ACADEMIA DE STUDII ECONOMICE BUCUREȘTI
FACULTATEA DE CIBERNETICĂ, STATISTICĂ ȘI INFORMATICĂ ECONOMICĂ

CLASIFICAREA CALITĂȚII AERULUI DUPĂ FACTORII DE MEDIU
(TEMA 2 – ANALIZĂ DISCRIMINANTĂ)

Profesor Coordonator: FURTUNĂ Titus Felix

Nume și prenume student: OPREA Ioana-Marcela

București

2025

Introducere

În contexul actual, calitatea aerului este un factor important pentru sănătatea oamenilor și pentru o bunăstare a comunității. Odată cu creșterea activității industriale, monitorizarea și clasificarea calității aerului au devenit priorități pentru cercetători. În proiectul de mai jos, este investigată calitatea aerului folosind o Analiză Liniară Discriminantă(LDA) și metode de Discriminare Bayesiană, aplicate pe un set de date reale ce conține măsurători pentru diverși factori de mediu.

Astfel, factorii folosiți în această analiză sunt:

- PM2.5 și PM10 – Particule în suspensie, fluide ori în stare solidă ce au compoziție și dimensiuni diverse și uneori sunt denumite "aerosoli";[1]
- NO2, SO2, CO – Gaze poluante;
- Temperatură;
- Umiditate;
- Densitatea populației;
- Apropierea de zone industriale.

Variabila țintă este calitatea aerului, împărțită în următoarele clase: „Good", „Moderate", „Poor", „Hazardous".

Astfel, obiectivul proiectului este de a oferi o înțelegere aprofundată a factorilor determinanți ai calității aerului și să identifice modele ce pot fi utilizate în predicții viitoare. Rezultatele pot avea un rol important în domeniul protecției mediului și pentru sănătatea publică.

Sursa datelor

Datele au fost preluate de pe platforma Kaggle, iar setul de date are surse precum World Health Organization (WHO) și World Bank Data, așa cum este precizat pe platformă, https://www.kaggle.com/datasets/mujtabamatin/air-quality-and-pollution-assessment .

[1] *PM 10 și PM 2.5 – ce înseamnă și ce efecte au, care sunt sursele de poluare în orașele din România și cum pot combate companiile contaminarea aerului*, in https://stratos.ro/pm10-si-pm2-5-ce-inseamna-si-ce-efecte-au/

Analiza discriminantă

Sub numele de analiză discriminantă, sunt reunite o serie de metode explicative, descriptive şi predictive, destinate studierii unei populații împărțite în clase. Fiecare individ este caracterizat printr-un ansamblu de variabile observate (independente) numite predictori şi o variabilă calitativă numită şi variabilă țintă identificând clasa din care face parte.

Seturile de date au fost împărțite în:

- Setul de învățare-testare – pollution.csv (corespunzător eșantionului de bază)
- Setul de aplicare – pollution_apply.csv (corespunzător eșantionului neinvestigat)

După ce am înlocuit cu media elementelor de pe coloană posibilele valori de 0 din set, pentru analiza liniară discriminantă am trecut prin următorii pași:

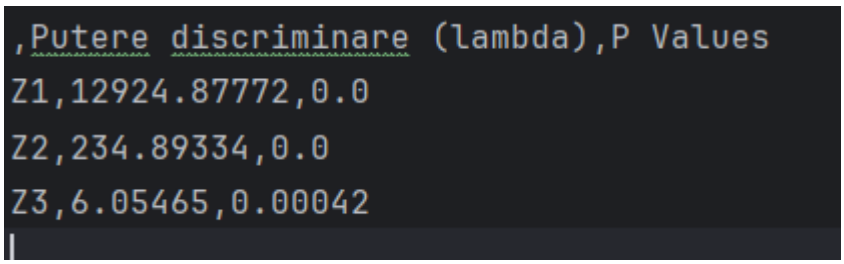
1. Am divizat datele în seturi de antrenament și testare. Mai exact, 70% din date vor fi pentru antrenare, iar 30% pentru testare.
2. Am contrsuit un model liniar pentru datele de antrenare.
3. Am calculat puterea de discriminare pentru predictori, obținând astfel puterea de discriminare pentru fiecare și p-value, așa cum se poate observa în figura de mai jos:

```
,Putere discriminare,p_values
Temperature,1210.8270340398924,1.1102230246251565e-16
Humidity,591.7459374345148,1.1102230246251565e-16
PM2.5,189.2329728803374,1.1102230246251565e-16
PM10,383.56712569950076,1.1102230246251565e-16
NO2,1526.768375342695,1.1102230246251565e-16
SO2,1118.5664760971529,1.1102230246251565e-16
CO,4714.425912536957,1.1102230246251565e-16
Proximity_to_Industrial_Areas,2110.950218309224,1.1102230246251565e-16
Population_Density,696.6330561938591,1.1102230246251565e-16
```

Predictori.csv

Astfel, rezultă faptul că CO, Proximity_to_Industrial_Areas și SO2 au cea mai mare putere de discriminare și sunt predictori importanți pentru diferențierea claselor.

4. Am calculat scorurile de discriminare proiectând datele din setul de antrenament pe axele de discriminare. De asemenea, am calculat puterea de discriminare pentru fiecare axă și p-value, date redate în figura de mai jos:

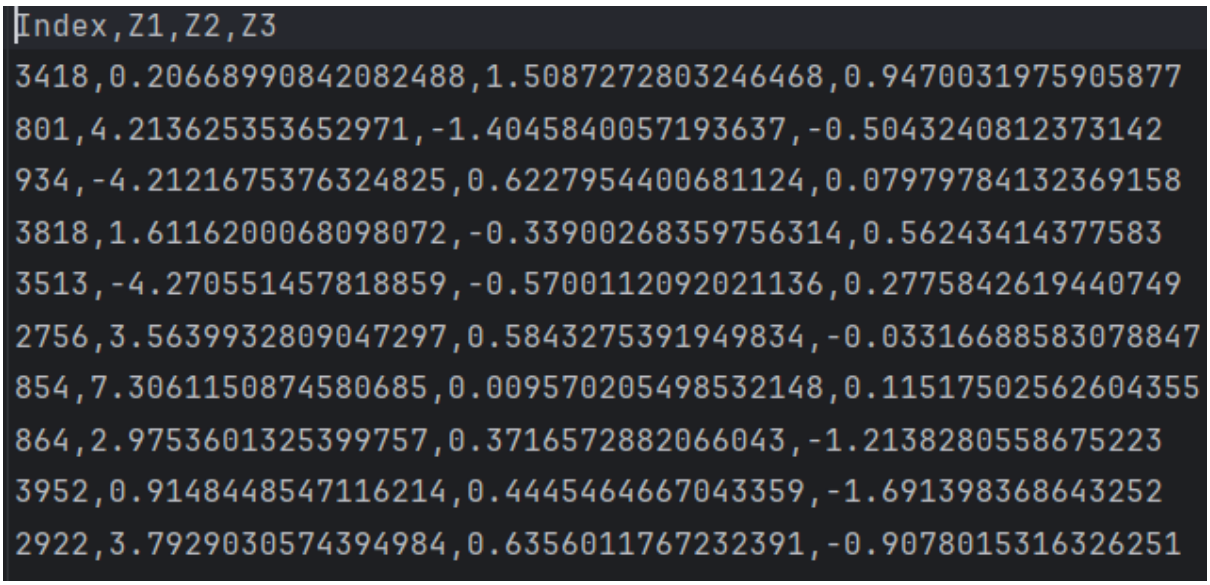


```
,Putere discriminare (lambda),P Values
Z1,12924.87772,0.0
Z2,234.89334,0.0
Z3,6.05465,0.00042
```

Discriminatori.csv

Astfel, se poate observa faptul că axa Z1 are cea mai mare putere de discriminare, ceea ce indică faptul că această axă explică cel mai bine variația dintre clase. P-value are o valoare mai mica de 0.05, ceea ce înseamnă că toate axele sunt semnificative pentru separarea grupelor.

5. În pasul următor, am calculat numărul de axe determinante. În urma formulei de calcul, a reieșit faptul că vom avea 3 axe de discriminare, întrucât minimul dintre numărul de clase-1 și numărul de predictori, este 3(numărul de clase-1).
6. Mai departe, vom calcula scorurile de discriminare pentru instanțele din setul de testare (mai exact poziția fiecărei instanțe pe fiecare axă de discriminare Z1,Z2,Z3), precum poate fi observat și în figura de mai jos, ce curpinde doar o parte din cele 1000 de instanțe.

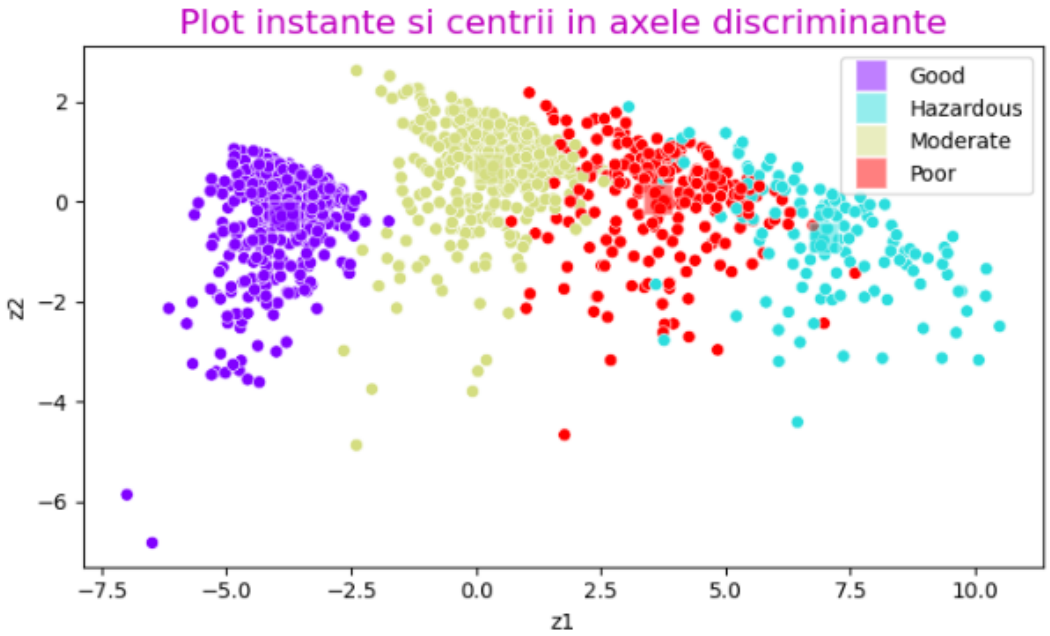


```
Index,Z1,Z2,Z3
3418,0.20668990842082488,1.5087272803246468,0.9470031975905877
801,4.213625353652971,-1.4045840057193637,-0.5043240812373142
934,-4.2121675376324825,0.6227954400681124,0.07979784132369158
3818,1.6116200068098072,-0.33900268359756314,0.56243414377583
3513,-4.270551457818859,-0.5700112092021136,0.2775842619440749
2756,3.5639932809047297,0.5843275391949834,-0.03316688583078847
854,7.3061150874580685,0.009570205498532148,0.11517502562604355
864,2.9753601325399757,0.3716572882066043,-1.2138280558675223
3952,0.9148448547116214,0.4445464667043359,-1.691398368643252
2922,3.7929030574394984,0.6356011767232391,-0.9078015316326251
```

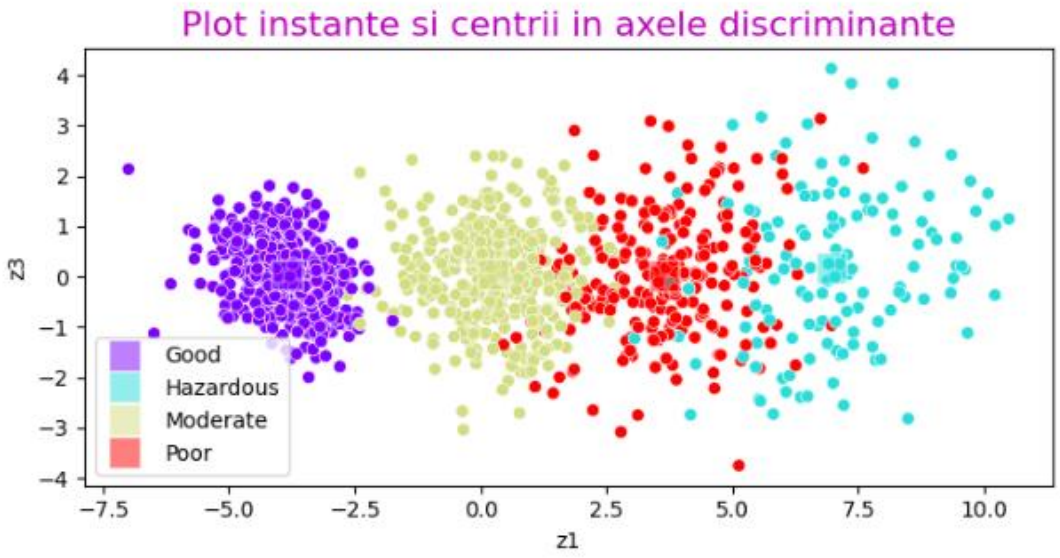
z.csv

7. Calculăm poziția medie a fiecărui grup pe axele de discriminare(centrul fiecărui grup), pentru a calsifica și vizuliza exact poziția pe fiecărui grup pe axa de discriminare.

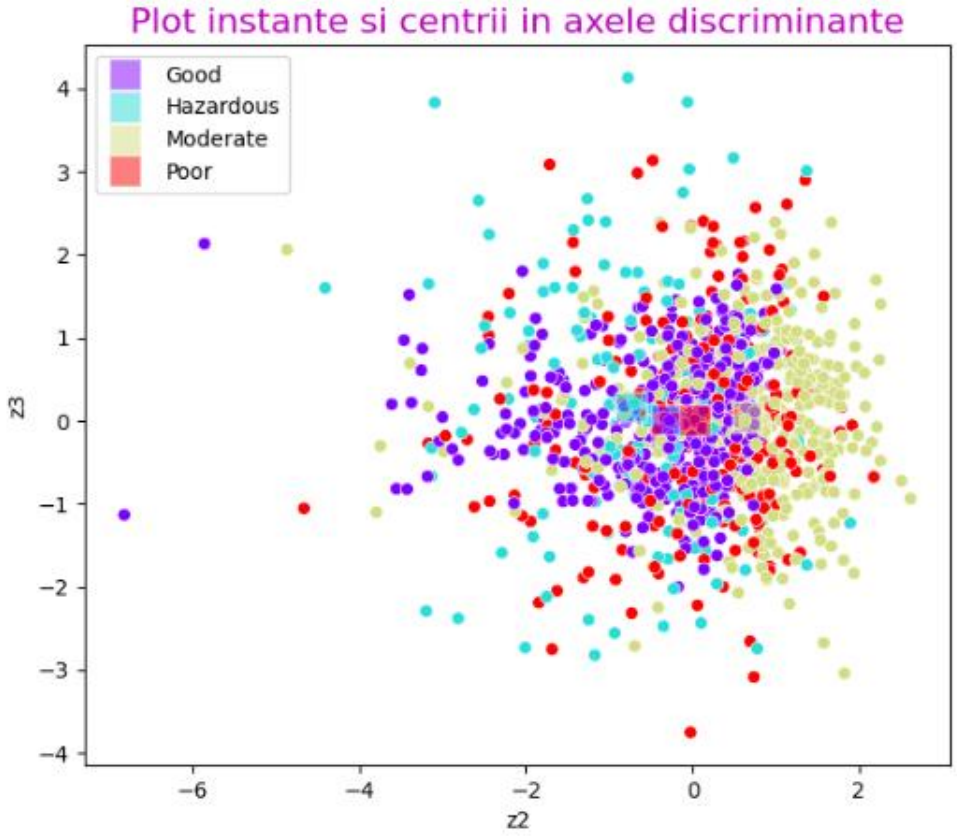
Mai mult, pentru a putea vizualiza instanțele pe grupe și centrii în axele de discriminare am realizat următoarele scatterploturi.



Acest scatterplot reprezintă proiecția fiecărei instanțe din setul de testare pe spațiul determinat de axele z1 și z2. Rprezintă cea mai bună combinație de axe, fapt ce reiese și din puterea de discriminare a acestora (Z1=12924.87772, Z2=234.89334).

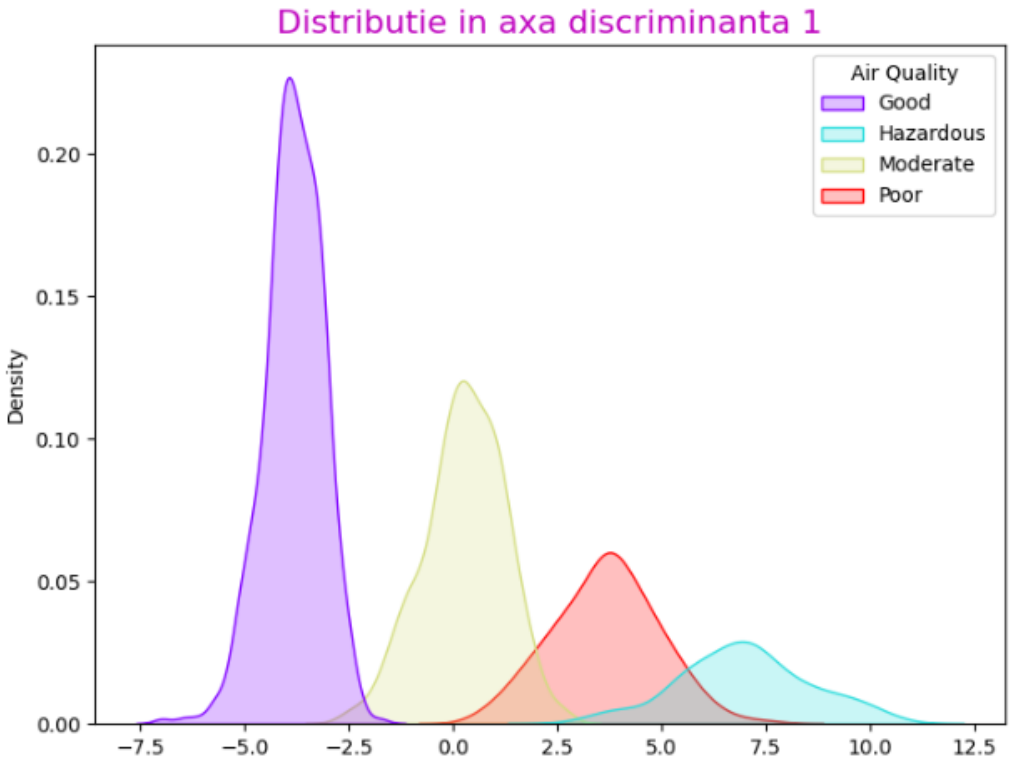


Scatterplotul de mai sus evidențiază proiecția instanțelor din setul de testare pe axele de discriminare z1 și z3. Întrucât puterea de discriminare a lui z3 este mai mică decât cea a lui z2, se poate observa o suprapunere mai mare între grupe, cu un grad de disipare mai mare.

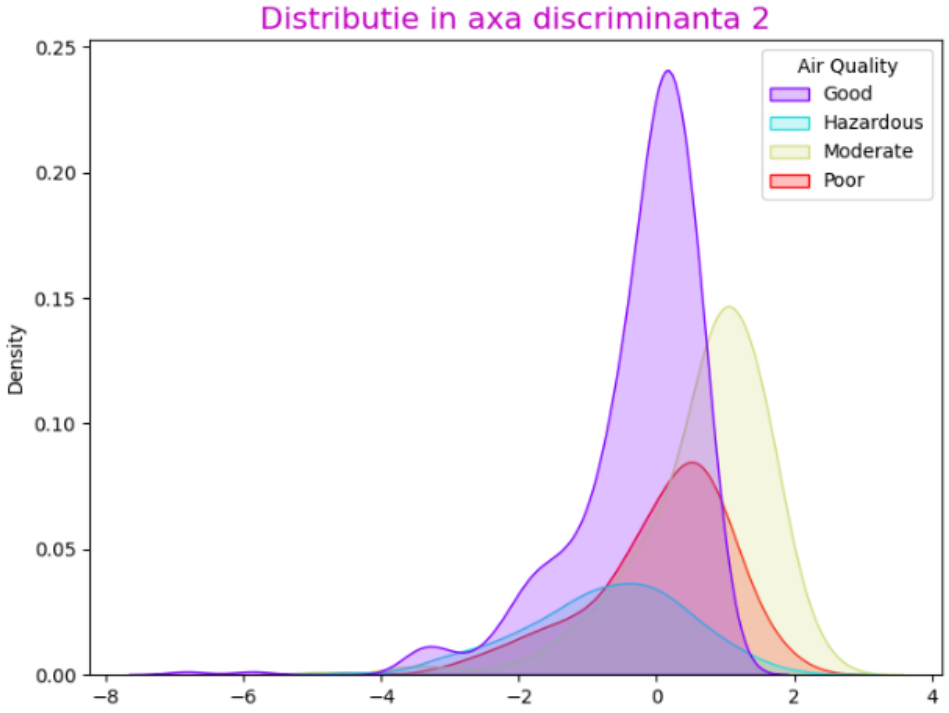


Ultimul scatterplot realizat, reprezintă proiecța instanțelor din setul de testare pe axele de discriminare z3 și z3. Întrucât aceste axe au puterea de discriminare cea mai mică se poate observa faptul că există o suprapunere foarte mare între cele 4 clase.

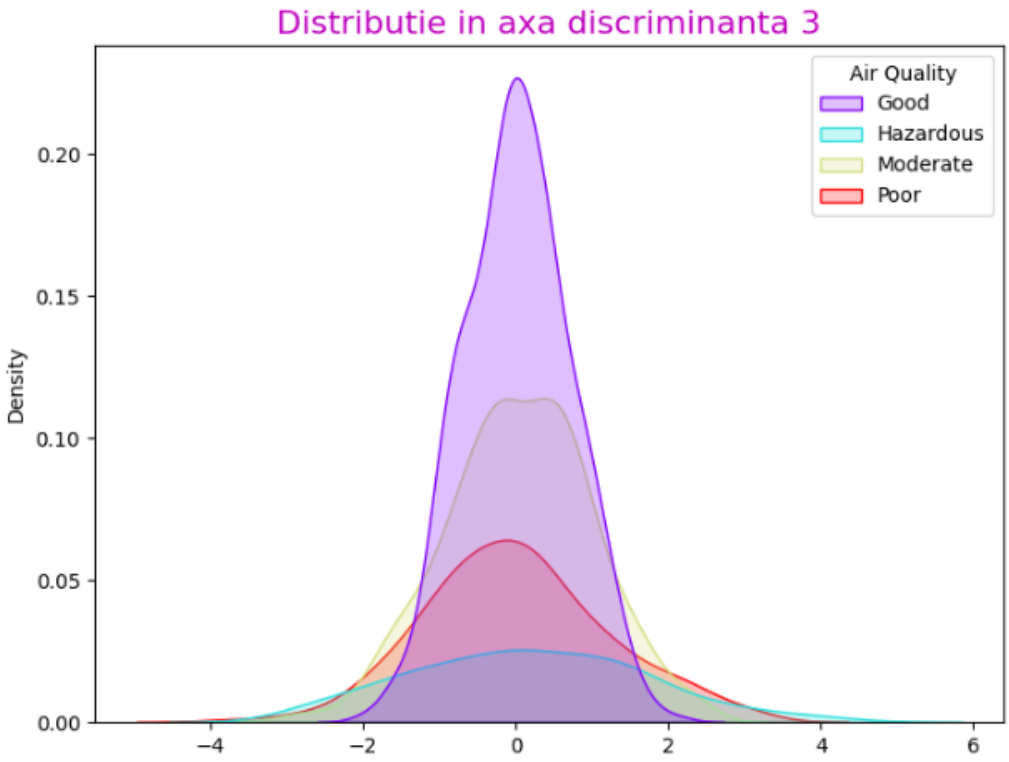
Mai departe, pentru a evidenția modul în care este distribuită fiecare instanță pe fiecare axă de discriminare am realizat următoarele ploturi:



Astfel, din primul plot reiese distribuția instanțelor pe prima axă de discriminare. Grație faptului că prezintă cea mai mare puteere de discriminare, se poate observa faptul că instanțele se împart cel mai clar în cele 4 clase.



În plotul de mai sus se poate observa faptul că instanțele nu sunt foarte clar distribuite pe a două axă de discriminare. Astfel, există o suprapunere între cele 4 grupe, cu excepția grupei Moderate care se separă.



În acest plot, se poate observa faptul că axa de determinare Z3 are o putere de discriminare foarte mică și un p-value diferit de zero, față de celelalte două axe de discriminare. Astfel în plotul de distribuție se poate observa o suprapunere totală a celor 4 clase.

În final, am plicat modelul antrenat pe setul de date de testare și am calculat matricea de confuzie, acuratețea global, acuratețea medie și indexul Cohen-Kappa.

Rezultatele din matricea de confuzie sunt următoarele:

```
,Good,Hazardous,Moderate,Poor,Acuratete
Good,481,0,0,0,100.0
Hazardous,0,104,0,27,79.389
Moderate,12,0,339,8,94.429
Poor,0,11,28,190,82.969
```

MatC_LDA.csv

Astfel, din matricea de mai sus rezultă faptul că 100% din instanțe din clasa Good au fost clasificate corect, în timp ce pentru clasa Hazardous au fost clasificate correct aproximativ 79%, pentru clasa Modarate 94%, iar pentru Poor 83%. Clasele Hazardous, Moderate și poor au anumite suprapuneri, ceea ce poate duce la diverse erori.

```
Acuratete globala,Acuratete medie,Index Cohen-Kappa
92.833,89.19675,0.8972647741616925
```

Acuratete_LDA.csv

Astfel din rezultatele obținute în urma clasificării putem afirma faptul că:

- 92.833% din instanțe din setul de testare au fost clasificate corect;
- Acuratețea medie de 89.197% sugerează că unele clase sunt mai greu de separat, ceea ce este confirmat de analiza matricei de confuzie;
- Valoarea 0.897 a indixului Cohen-Kappa indică un acord foarte bun între predicțiile modelului și valorile reale.

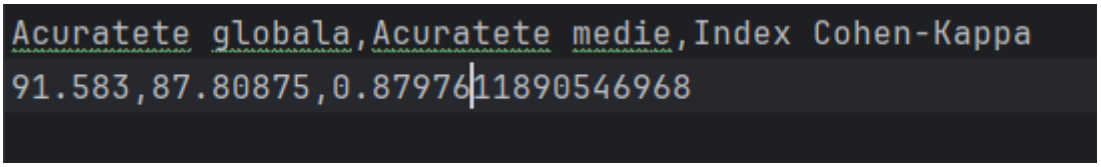
Discriminarea Bayesiană

Pentru discriminarea Bayesiană, am creat un nou model bazat pe algoritmul Bayesian Naiv, cu distribuție Gaussiană, pe care l-am antrenat folosind setul de antrenamemt. Pentru început, se presupune că fiecare predictor urmează o distribuție normală, apoi sunt calculate probabilități pentru fiecare clasă, iar instanțele sunt clasificate pe baza clasei cu probabilitatea maximă.

În urma clasificării, se calculează matricea de confuzie și urmatorii indicatori: acuratețea globală, acuratețea medie și indexul Cohen-Kappa.

```
,Good,Hazardous,Moderate,Poor,Acuratete
Good,473,0,8,0,98.337
Hazardous,0,100,0,31,76.336
Moderate,3,0,336,20,93.593
Poor,0,12,27,190,82.969
```

MatC_B



Acuratete_B.csv

Astfel, se poate observa faptul că acuratețea obținută folosind discriminarea Bayesiană este mai mică, deci vom folosi modelul LDA pentru a clasifica setul de aplicare. Predicțiile vor fi salvate în fișierul Predictii.csv.

Concluzie

Proiectul reprezintă un exemplu bine realizat al aplicării tehnicilor statistice avansate, cum ar fi Analiza Liniară Discriminantă (LDA) și Discriminarea Bayesiană, în scopul clasificării calității aerului. Prin utilizarea unor seturi de date reale, proiectul adresează o problemă esențială pentru sănătatea publică și protecția mediului: înțelegerea factorilor care determină calitatea aerului și utilizarea acestor factori pentru predicții viitoare.