipsă:
Country Name         349
Ozone AQI Value     1870
NO2 AQI Category    1893
dtype: int64

Air Pollution - Valori lipsă în setul de testare:
Country Name         78
City Name             1
Ozone AQI Value     476
NO2 AQI Category    453
dtype: int64
```

Aceastea sunt datele lipsa din ambele seturi de date. Pentru imputare am folosit media aritmetica pentru atributele numerice(“OBS_1” si “Ozone AQI Value), folosind(strategy=”mean”), astfel incat distributia sa fie mentinuta cat mai aproape de cea initiala, iar pentru “NO2 AQI Category” si “Country name” am folosit valoarea cea mai frecventa(strategy=”most_frequent”) presupunand ca aceasta reflecta cel mai probabil contextual lipsa.

2. Valori extreme pentru un atribut într-un eșantion Pentru a identifica valorile extreme, am folosit regula intercuartilica: un punct este considerat outlier daca se afla in afara intervalului Q1−1.5×IQR,Q3+1.5×IQRQ1 - 1.5 × IQR, Q3 + 1.5 × IQRQ1−1.5×IQR,Q3+1.5×IQR, unde Q1 este percentila 25, Q3 percentila 75, iar IQR = Q3 − Q1. Valorile extreme au fost apoi marcate ca lipsa (NaN), iar imputarea s-a realizat folosind media aritmetica a fiecarei coloane.

3. Atribute redundante (puternic corelate)
In urma analizei de la pasul 3.1, am identificat atribute numerice puternic corelate intre ele, ceea ce indica redundanta informationala. Astfel, am aplicat o eliminare a atributelor care au un coefficient de corelatie mai mare decat:
-0.95 in setul PIRvision(“OBS_48”, “OBS_49”, “OBS_50”)
-0.9 in serul Poluare Aer(“PM2.5 AQI Value”, “Volatile_Organic_Compounds”)

4. Plaje valorice de mărimi diferite pentru atributele numerice
Pentru a preveni diferentele de scala sa influenteze negative antrenarea modelelor de clasificare(in general regresia logistica) am aplicat standardizarea atributelor numerice, transformand fiecare coloana numerica astfel incat sa abia media = 0 si std = 1.

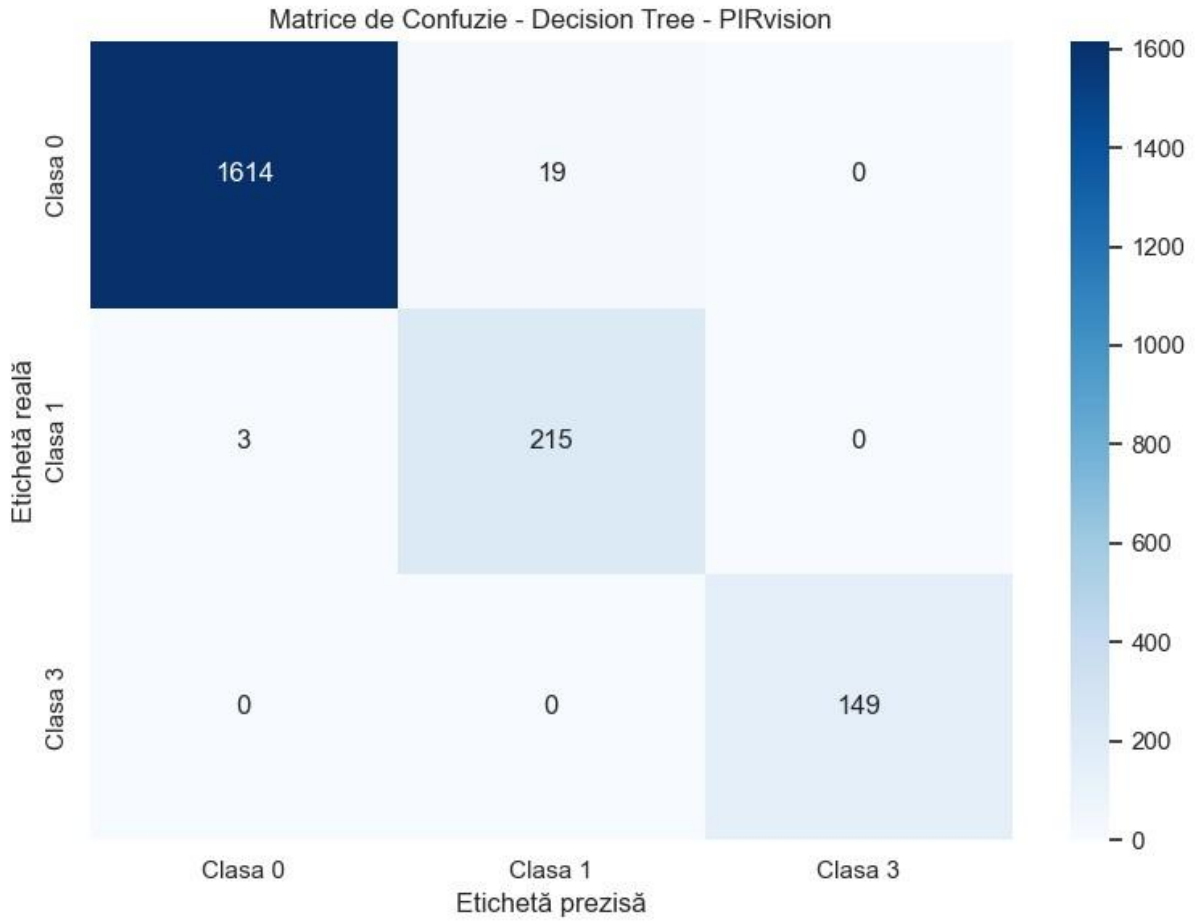
Utilizarea algoritmilor de Învățare Automată
Arbori de Decizie – PIRvision Hiperparametrii alesi:

-criterion=”gini” – pentru separarea nodurilor -max_depth=10 – limitam adancimea pentru a evita overfitting

-min_samples_leaf=5 – minim 5 exemeple per frunza, pentru regularizare

-class_weight=”balanced” – ajusteaza automata pentru clase dezechilibrate

-random_state=42 – pentru reproducibilitate




```
=== Evaluare Decision Tree - PIRvision ===
Acuratețe: 0.9890
              precision    recall  f1-score   support

     Clasa 0       1.00      0.99      0.99      1633
     Clasa 1       0.92      0.99      0.95       218
     Clasa 3       1.00      1.00      1.00       149

    accuracy                           0.99      2000
   macro avg       0.97      0.99      0.98      2000
weighted avg       0.99      0.99      0.99      2000
```

Am obtinut o acuratete de 98.9%

Modelul se descurca excellent in identificarea claselor, inclusive cele minoritate.

Confuziile apar rareori intre clasele 0 si 1.

Parametrul class_weight=”balanced” a fost efficient in combaterea dezechilibrului dintre clase(clasa 0 avea de 8 ori mai multe exemple declat clasa 3)

Max_depth=10 a prevenit overfitting

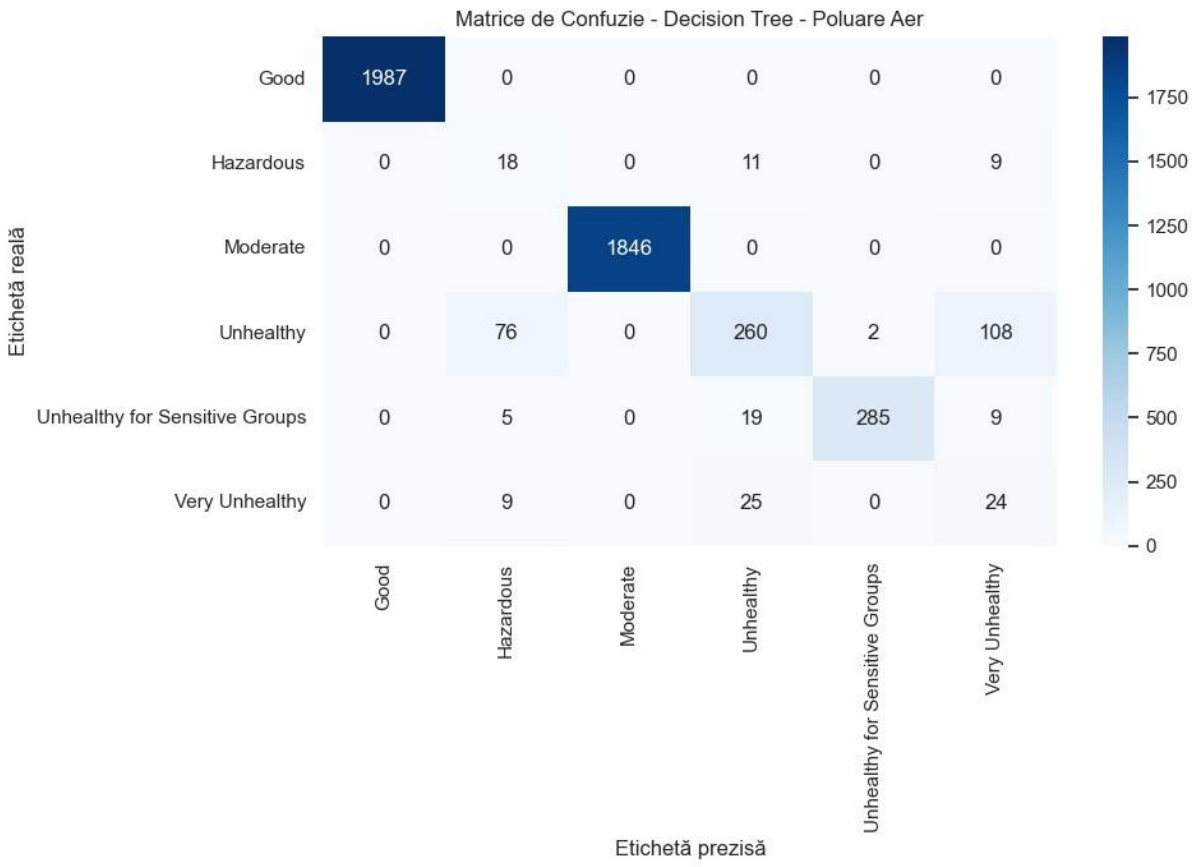
Arbori de Decizie – Air Pollution Hiperparametrii alesi:

-criterion=”gini” – pentru separarea nodurilor -max_depth=10 – limitam adancimea pentru a evita overfitting

-min_samples_leaf=5 – minim 5 exemeple per frunza, pentru regularizare

-class_weight=”balanced” – ajusteaza automata pentru clase dezechilibrate

-random_state=42 – pentru reproducibilitate



```
=== Evaluare Decision Tree - Poluare Aer ===
Acuratețe: 0.9418
                            precision    recall  f1-score   support

                      Good       1.00      1.00      1.00      1987
                 Hazardous       0.17      0.47      0.25        38
                  Moderate       1.00      1.00      1.00      1846
                 Unhealthy       0.83      0.58      0.68       446
Unhealthy for Sensitive Groups       0.99      0.90      0.94       318
            Very Unhealthy       0.16      0.41      0.23        58

                  accuracy                           0.94      4693
                 macro avg       0.69      0.73      0.68      4693
              weighted avg       0.97      0.94      0.95      4693
```

Am obtinut o acuratete de 94.18%

Clasele “Good” si “Moderate” sunt foarte aproape perfect clasificate, avand un volum mare de date si valori numerice bine separate.

Clasele “Hazardous” si “Very Unhealthy” sunt greu de separate din cauza numarului redus de exemple si suprapunerii cu alte clase.

Parametrul class_weight=”balanced” a imbunatatit vizibil scorurile pentru clasele mici

Păduri Aleatoare – PIRvision

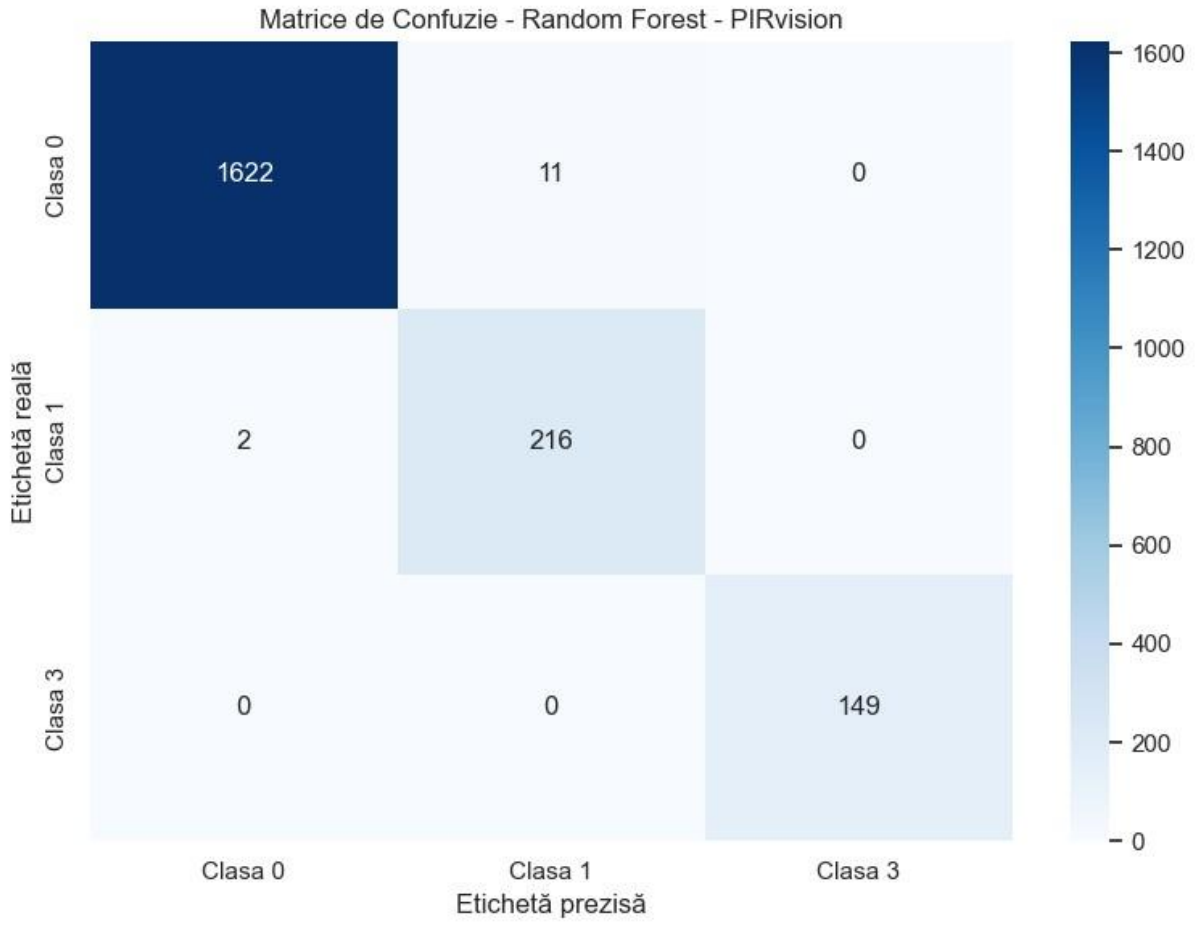
-n_estimators=100 – modelul este compus din 100 arbori -max_depth=12 – limitam adancimea fiecarui arbore pentru a evita overfitting

-min_samples_leaf=5 – minim 5 exemeple per frunza, pentru regularizare

-max_features=”sqrt” – selectare randomizata a atributelor

-class_weight=”balanced” – ajusteaza automata pentru clase dezechilibrate

-random_state=42 – pentru reproducibilitate



```
=== Evaluare Random Forest - PIRvision ===
Acuratețe: 0.9935
              precision    recall  f1-score   support

     Clasa 0       1.00      0.99      1.00      1633
     Clasa 1       0.95      0.99      0.97       218
     Clasa 3       1.00      1.00      1.00       149

    accuracy                           0.99      2000
   macro avg       0.98      0.99      0.99      2000
weighted avg       0.99      0.99      0.99      2000
```

Am obtinut o acuratete de 99.35%

Modelul Random Forest a oferit performante mai bune decat arborele de decizie

Clasa 3 a fost clasificata perfect

Clasa 1 cu scor F1 de 0.97

Class_weight=”balanced” a contribuit la cresterea scorului pentru clasele minoritare

Paduri aleatoare – Air Pollution

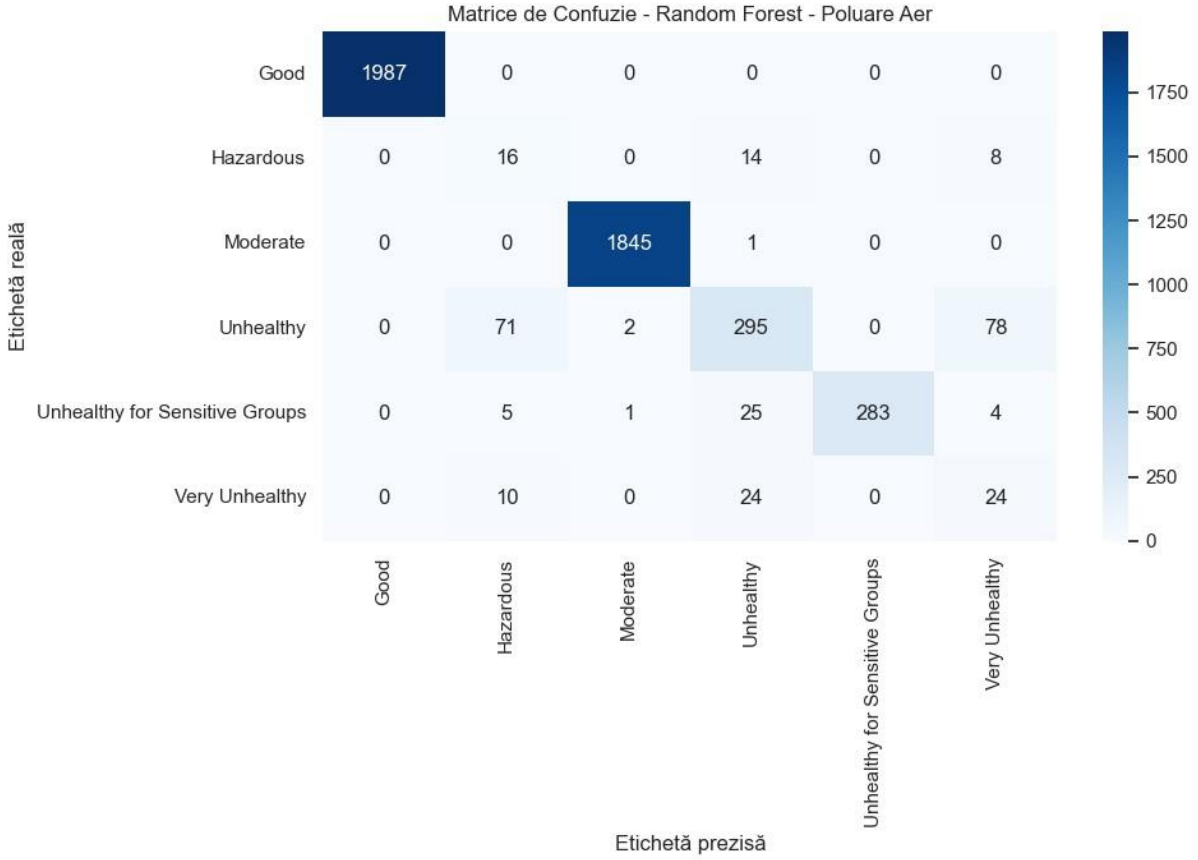
-n_estimators=100 – modelul este compus din 100 arbori -max_depth=14 – limitam adancimea fiecarui arbore pentru a evita overfitting

-min_samples_leaf=5 – minim 5 exemeple per frunza, pentru regularizare

-max_features=”sqrt” – selectare randomizata a atributelor

-class_weight=”balanced” – ajusteaza automata pentru clase dezechilibrate

-random_state=42 – pentru reproducibilitate



```
=== Evaluare Random Forest - Poluare Aer ===
Acuratețe: 0.9482
                              precision    recall  f1-score   support

                        Good       1.00      1.00      1.00      1987
                   Hazardous       0.16      0.42      0.23        38
                    Moderate       1.00      1.00      1.00      1846
                   Unhealthy       0.82      0.66      0.73       446
Unhealthy for Sensitive Groups       1.00      0.89      0.94       318
              Very Unhealthy       0.21      0.41      0.28        58

                    accuracy                           0.95      4693
                   macro avg       0.70      0.73      0.70      4693
                weighted avg       0.97      0.95      0.96      4693
```

Am obtinut o acuratete de 94.82%

Clasele dominante(Good, Moderate) sunt clasificate excellent(F1=1.00)

Clasele rare precum “Hazardous” si “Very Unhealthy” raman dificil de distins

Fata de arborele de decizie, scorurile pe clasele medii s-au imbunatatit.

Random Forest ofera o generalizare mai buna datorita agregarii multiplelor modele slab corelate.

Regresie Logica - PIRvision

Tipul de encodare folosit:
Variabila tinta Class (numerica, dar tratata drept categorie) a fost transformata cu LabelEncoder.

Atributele de intrare (predictori) sunt deja numerice, deci nu a fost necesara codificare suplimentara (ex. One-Hot Encoding).

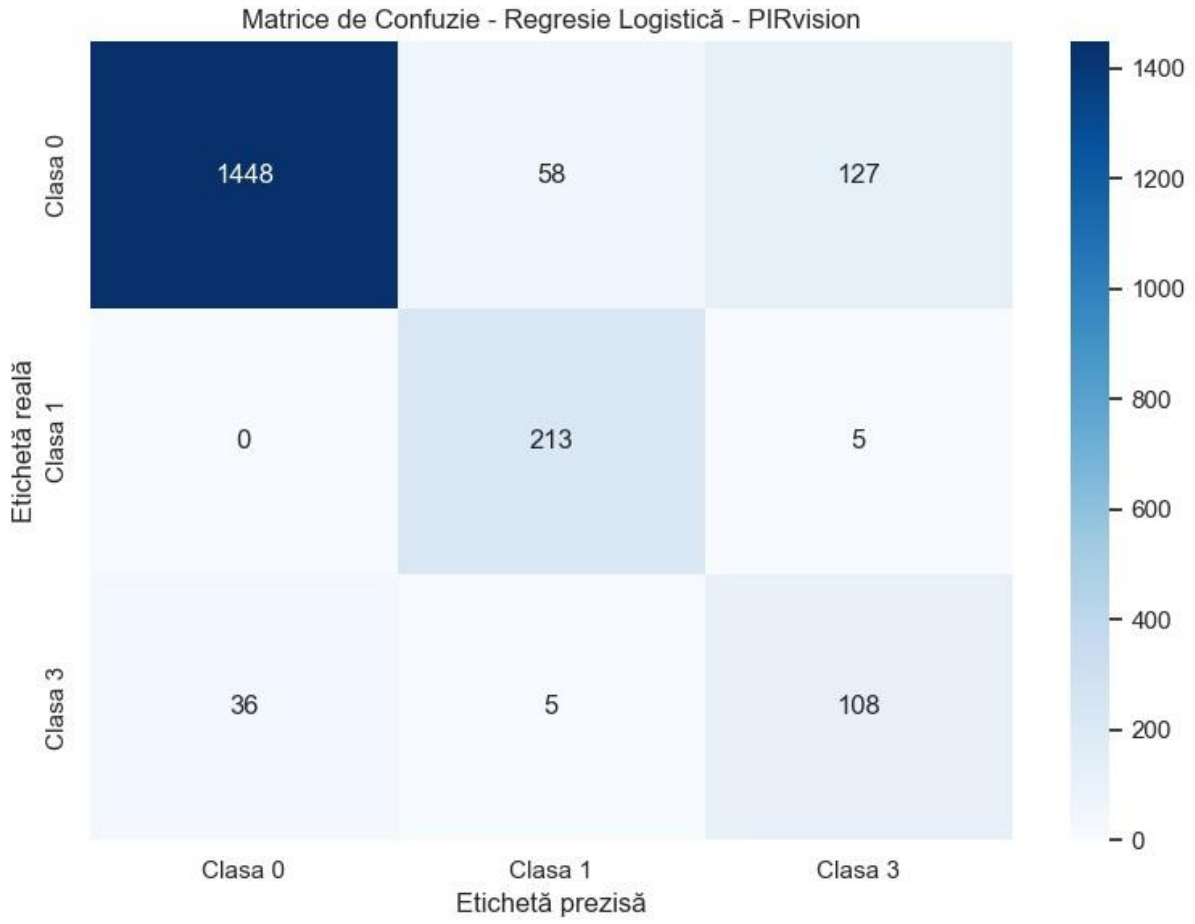
Optimizator (Gradient Descent):
A fost folosit lbfgs, un optimizator de tip quasi-Newton, recomandat pentru regresie logistica multiclasa.

Parametrul max_iter=1000 a fost setat pentru a asigura convergenta in cazul unui set de date cu multe variabile.

Regularizare:
S-a folosit regularizarea L2 (implicita), care penalizeaza coeficientii mari pentru a reduce overfitting-ul.

Nu s-a setat explicit un coeficient C, deci a fost folosita valoarea implicita C=1.0, adica regularizare moderata.

class_weight="balanced" a fost utilizat pentru a corecta dezechilibrul intre clase.




```
=== Evaluare Regresie Logistică - PIRvision ===
Acuratețe: 0.8845
              precision    recall  f1-score   support

     Clasa 0       0.98      0.89      0.93      1633
     Clasa 1       0.77      0.98      0.86       218
     Clasa 3       0.45      0.72      0.56       149

    accuracy                           0.88      2000
   macro avg       0.73      0.86      0.78      2000
weighted avg       0.91      0.88      0.89      2000
```

Am obtinut o acuratete de 88.45% Clasa majoritara(0) are o precizie excelenta, dar recall-ul scade la 0.89(unele exemple fiind confundate cu Clasa 3) Clasa 1 este foarte bine identificata(recall = 0.98)

Clasa 3 este slab recunoscuta(precision = 0.45), cee ace reflecta dificultatea regresiei logistice in a invata frontier clare pentru claase mici, mai alex in contexte nelineara

Regresie Logistica – Poluare Aer Tipul

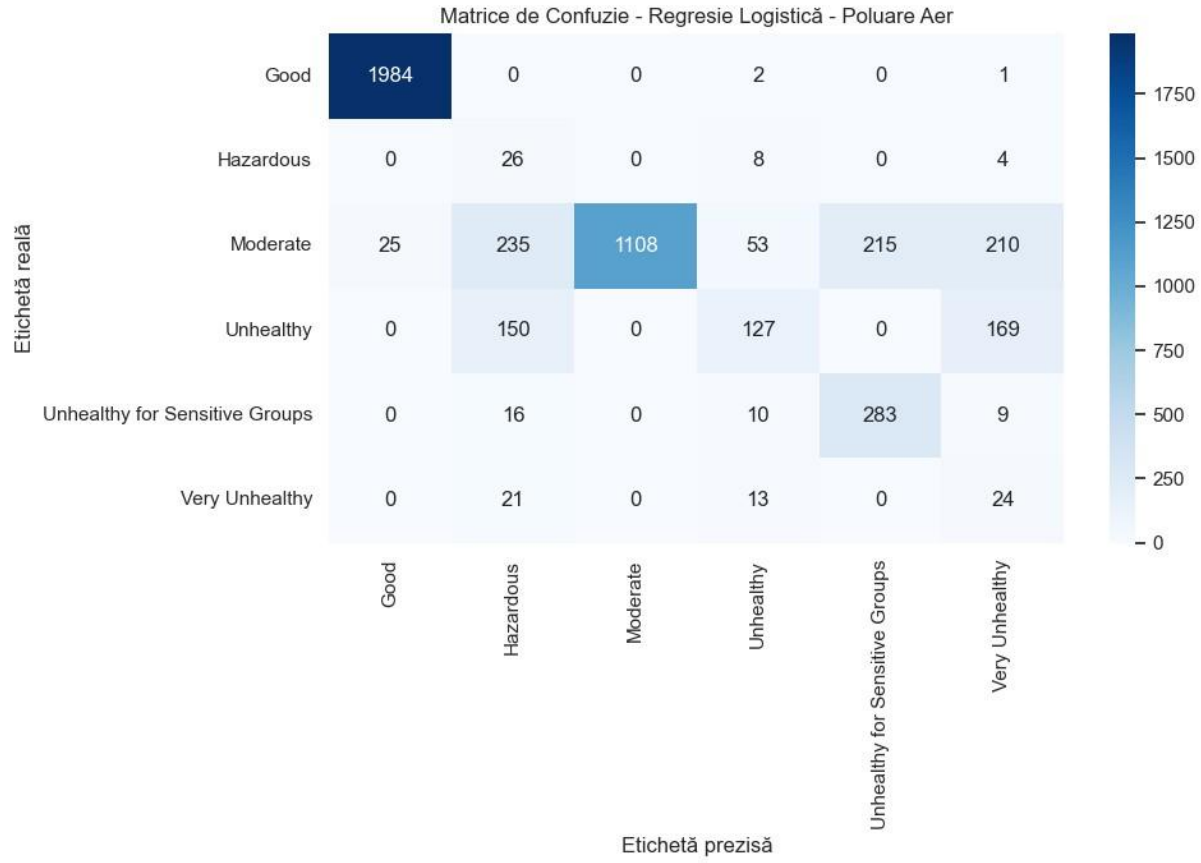
de encodare folosit:

Variabila tinta Overall AQI Level, fiind de tip categoric, a fost codificata numeric folosind LabelEncoder.
Atributele de intrare folosite in model au fost exclusiv numerice (e.g. CO AQI Value, Ozone AQI Value, NO2 AQI Value, etc.), deci nu a fost necesara codificarea suplimentara prin OneHotEncoder.

Optimizator (Gradient Descent):

Modelul a fost antrenat folosind algoritmul lbfgs, un optimizator de tip quasi-Newton, recomandat pentru regresie logistica multiclasa.
S-a setat max_iter=1000 pentru a permite convergenta in cazul unui volum mare de date si multiple clase tinta.
Regularizare:

S-a utilizat regularizarea implicita L2, care penalizeaza coeficientii mari si reduce overfitting-ul.
Parametrul C a fost lasat la valoarea implicita (C=1.0), echivalent cu regularizare moderata.
Pentru a gestiona dezechilibrul intre clase, a fost setat class_weight="balanced".



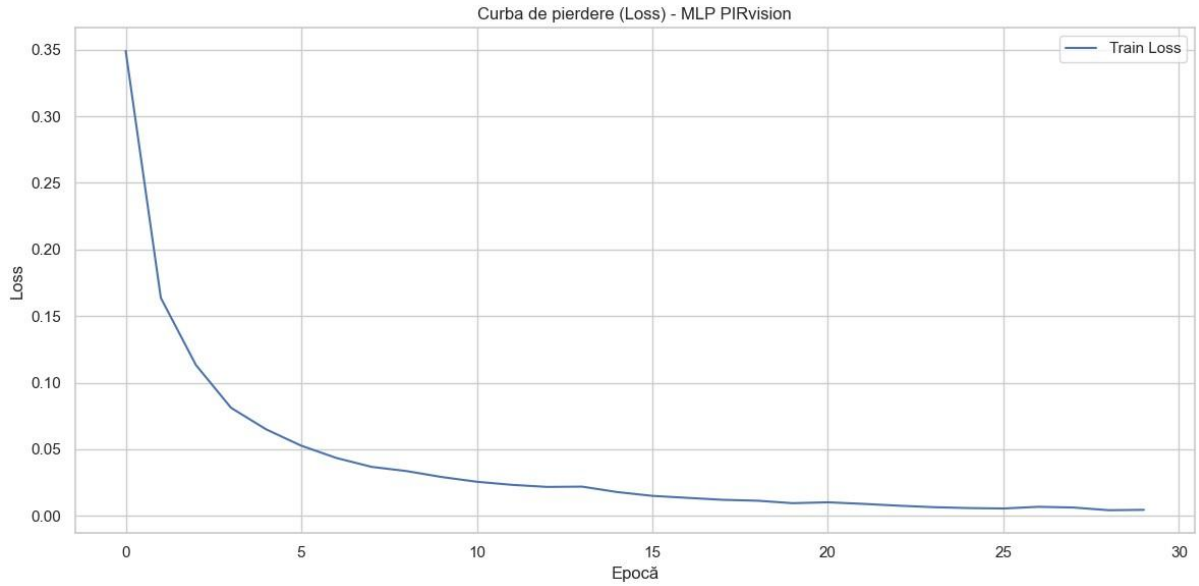

```
=== Evaluare Regresie Logistică - Poluare Aer ===
Acuratețe: 0.7569
                              precision    recall  f1-score   support

                        Good       0.99      1.00      0.99      1987
                   Hazardous       0.06      0.68      0.11        38
                    Moderate       1.00      0.60      0.75      1846
                   Unhealthy       0.60      0.28      0.39       446
Unhealthy for Sensitive Groups       0.57      0.89      0.69       318
              Very Unhealthy       0.06      0.41      0.10        58

                    accuracy                           0.76      4693
                   macro avg       0.54      0.65      0.51      4693
                weighted avg       0.91      0.76      0.80      4693
```

Am obtinut o acuratete de 75.69%

Modelul reuseste sa clasifice corect clasele dominante precum “Good”(precizie si recall ~1.00), dar performanta este scazuta pentru clasele rare precum “Hazardous” sau “Very Unhealthy”, cu scoruri f1-score de 0.11 si 0.10

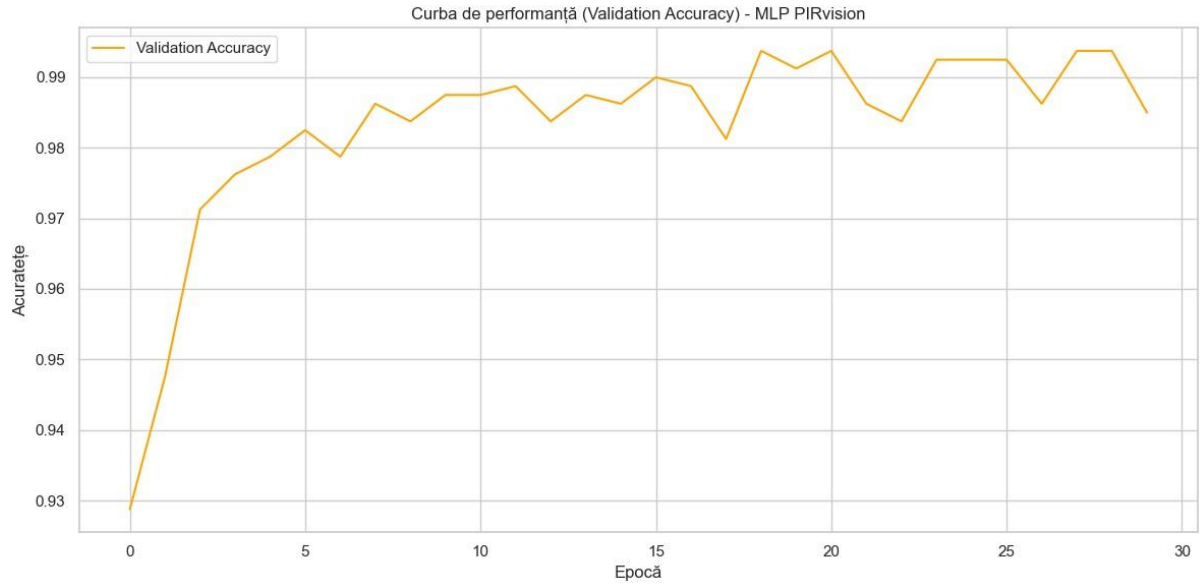
Rezultatele reflecta un dezechilibru al claselor si o dificultate ridicata in generalizarea pe clasele minoritare, in special din cauza suprapunerii valorilor atributelor numerice.

MLP – PIRvision



Curba de pierdere (Loss):

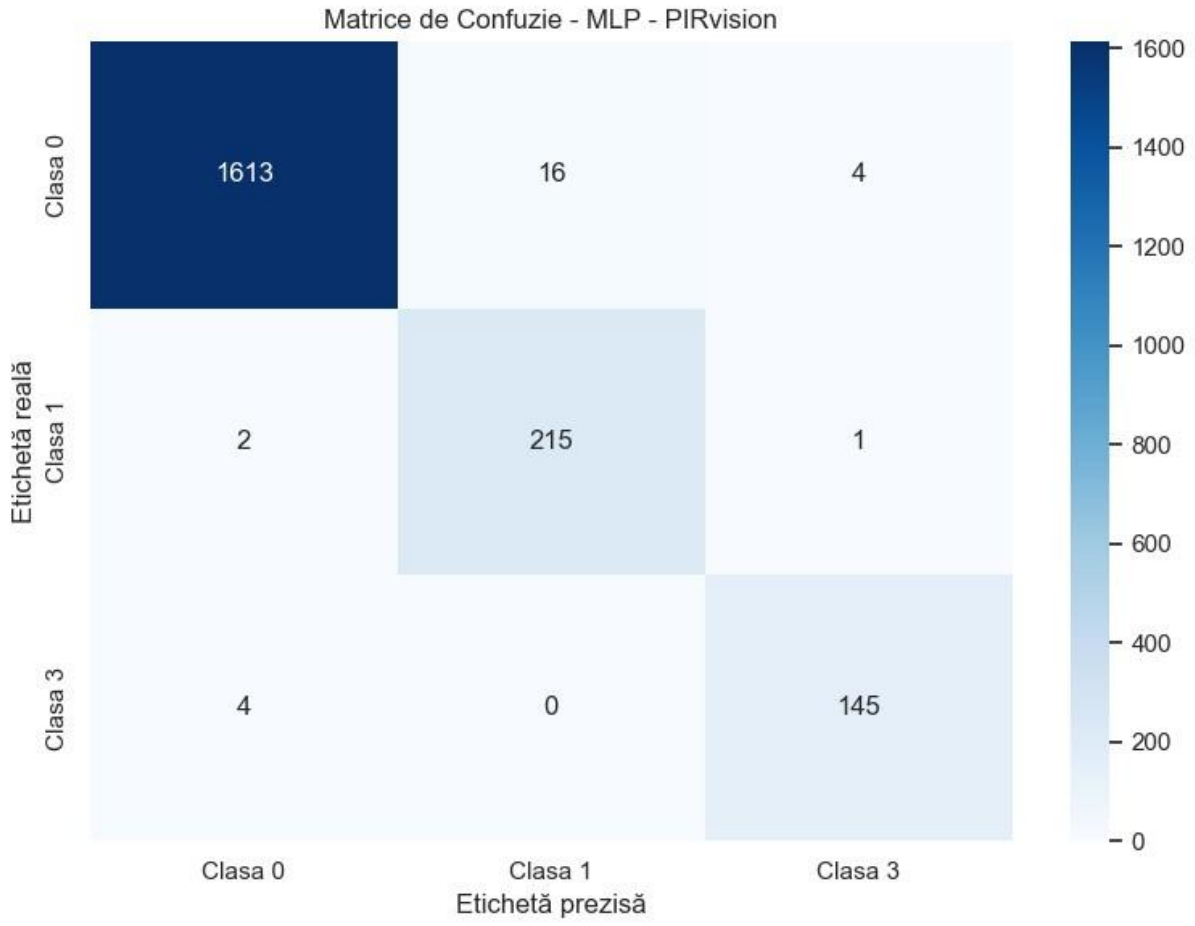
Modelul a invatat eficient din datele de antrenament. Valoarea functiei de pierdere scade constant in primele epoci si se stabilizeaza in jurul unei valori foarte mici dupa aproximativ 20 de epoci. Aceasta evolutie sugereaza ca modelul converge corect, fara fluctuatii majore sau semne de suprainvatare.



Curba de acuratete (Validation Accuracy):

Acuratetea pe setul de validare creste rapid in primele epoci

si se mentine la un nivel ridicat (intre 0.98 si 0.993). Oscilatiile minore sunt normale si nu indica overfitting. Nu exista o scadere semnificativa care sa sugereze pierderea capacitatii de generalizare.




```
=== Evaluare MLP - PIRvision ===
Acuratețe: 0.9865
              precision    recall  f1-score   support

     Clasa 0       1.00      0.99      0.99      1633
     Clasa 1       0.93      0.99      0.96       218
     Clasa 3       0.97      0.97      0.97       149

    accuracy                           0.99      2000
   macro avg       0.96      0.98      0.97      2000
weighted avg       0.99      0.99      0.99      2000
```

Reteaua neuronala utilizata are o structura compusa din:

- 2 straturi ascunse cu:
  - Primul strat: 64 neuroni
  - Al doilea strat: 32 neuroni
- Functia de activare folosita: ReLU (Rectified Linear Unit)
- Strat de iesire: clasificare multiclasa (activare implicita softmax)

Optimizator si hiperparametri

- Optimizator: Adam (adaptive gradient-based optimizer)
- Learning rate initial: 0.001
- Numar maxim de epoci (max_iter): 300
- Batch size: 64
- Early stopping: activat (pentru a preveni overfitting-ul)
- Random state: 42

Regularizare

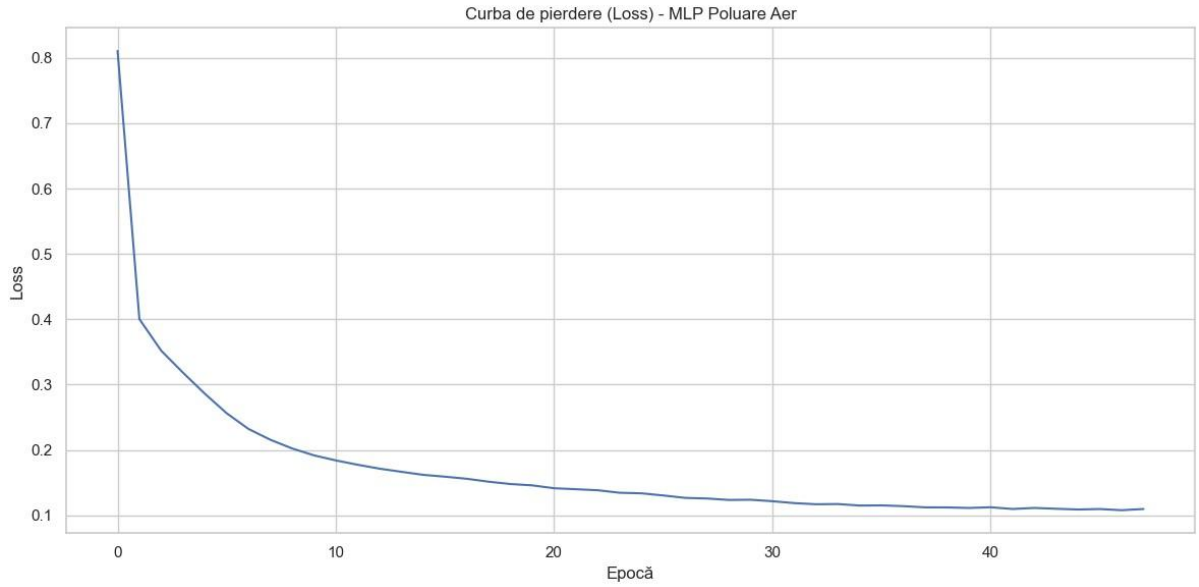
- A fost folosita regularizare L2 cu coeficient alpha = 0.001
- Early stopping monitorizeaza performanta pe un set de validare intern si opreste antrenarea cand modelul nu se mai imbunatateste

Rezultate obtinute

- Acuratete totala: 0.9865

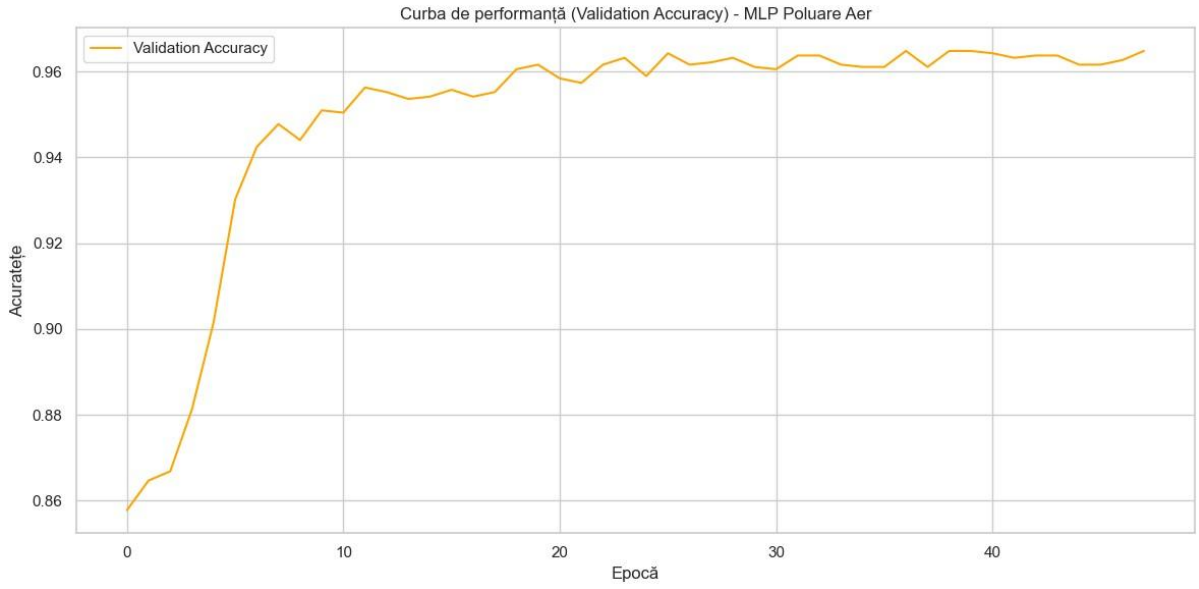
- Matricea de confuzie a aratat ca modelul clasifica corect majoritatea observatiilor pentru toate cele 3 clase (0, 1, 3)
- Curba de pierdere (loss) arata o scadere constanta a erorii in timpul antrenarii, fara semne evidente de overfitting
- Graficul curbei de pierdere este inclus si arata convergenta rapida in aproximativ 30 epoci

MLP - Poluare Aer



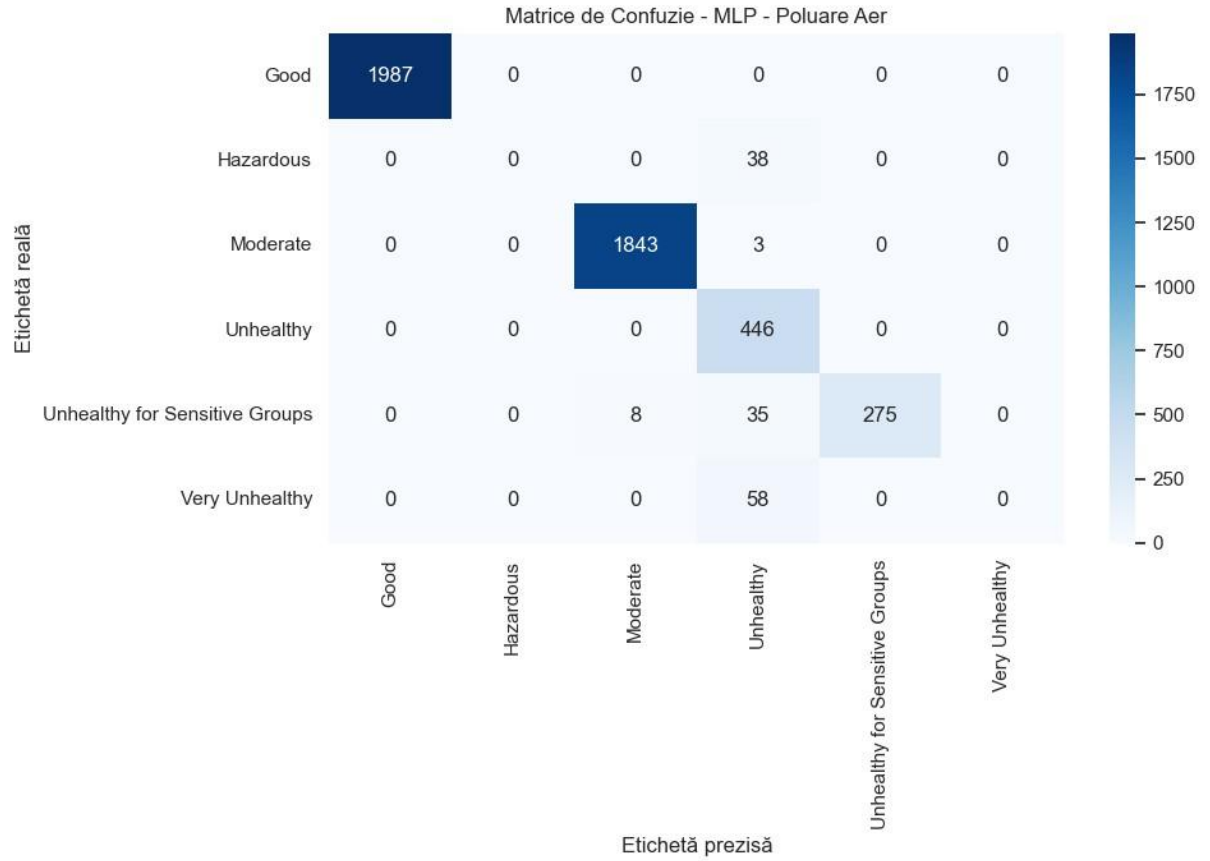
1. Curba de pierdere (Loss)

Graficul arata o scadere constanta a erorii (loss-ului) pe parcursul epocilor, ceea ce indica faptul ca modelul invata in mod eficient. Nu exista fluctuatii sau stagnari semnificative, iar curba devine plata spre final, semn ca modelul se apropie de o convergenta stabila. Acest comportament sugereaza ca modelul nu sufera de overfitting si ca antrenarea s-a realizat in conditii controlate.



2. Curba de performanta (Validation Accuracy)

Aceasta curba urca rapid in primele epoci, atingand o acuratete de validare ridicata (~96%) si ramanand constanta ulterior. Acest lucru confirma ca modelul generalizeaza bine si ca nu exista semne evidente de overfitting (nu scade acuratetea de validare in timp). Micile oscilatii sunt normale si pot fi cauzate de variatia setului de validare.




```
=== Evaluare MLP - Poluare Aer ===
Acuratețe: 0.9697
                            precision    recall  f1-score   support

                      Good       1.00      1.00      1.00      1987
                 Hazardous       0.00      0.00      0.00        38
                  Moderate       1.00      1.00      1.00      1846
                 Unhealthy       0.77      1.00      0.87       446
Unhealthy for Sensitive Groups       1.00      0.86      0.93       318
            Very Unhealthy       0.00      0.00      0.00        58

                  accuracy                           0.97      4693
                 macro avg       0.63      0.64      0.63      4693
              weighted avg       0.96      0.97      0.96      4693
```

Arhitectura rețelei:

- Număr straturi ascunse: 2

- Dimensiunea straturilor: primul strat cu 64 de neuroni, al doilea cu 32 de neuroni
- Funcția de activare: ReLU (activation='relu') Optimizator:
- Tip optimizator: Adam (solver='adam')
- Learning rate inițial: 0.001 (learning_rate_init=0.001)
- Număr maxim de epoci: 300 (max_iter=300)
- Dimensiune batch: 64 (batch_size=64) Metode de regularizare:
- Regularizare L2 cu coeficient alpha=0.001 (implicit în scikit-learn pentru MLPClassifier)
- Activarea opțiunii early_stopping=True pentru a preveni overfitting-ul prin oprirea antrenării atunci când performanța pe setul de validare nu mai crește

Tabel comparativ al performanței algoritmilor de clasificare pe seturile de date PIRvision și Poluare Aer

	Model	Accuracy	Precizie clasa 0	Recall clasa 0	F1 clasa 0	Precizie clasa 1	Recall clasa 1	F1 clasa 1	Precizie clasa 2	Recall clasa 2	F1 clasa 2	Precizie clasa 3	Recall clasa 3	F1 clasa 3	Precizie clasa 4	Recall clasa 4	F1 clasa 4
0	Decision Tree - PIRvision	0.9890	0.9981	0.9884	0.9932	0.9188	0.9862	0.9513	**1.0000**	**1.0000**	**1.0000**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	Random Forest - PIRvision	**0.9935**	0.9988	0.9933	0.9960	**0.9515**	**0.9908**	**0.9708**	**1.0000**	**1.0000**	**1.0000**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	Logistic Regression - PIRvision	0.8845	0.9757	0.8867	0.9291	0.7717	0.9771	0.8623	0.4500	0.7248	0.5553	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	MLP - PIRvision	0.9865	0.9963	0.9878	0.9920	0.9307	0.9862	0.9577	0.9667	0.9732	0.9699	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4	Decision Tree - Poluare Aer	0.9416	**1.0000**	**1.0000**	**1.0000**	0.1442	0.3947	0.2113	**1.0000**	**1.0000**	**1.0000**	**0.8179**	0.5942	0.6883	**1.0000**	**0.8899**	**0.9418**
5	Random Forest - Poluare Aer	0.9482	**1.0000**	**1.0000**	**1.0000**	0.1485	0.3947	0.2158	0.9984	**1.0000**	0.9992	0.8154	0.6637	0.7318	**1.0000**	**0.8899**	**0.9418**
6	Logistic Regression - Poluare Aer	0.7569	0.9876	0.9985	0.9930	0.0580	0.6842	0.1070	**1.0000**	0.6002	0.7502	0.5962	0.2848	0.3854	0.5683	**0.8899**	0.6936
7	MLP - Poluare Aer	0.9697	**1.0000**	**1.0000**	**1.0000**	0.0000	0.0000	0.0000	0.9957	0.9984	0.9970	0.7690	**1.0000**	**0.8694**	**1.0000**	0.8648	0.9275

1. Setul de date PIRvision

Pe acest set, toti cei patru algoritmi (Decision Tree, Random Forest, Logistic Regression, MLP) au obtinut scoruri ridicate, dar Random Forest este modelul cel mai performant:

-A obtinut cea mai mare acuratete (0.9935) si cele mai mari scoruri F1, precizie si recall pentru clasele predominante (clasa 0, clasa 1, clasa 2).

-Explicatie: Random Forest combina mai multi arbori de decizie si reduce suprainvatarea. Acest lucru ii permite sa generalizeze mai bine pe un set ca PIRvision, care pare sa aiba structuri de date bine definite intre clase.

-In plus, clasele 3, 4, 5 au scoruri 0 pentru toti algoritmii (ceea ce sugereaza ca ele sunt fie foarte rare, fie absente in setul de test). Acest lucru explica de ce si modelele foarte bune nu le pot prezice corect.

2. Setul de date Poluare Aer

Aici, MLP este modelul cu cea mai mare acuratete (0.9697), urmat de Random Forest (0.9482).

-Random Forest are o performanta buna si echilibrata, dar MLP reuseste sa atinga scoruri perfecte (1.0) pentru clasele 0 si 3 si scoruri ridicate pentru clasele 2 si 4.

-MLP functioneaza bine pentru ca poate modela relatii complexe nelineare in datele de mediu (care includ multi factori corelati – temperatura, poluanti etc.).

-Clasele 1 si 5 sunt problematice pentru toti algoritmii (mai ales Logistic Regression si MLP, care au scoruri foarte slabe sau 0). Acest lucru sugereaza ca aceste clase sunt dezechilibrate sau greu de separat, posibil din cauza suprapunerii in spatiul de trasaturi (features).

Concluzie generala

-Random Forest este cel mai constant algoritm performant pe ambele seturi de date – ofera o combinatie buna intre acuratete si robustete in fata dezechilibrelor de clasa.

-MLP reuseste performante excelente pe setul "Poluare Aer", datorita capacitatii sale de a invata modele complexe. Insa are dificultati cand anumite clase lipsesc sau sunt foarte rare.
-Logistic Regression este cel mai slab model pe ambele seturi, in special pe cel de poluare, deoarece relatiile dintre variabile sunt probabil nelineare si nu pot fi modelate bine de un model liniar.