ML
Tema 2 — Inteligență Artificială

Alexandru Sima (332CA)

26 mai 2025

Rezumat

Analiza a 2 seturi mari de date — indici de calitate ai aerului și informații despre popularitatea unor știri — folosind tehnici de învățare automată, prelucrarea acestora în vederea antrenării unor modele de clasificare. Clasificarea acestora folosind **Arbori de decizie**, **Păduri aleatoare**, **Regresii logistică** și **Rețele neurale adânci**. Comparații ale performanțelor modelelor.

Cuprins

1 Poluarea aerului

Primul set de date conține date despre diferiți parametri măsurați ai aerului, în peste 20.000 de orașe din întreaga lume. Prin antrenarea unui model de învățare automată, se dorește clasificarea orașelor în funcție de gradul de riscuri pentru sănătate.

1.1 Analiza datelor

1.1.1 Analiza valorică

Setul de date conține 23.463 de înregistrări, fiecare având 15 atribute, dintre care 7 numerice și 8 categorice (incluzând și atribul țintă *AQI_Category*).

Din statisticile obținute pentru atributele numerice la (1), observăm că numai *Ozone_Value* conține valori lipsă și că plajele de valori sunt destul de variate (spre exemplu, valorile *AQI_Value* se situează în principal în plaja de 40-80, pe când *CO_Value* are valori foarte apropiate de 0). Aceste fapte se observă și trasând *boxplot*-ul valorilor (2), de unde se poate observa mai clar "tendințele" outlierilori: deși în anumite cazuri (*CO_Value*, *NO2_Value*, *SO2* într-o anumită măsură), valorile mari sunt în mod flagrant eronate, în celelalte cazuri valorile sunt distribuite relativ unifom cu mult în afara plajei inter-cuartilă[1], ceea ce ar determina eliminarea a mult prea multe valori considerate outlier. De aceea, se vor considera outlieri doar valorile care depășesc $1,5 \cdot \text{IQR}$, cu "cuartilele"[2] mult mai depărtate: $0,1$, respectiv $0,9$.

Analizând atributele categorice la (3), se observă că există valori lipsă pentru *Ozone_Category* și *City*, deși atributul din urmă poate fi complet eliminat, fiecare coloană reprezentând câte un oraș diferit, acesta neputând fi folosit pentru niciun fel de corelație. Din histogramele atributelor realizate la (4 și 5), se observă că valorile atributelor nu sunt distribuite uniform, inclusiv în cazul **AQI_Category** (atribut țintă), ceea ce face clasificarea mai dificilă.

1.1.2 Analiza corelației atributelor

Aplicând testul Pearson pentru a determina corelația liniară dintre atributele numerice, se obține matricea din (6). Se poate presupune astfel că atributele *AQI_Value*, *PM25_Value* și *VOCs* sunt foarte puternic corelate între ele, având coeficientul de corelație ≥ 0.98 și că atributul *NO2_Value* nu este corelat cu niciun altul, ceea ce ar putea indica lipsa de relevanță a acestui atribut în determinarea calității aerului. Într-adevăr, primele 3 atribute sunt puternic corelate, acest fapt observându-se trasând graficele valorilor (8). Testul Pearson oferă însă doar informații despre corelația liniară, astfel că, aplicând testul Spearman, se obține matricea din (7), care arată existența unor corelații între *NO2_Value* și alți parametri, fiind deci, până la urmă, relevant.

În ceea ce privește corelația dintre atributele categorice, analiza este complicată de inegalitatea repartiției valorilor: deși testul χ^2 de la (9) indică o corelație puternică între toate atributele, mai puțin *City* cu oricare altul (evident) și perechile *CO_Category* - *NO2_Category* și *Ozone_Category* - *PM25_Category*, anumite corelații fiind doar aparente (10). Totuși, se remarcă o corelație pură: *PM25_Category* - *Emissions* (11), deci unul dintre cele 2 atribute este superfluu. Am ales să elimin atributul *PM25_Category*.

[1] IRQ — Interquartile Range; plaja de valori dintre prima (25%) și a treia (75%) cuartilă. Afișată în boxplot-uri printr-un segment.

[2] impropriu numite astfel

	AQI_Value	CO_Value	Ozone_Value	NO2_Value
count	23463.000000	23463.000000	21117.000000	23463.000000
mean	72.010868	1.368367	35.239665	43.084153
std	56.055220	1.832064	28.149280	196.079179
min	6.000000	0.000000	0.000000	0.000000
25%	39.000000	1.000000	21.000000	0.000000
50%	55.000000	1.000000	31.000000	1.000000
75%	79.000000	1.000000	40.000000	4.000000
max	500.000000	133.000000	222.000000	1003.063334

	PM25_Value	VOCs	SO2
count	23463.000000	23463.000000	23463.000000
mean	68.519755	185.053110	4.447841
std	54.796443	140.486759	5.953601
min	0.000000	12.415670	-18.528019
25%	35.000000	103.267345	0.735052
50%	54.000000	142.972272	4.286825
75%	79.000000	204.227896	7.916001
max	500.000000	1280.988229	234.692971

Figura 1: Statistici despre atributele numerice ale setului de date

1.2 Preprocesarea datelor

Analizăm cunoștințele acumulate în urma analizei de mai sus a datelor, deducem că putem elimina anumite valori, pentru a îmbunătăți performanța modelului, fără a afecta major acuratețea.

Transformarea datelor se face folosind pipeline-uri[3], pentru a fi consecventă — aceeași transformare trebuie aplicată și datelor de antrenament, și celor de test. Singura excepție este eliminarea outlierilor, care nu are sens decât pentru setul de antrenament.

1.2.1 Eliminarea valorilor extreme

Conform analizei de la (1.1.2), valorile au o dispersie foarte ridicată, deci noțiunea de outlier trebuie restricționată, pentru a nu pierde din acuratețe.

1.2.2 Imputarea valorilor lipsă

Valorile lipsă, inclusiv cele eliminate la pasul anterior, sunt imputate folosind un *SimpleImputer*[4], care completează valorile lipsă folosind mediana, respectiv moda (în cazul atributelor categorice). Imputarea multivariată (prin învățarea valorilor lipsă din celelalte atribute), implementată prin *IterativeImputer*[5], nu a dat rezultate mai bune, iar timpul de antrenare a crescut semnificativ.

[3] `https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline`
[4] `https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer`
[5] `https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer`

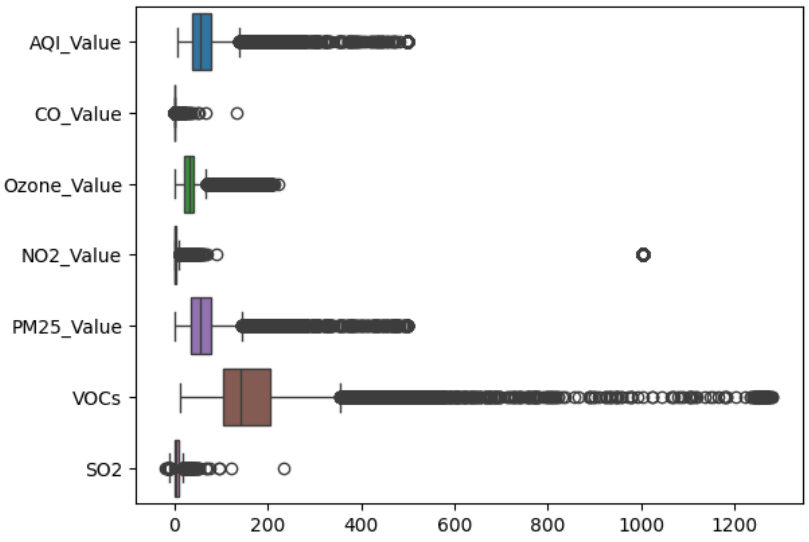

Figura 2: Boxplot pentru atributele numerice ale setului de date

1.2.3 Eliminarea atributelor redundante

Conform deciziilor de la (1.1.2), se elimină atributele *AQI_Value*, *PM25_Value* și *PM25_Category*, datorită corelațiilor și *City*, datorită irelevanței.

1.2.4 Normalizarea datelor

Atributele numerice sunt normalizate folosind *StandardScaler*[6], care le transformă astfel încât să aibă media 0 și deviația standard 1. Acest pas este necesar pentru a asigura contribuția echitabilă a fiecărui atribut în regresia modeleleor de învățare. Normalizarea este importantă mai ales în cazul regresiei logistice, deoarece valori mari (cu atât mai mult exponențiate) pot eclipsa alți parametri.

1.2.5 Codificarea atributelor categorice și a atributului țintă

Inițial, am decis ca atributele categorice să fie codificate folosind *OrdinalEncoder*[7], care le transformă în numere întregi, fiecare valoare unică având un număr corespunzător, dar, codificând în schimb prin *OneHotEncoder*[8], se obțin rezultate mai bune, probabil existând o mai mare relevanță a unor valori specifice ale unei clase.

[6] `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler`
[7] `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OrdinalEncoder`
[8] `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder`

	count	unique	top	freq
Country	23036	175	United States of America	2872
City	23462	23462	Marang	1
CO_Category	21117	2	Good	21115
Ozone_Category	23463	5	Good	21069
NO2_Category	23463	2	Good	23448
PM25_Category	23463	6	L0	10208
Emissions	23463	6	L0	10208
AQI_Category	23463	6	Good	9936

Figura 3: Statistici despre atributele categorice ale setului de date

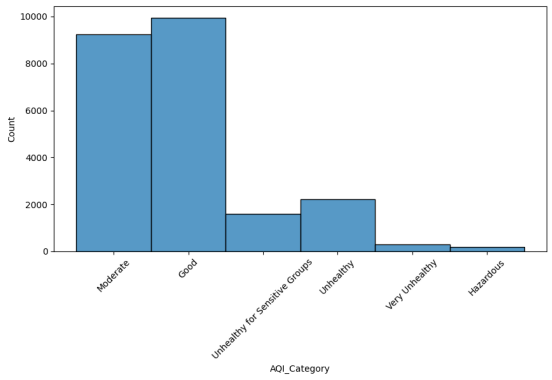

Figura 4: Histogramă pentru atributul țintă al setului de date

1.3 Învățarea automată

1.3.1 Arbori de decizie

Pentru a antrena un model de tip arbore de decizie, se folosește *DecisionTreeClassifier*[9], care îl construiește pe baza datelor de antrenament. Acesta este foarte performant, reușind, cu parametrizarea implicită, o acuratețe de $\approx 100\%$ pe setul de test, clasificând eronat, în medie, 3 intrări, conform matricei de confuzie de la (12).

1.3.2 Păduri aleatoare

Modelul de tip pădure aleatoare folosit este *RandomForestClassifier*[10], care reușește performanțe similare (conform (13)), având însă o performanță mai scăzută d.p.d.v. temporal ($\approx 1,6s$ vs $\approx 0,2s$).

[9] `https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier`
[10] `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier`

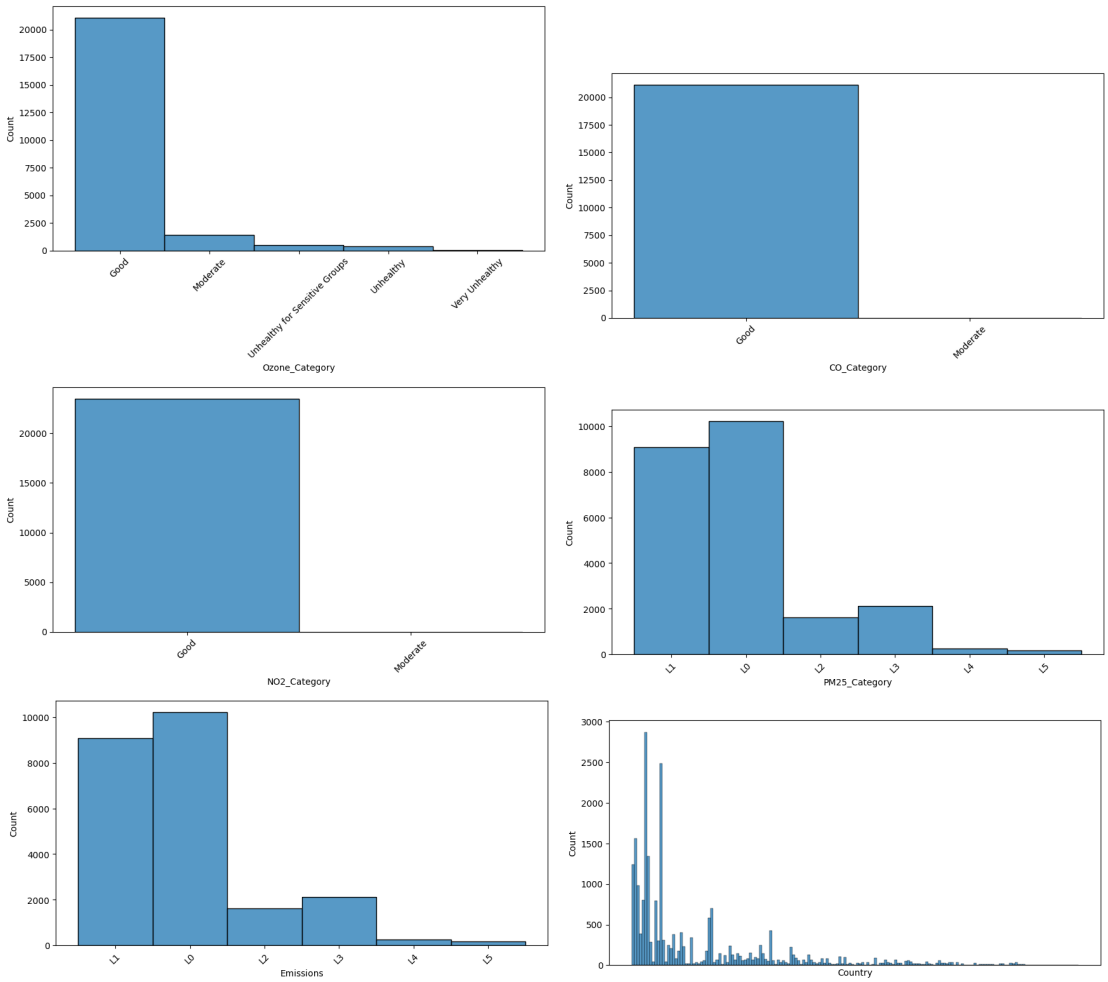

Figura 5: Histograme pentru atributele categorice ale setului de date . *City* este ignorat din motivele enunțate anterior.

1.3.3 Regresie logistică

Modelul de regresie logistică este implementat manual. Cum regresia logistică este folosită în mod obișnuit pentru clasificare binară, aceasta trebuie adaptată, realizând o clasificare de tip *one-vs-rest*, antrenându-se câte un regresor pentru fiecare clasa, iar clasa prezisă fiind cea a cărei regresor întoarce o valoare maximă. Nu am implementat regularizare, datele fiind în principiu normalizate, iar, prin testare, neobținându-se rezultate mai bune. Pe setul de test, se obține o acuratețe de $\approx 87\%$ (conform (14)).

1.3.4 Rețele neurale adânci

Modelul de rețea neurală adâncă folosit este *MLPClassifier*[11], care reușește o acuratețe de $\approx$ 97% (conform (15)) pe setul de test, cu 2 straturi ascunse, fiecare a câte 64 de neuroni, restul

[11] `https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier`

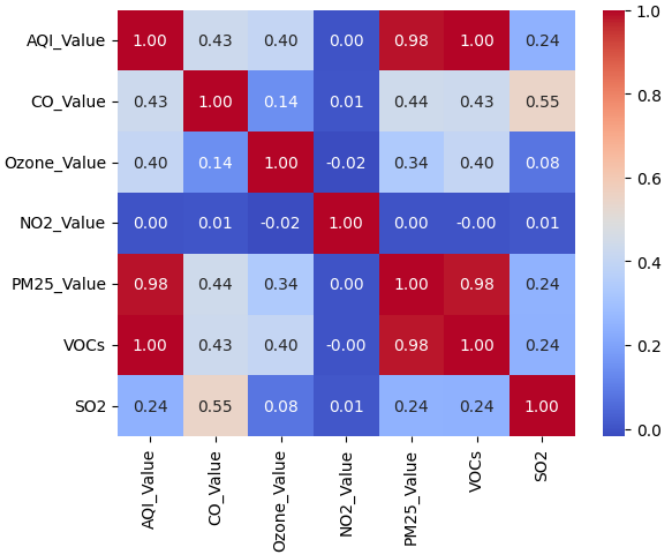

Figura 6: Corelația dintre atributele setului de date, folosind coeficientul Pearson

parametrilor fiind cei impliciți.

1.4 Comparații

Comparând performanțele modelelor, putem observa că arborii de decizie oferă cele mai bune rezultate, urmate de pădurile aleatoare, rețelele neurale și, în cele din urmă, regresia logistică. Se observă că clasele cel mai frecvent clasificate greșit sunt *Hazardous* și *Very Unhealthy*, care au și cel mai mic suport (sub 60 de valori fiecare). În particular, regresia logistică nu clasifică corect niciuna dintre aceste clase.

Class	**Precision**	**Recall**	**F1-score**	**Support**
Good	**1.00**	**1.00**	**1.00**	**1987**
Hazardous	**1.00**	**0.97**	**0.99**	38
Moderate	**1.00**	**1.00**	**1.00**	1846
Unhealthy	**1.00**	**1.00**	**1.00**	446
Unhealthy for Sensitive Groups	**1.00**	**1.00**	**1.00**	318
Very Unhealthy	**0.98**	**1.00**	**0.99**	58
Accuracy			**1.00**	**4693**
Macro avg	**1.00**	**1.00**	**1.00**	**4693**
Weighted avg	**1.00**	**1.00**	**1.00**	**4693**

Tabela 1: Raport de clasificare pentru arbori de decizie

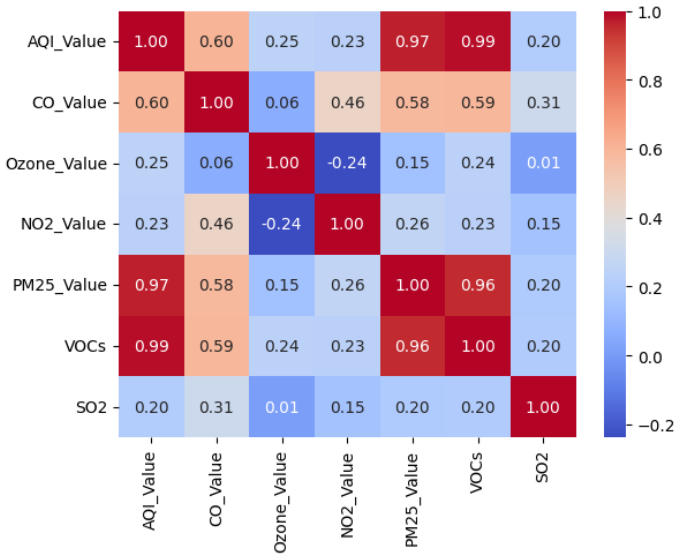

Figura 7: Corelația dintre atributele setului de date, folosind coeficientul Spearman

Class	**Precision**	**Recall**	**F1-score**	**Support**
Good	**1.00**	**1.00**	**1.00**	**1987**
Hazardous	**1.00**	0.84	0.91	38
Moderate	**1.00**	**1.00**	**1.00**	1846
Unhealthy	**1.00**	**1.00**	**1.00**	446
Unhealthy for Sensitive Groups	**1.00**	**1.00**	**1.00**	318
Very Unhealthy	0.90	0.98	0.94	58
Accuracy			**1.00**	**4693**
Macro avg	0.98	0.97	0.98	**4693**
Weighted avg	**1.00**	**1.00**	**1.00**	**4693**

Tabela 2: Raport de clasificare pentru păduri aleatoare

Class	**Precision**	**Recall**	**F1-score**	**Support**
Good	0.96	0.96	0.96	**1987**
Hazardous	0.00	0.00	0.00	38
Moderate	0.85	0.97	0.91	1846
Unhealthy	0.76	0.74	0.75	446
Unhealthy for Sensitive Groups	0.94	0.44	0.60	318
Very Unhealthy	0.00	0.00	0.00	58
Accuracy			0.89	**4693**
Macro avg	0.59	0.52	0.54	**4693**
Weighted avg	0.88	0.89	0.87	**4693**

Tabela 3: Raport de clasificare pentru regresia logistică

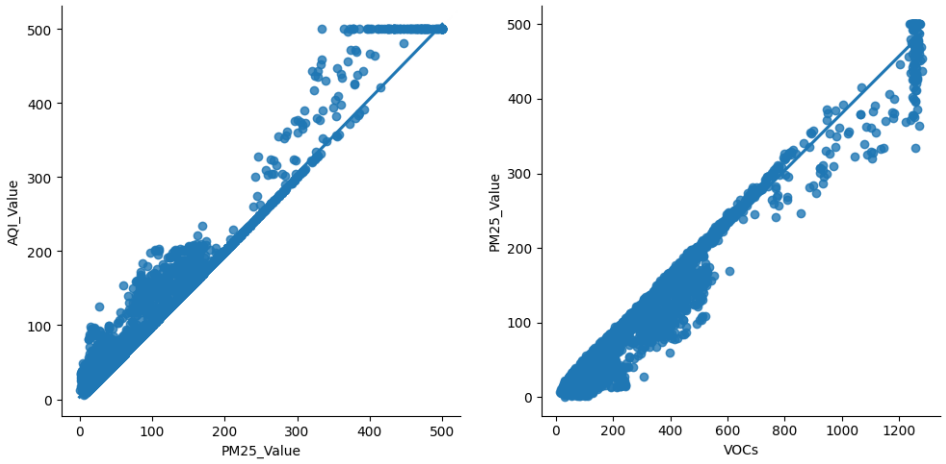

Figura 8: Corelația liniară dintre *AQI_Value* și *PM25_Value*, respectiv *PM25_Value* și *VOCs*

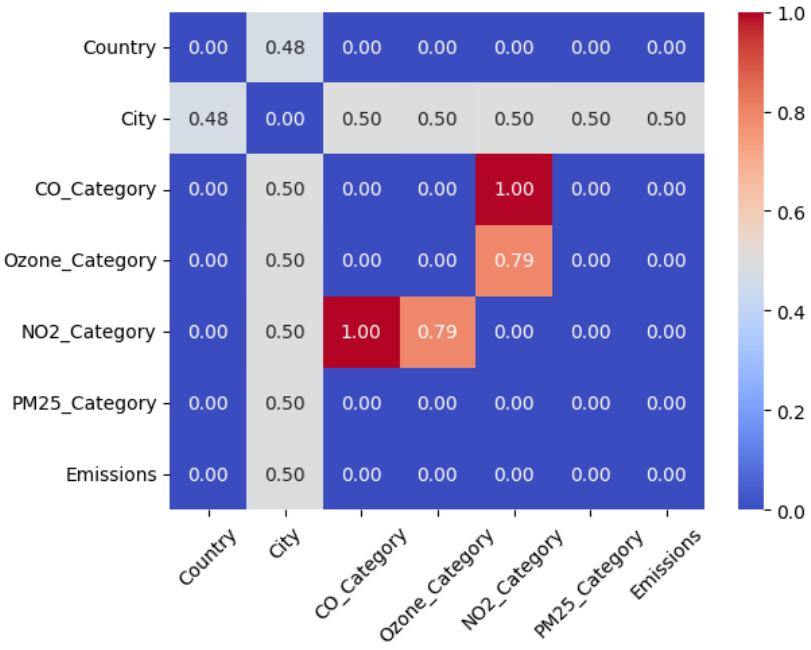

Figura 9: Testul χ^2 pentru atributele categorice ale setului de date

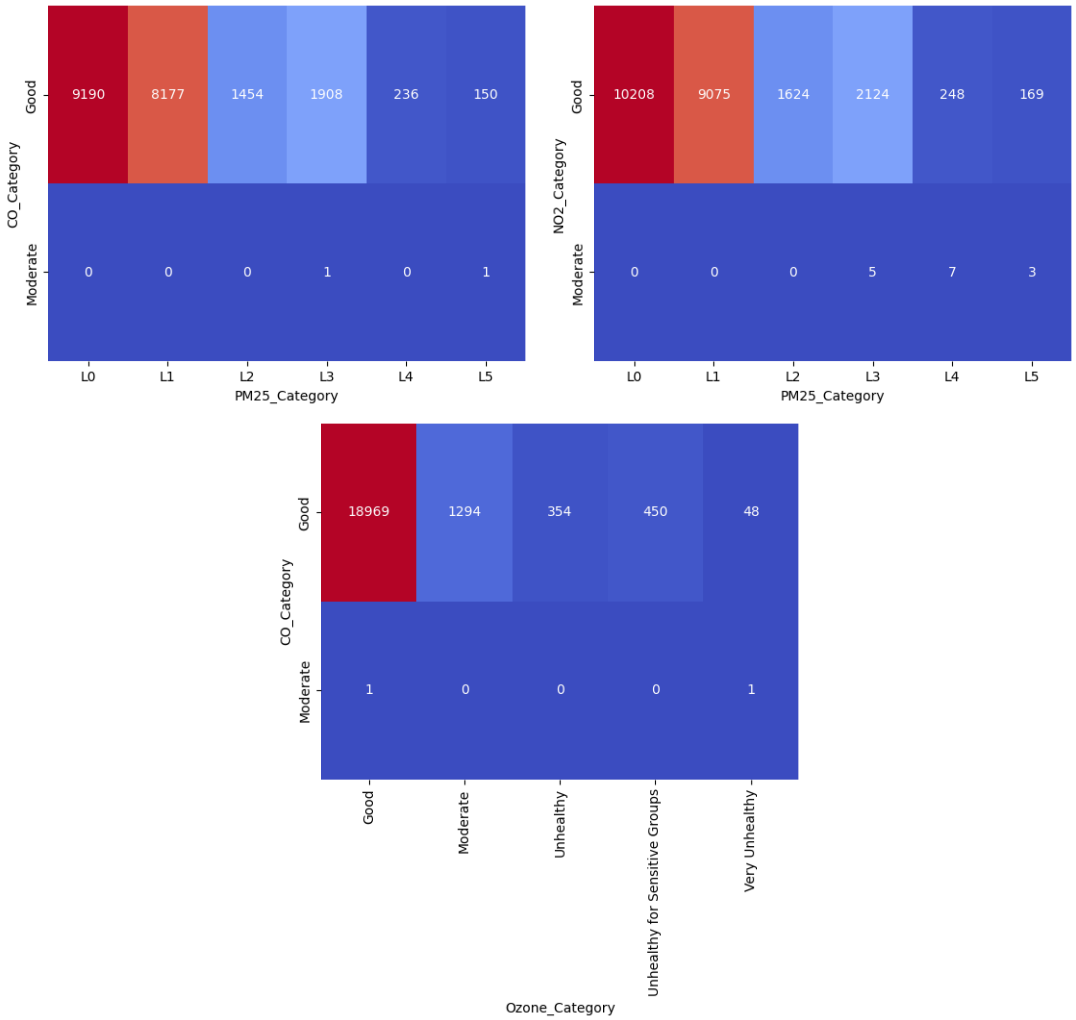

Figura 10: Câteva corelații puternice între atributele categorice ale setului de date, datorate inegalității

Class	Precision	Recall	F1-score	Support
Good	**1.00**	0.96	0.98	**1987**
Hazardous	0.16	0.95	0.28	38
Moderate	**1.00**	0.96	0.98	1846
Unhealthy	0.99	0.96	0.98	446
Unhealthy for Sensitive Groups	0.99	0.96	0.98	318
Very Unhealthy	0.96	0.90	0.93	58
Accuracy			0.96	**4693**
Macro avg	0.85	0.95	0.85	**4693**
Weighted avg	0.99	0.96	0.97	**4693**

Tabela 4: Raport de clasificare pentru rețele neurale adânci

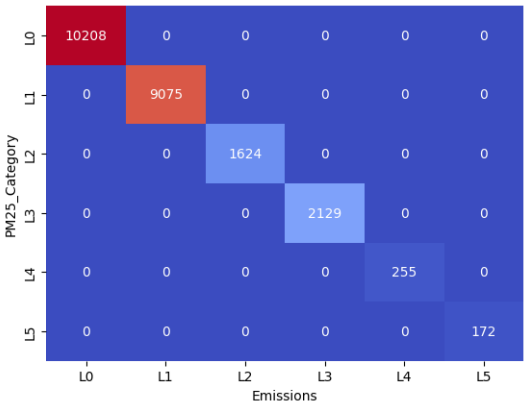

Figura 11: Corelație totală între *PM25_Category* și *Emissions*

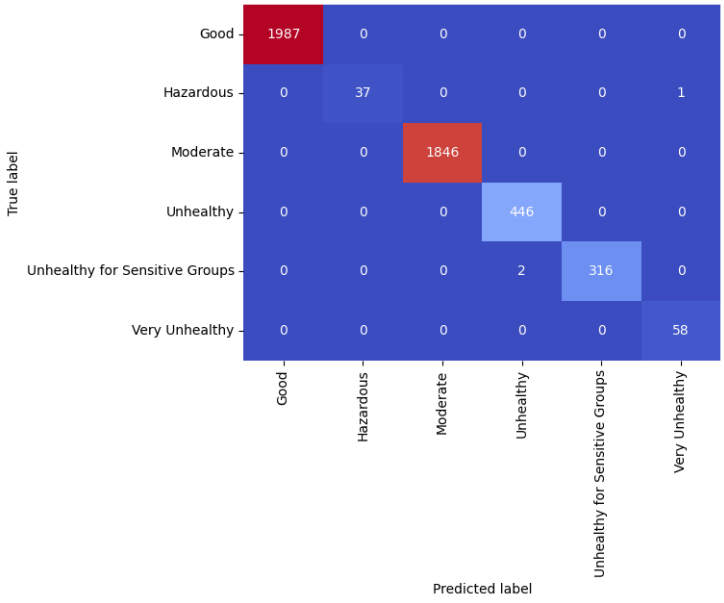

Figura 12: Matricea de confuzie a modelului de tip arbore de decizie

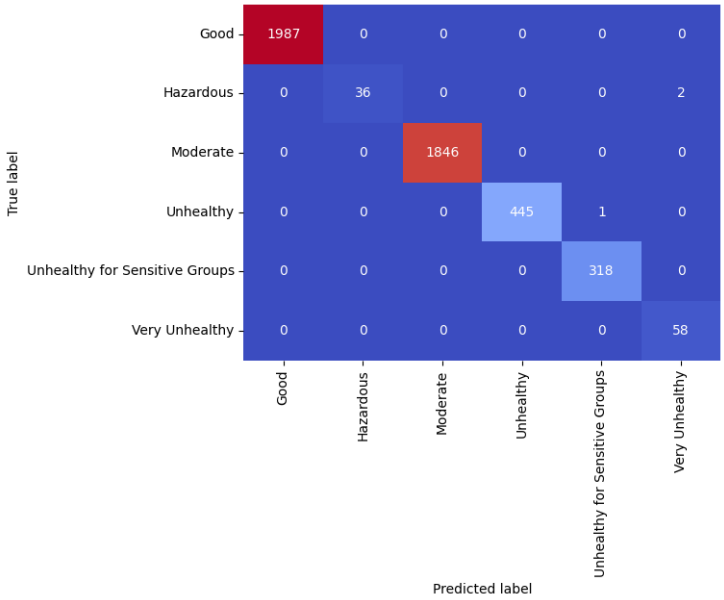

Figura 13: Matricea de confuzie a modelului de tip pădure aleatoare

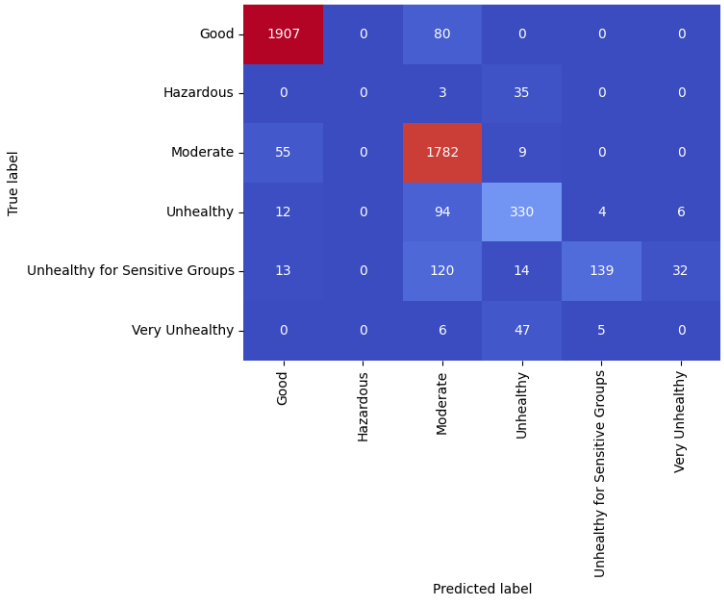

Figura 14: Matricea de confuzie a modelului de regresie logistică

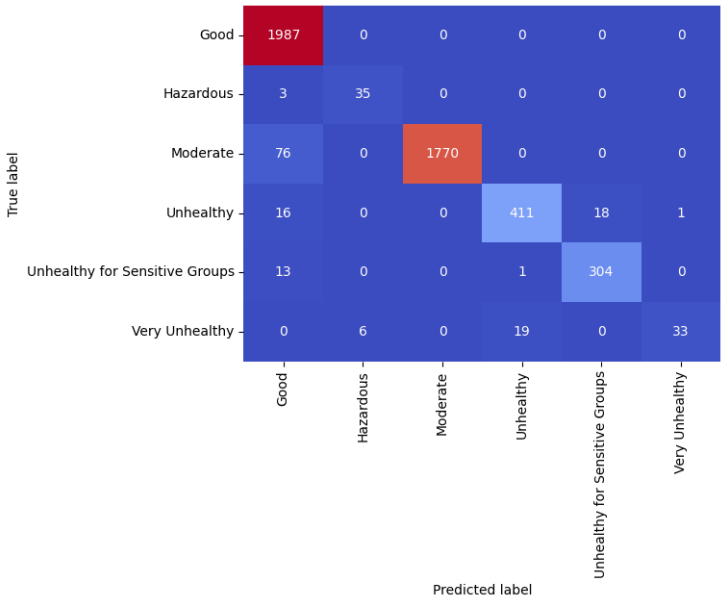

Figura 15: Matricea de confuzie a modelului de tip rețea neurală adâncă

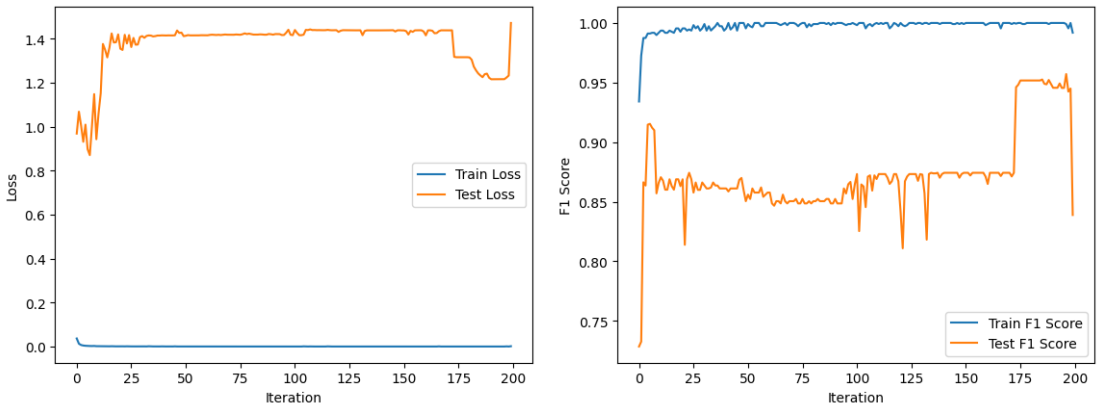

Figura 16: Curbe de învățare și de acuratețe pentru rețea

2 Popularitatea știrilor

Al doilea set de date conține informații despre o selecție de aproape 40.000 de articole știri de pe site-ul `https://www.mashable.com`, cuprinzând informații precum numărul de linkuri, de vizualizări, momentul publicării, categoria articolului și date despre cuvintele folosite. Ca și în cazul setului de date anterior, se dorește clasificarea articolelor în funcție de popularitatea acestora, aceasta cerință fiind mai dificilă, deoarece volumul de date este mult mai mare.

2.1 Analiza datelor

2.1.1 Analiza valorică

Setul de date conține 39.644 de înregistrări, fiecare având 64 de atribute. 47 dintre acestea sunt numerice, iar restul de sunt categorice (incluzând atributul țintă *popularity_category*).

Analizând repartițiile atributelor numerice de la (17), (18), (19) și (20), se observă că singura coloană cu valori lipsă este cea de *content_density* și, ca și în cazul anterior, că există o distribuție foarte largă a valorilor, pe intervale de ordine de mărime foarte diferite: de exemplu, *topic_relevance* și **_rate* sunt rapoarte (în intervalul $0-1$), pe când *keyword_best_avg_shares* este de ordinul sutelor de mii. Având acestea în vedere, se vor considera outlierii ca în cazul setului de date precedent, apoi se va aplica o procedură de normalizare a datelor.

În ceea ce privește atributele categorice, se observă din (21) că există valori lipsă pentru *channel_lifestyle* și că, în afară de *url* care are numai valori unice (precum *City* în cazul poluării), toate atributele au doar 2 valori posiblie (de obicei "Yes" / "No"). De aceea, o codificare de tip one-hot nu își are rostul, adăugând câte un atribut în plus pentru fiecare valoare posibilă. Analizând histogramele de la (22), (23) și (24), se observă, din nou, o distribuție inegală, inclusiv în cazul atributului țintă **popularity_category**, fiind totuși mai uniformă decât în cazul setului de date precedent.

2.1.2 Analiza corelației atributelor

Analizând corelația dintre atributele numerice folosind coeficientul Pearson, se obține matricea din (25). Trasând graficele de corelație dintre atributele cu un indice ridicat ($|p| \geq 0.9$), se observă în (26) că există niște corelații date în mod eronat de outlieri: *non_stop_word_ratio* - *unique_non_stop_word_ratio* și *unique_word_ratio* - *non_stop_word_ratio*. În schimb, corelații liniare evidente relevă din (27) între perechile *keyword_worst_max_shares* - *keyword_worst_avg_shares* și *content_word_count* - *content_density*. Am ales să elimin *keyword_worst_avg_shares* și *content_word_count*.

În ceea ce privește corelația între atributele categorice (calculată la (28)), aceasta este logică din moment ce majoritatea provin dintr-o codificare one-hot a unor atribute (de exemplu, *day_monday* adevărat determină implicit ca toate celelalte zile ale săptămânii să fie false). Nu are sens să eliminăm aceste atribute. Totuși, există o redundanță evidentă: *day_** vs *is_weekend*. În funcție de scop, se poate alege păstrarea anumitor atribute. În cazul acesta, am ales să elimin *is_weekend*, păstrând mai multe informații (ce zi este, nu doar dacă este weekend sau nu) în speranța că se va atinge o acuratețe mai mare.

2.2 Preprocesarea datelor

Modul de preprocesare a datelor este similar cu cel descris la (1.2). Singurele diferențe sunt selectarea atributelor relevante și codificarea atributelor categorice.

2.2.1 Eliminarea atributelor redundante

Conform rezultatelor de mai sus, se elimină atributele *url*, *is_weekend*, *keyword_worst_avg_shares* și *content_word_count*.

2.2.2 Codificarea atributelor categorice și a atributului țintă

Atributele categorice sunt codificate folosind *OrdinalEncoder*[12], care le transformă în numere întregi, fiecare valoare unică având un număr corespunzător. Deoarece toate atributele categorice au doar 2 valori posibile, acest lucru nu afectează performanța modelului, iar codificarea este mai eficientă decât codificarea one-hot. Atributul țintă este, totuși, codificat folosind *LabelEncoder*.

2.3 Învățarea automată

2.3.1 Arbori de decizie

Modelul de tip arbore de decizie folosit este parametrizat prin:

```
min_samples_split=5
min_samples_leaf=5
criterion="entropy"
```

Acesta reușește o acuratețe de $\approx 89\%$ pe setul de test.

2.3.2 Păduri aleatoare

Modelul de tip pădure aleatoare folosit este parametrizat prin:

```
n_estimators=500
min_samples_split=5
min_samples_leaf=5
criterion="entropy"
max_features=1.0
```

Acesta reușește o acuratețe de $\approx 89\%$ pe setul de test.

2.3.3 Regresie logistică

Modelul de regresie logistică este cel implementat la (1.3.3). Acesta reușește o acuratețe de $\approx 88\%$ pe setul de test.

2.3.4 Rețele neurale adânci

Modelul de rețea neurală adâncă este parametrizat prin:

```
iters=200
hidden_layer_sizes=[100, 100]
activation="relu"
solver="adam"
learning_rate_init=0.001
```

2.4 Comparații

[12] `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OrdinalEncoder`

Class	Precision	Recall	F1-score	Support
Moderately Popular	0.93	**0.92**	**0.93**	2401
Popular	**0.80**	0.83	0.81	1074
Slightly Popular	**0.95**	**0.94**	**0.95**	**3799**
Unpopular	0.42	**0.93**	0.58	218
Viral	0.93	0.42	**0.58**	437
Accuracy			**0.89**	**7929**
Macro avg	0.81	**0.81**	**0.77**	**7929**
Weighted avg	**0.91**	**0.89**	**0.89**	**7929**

Tabela 5: Raport de performanță pentru modelul de tip arbore de decizie

Class	Precision	Recall	F1-score	Support
Moderately Popular	**0.94**	**0.92**	**0.93**	2401
Popular	**0.80**	**0.84**	**0.82**	1074
Slightly Popular	**0.95**	**0.94**	**0.95**	**3799**
Unpopular	**0.43**	**0.93**	**0.59**	218
Viral	**0.96**	0.41	0.57	437
Accuracy			**0.89**	**7929**
Macro avg	**0.82**	**0.81**	**0.77**	**7929**
Weighted avg	**0.91**	**0.89**	**0.89**	**7929**

Tabela 6: Raport de performanță pentru modelul de tip pădure aleatoare

Class	Precision	Recall	F1-score	Support
Moderately Popular	0.59	0.37	0.46	2401
Popular	0.38	0.47	0.42	1074
Slightly Popular	0.73	0.72	0.72	**3799**
Unpopular	0.05	0.19	0.08	218
Viral	0.05	0.05	0.05	437
Accuracy			0.53	**7929**
Macro avg	0.36	0.36	0.35	**7929**
Weighted avg	0.58	0.53	0.55	**7929**

Tabela 7: Raport de performanță pentru modelul de regresie logistică

Class	Precision	Recall	F1-score	Support
Moderately Popular	0.70	0.70	0.70	2401
Popular	0.71	0.58	0.64	1074
Slightly Popular	0.80	0.84	0.82	**3799**
Unpopular	0.67	0.34	0.45	218
Viral	0.55	0.69	0.61	437
Accuracy			0.74	**7929**
Macro avg	0.69	0.63	0.65	**7929**
Weighted avg	0.74	0.74	0.74	**7929**

Tabela 8: Raport de performanță pentru modelul de rețeaua neurală

	days_since_published	title_word_count	content_word_count	unique_word_ratio
count	39644.000000	39644.000000	39644.000000	39644.000000
mean	354.530471	10.398749	546.514731	0.548216
std	214.163767	2.114037	471.107508	3.520708
min	8.000000	2.000000	0.000000	0.000000
25%	164.000000	9.000000	246.000000	0.470870
50%	339.000000	10.000000	409.000000	0.539226
75%	542.000000	12.000000	716.000000	0.608696
max	731.000000	23.000000	8474.000000	701.000000

	non_stop_word_ratio	unique_non_stop_ratio	external_links	internal_links
count	39644.000000	39644.000000	39644.000000	39644.000000
mean	0.996469	0.689175	192.250110	3.293638
std	5.231231	3.264816	905.415876	3.855141
min	0.000000	0.000000	0.000000	0.000000
25%	1.000000	0.625739	4.000000	1.000000
50%	1.000000	0.690476	8.000000	3.000000
75%	1.000000	0.754630	15.000000	4.000000
max	1042.000000	650.000000	6078.616775	116.000000

	image_count	video_count	avg_word_length	keyword_count
count	39644.000000	39644.000000	39644.000000	39644.000000
mean	4.544143	1.249874	4.548239	7.223767
std	8.309434	4.107855	0.844406	1.909130
min	0.000000	0.000000	0.000000	1.000000
25%	1.000000	0.000000	4.478404	6.000000
50%	1.000000	0.000000	4.664082	7.000000
75%	4.000000	1.000000	4.854839	9.000000
max	128.000000	91.000000	8.041534	10.000000

	keyword_worst_min_shares	keyword_worst_max_shares	keyword_worst_avg_shares	keyword_best_min_shares
count	39644.000000	39644.000000	39644.000000	39644.000000
mean	26.106801	1153.951682	312.366967	13612.354102
std	69.633215	3857.990877	620.783887	57986.029357
min	-1.000000	0.000000	-1.000000	0.000000
25%	-1.000000	445.000000	141.750000	0.000000
50%	-1.000000	660.000000	235.500000	1400.000000
75%	4.000000	1000.000000	357.000000	7900.000000
max	377.000000	298400.000000	42827.857143	843300.000000

Figura 17: Statistici despre atributele numerice ale setului de date (1)

	keyword_best_max_shares	keyword_best_avg_shares	keyword_avg_min_shares	keyword_avg_max_shares
count	39644.000000	39644.000000	39644.000000	39644.000000
mean	752324.066694	259281.938083	1117.146610	5657.211151
std	214502.129573	135102.247285	1137.456951	6098.871957
min	0.000000	0.000000	-1.000000	0.000000
25%	843300.000000	172846.875000	0.000000	3562.101631
50%	843300.000000	244572.222223	1023.635611	4355.688836
75%	843300.000000	330980.000000	2056.781032	6019.953968
max	843300.000000	843300.000000	3613.039819	298400.000000

	keyword_avg_avg_shares	ref_min_shares	ref_max_shares	ref_avg_shares
count	39644.000000	39644.000000	39644.000000	39644.000000
mean	3135.858639	3998.755396	10329.212662	6401.697580
std	1318.150397	19738.670516	41027.576613	24211.332231
min	0.000000	0.000000	0.000000	0.000000
25%	2382.448566	639.000000	1100.000000	981.187500
50%	2870.074878	1200.000000	2800.000000	2200.000000
75%	3600.229564	2600.000000	8000.000000	5200.000000
max	43567.659946	843300.000000	843300.000000	843300.000000

	topic_0_relevance	topic_1_relevance	topic_2_relevance	topic_3_relevance
count	39644.000000	39644.000000	39644.000000	39644.000000
mean	0.184599	0.141256	0.216321	0.223770
std	0.262975	0.219707	0.282145	0.295191
min	0.000000	0.000000	0.000000	0.000000
25%	0.025051	0.025012	0.028571	0.028571
50%	0.033387	0.033345	0.040004	0.040001
75%	0.240958	0.150831	0.334218	0.375763
max	0.926994	0.925947	0.919999	0.926534

	topic_4_relevance	content_subjectivity	content_sentiment	positive_word_rate
count	39644.000000	39644.000000	39644.000000	39644.000000
mean	0.234029	0.443370	0.119309	0.039625
std	0.289183	0.116685	0.096931	0.017429
min	0.000000	0.000000	-0.393750	0.000000
25%	0.028574	0.396167	0.057757	0.028384
50%	0.040727	0.453457	0.119117	0.039023
75%	0.399986	0.508333	0.177832	0.050279
max	0.927191	1.000000	0.727841	0.155488

Figura 18: Statistici despre atributele numerice ale setului de date (2)

	negative_word_rate	non_neutral_positive_rate	non_neutral_negative_rate	avg_positive_sentiment
count	39644.000000	39644.000000	39644.000000	39644.000000
mean	0.016612	0.682150	0.287934	0.353825
std	0.010828	0.190206	0.156156	0.104542
min	0.000000	0.000000	0.000000	0.000000
25%	0.009615	0.600000	0.185185	0.306244
50%	0.015337	0.710526	0.280000	0.358755
75%	0.021739	0.800000	0.384615	0.411428
max	0.184932	1.000000	1.000000	1.000000

	min_positive_sentiment	max_positive_sentiment	avg_negative_sentiment	min_negative_sentiment
count	39644.000000	39644.000000	39644.000000	39644.000000
mean	0.095446	0.756728	-0.259524	-0.521944
std	0.071315	0.247786	0.127726	0.290290
min	0.000000	0.000000	-1.000000	-1.000000
25%	0.050000	0.600000	-0.328383	-0.700000
50%	0.100000	0.800000	-0.253333	-0.500000
75%	0.100000	1.000000	-0.186905	-0.300000
max	1.000000	1.000000	0.000000	0.000000

	max_negative_sentiment	title_subjectivity	title_sentiment	title_subjectivity_magnitude
count	39644.000000	39644.000000	39644.000000	39644.000000
mean	-0.107500	0.282353	0.071425	0.341843
std	0.095373	0.324247	0.265450	0.188791
min	-1.000000	0.000000	-1.000000	0.000000
25%	-0.125000	0.000000	0.000000	0.166667
50%	-0.100000	0.150000	0.000000	0.500000
75%	-0.050000	0.500000	0.150000	0.500000
max	0.000000	1.000000	1.000000	0.500000

	title_sentiment_magnitude	engagement_ratio	content_density
count	39644.000000	39644.000000	35680.000000
mean	0.156064	1054.066316	1986.559830
std	0.226294	3496.605663	2209.101848
min	0.000000	0.041667	32.759785
25%	0.000000	220.000000	746.069031
50%	0.000000	465.500000	1314.549818
75%	0.250000	900.000000	2506.707506
max	1.000000	221200.000000	58857.969230

Figura 19: Statistici despre atributele numerice ale setului de date (3)

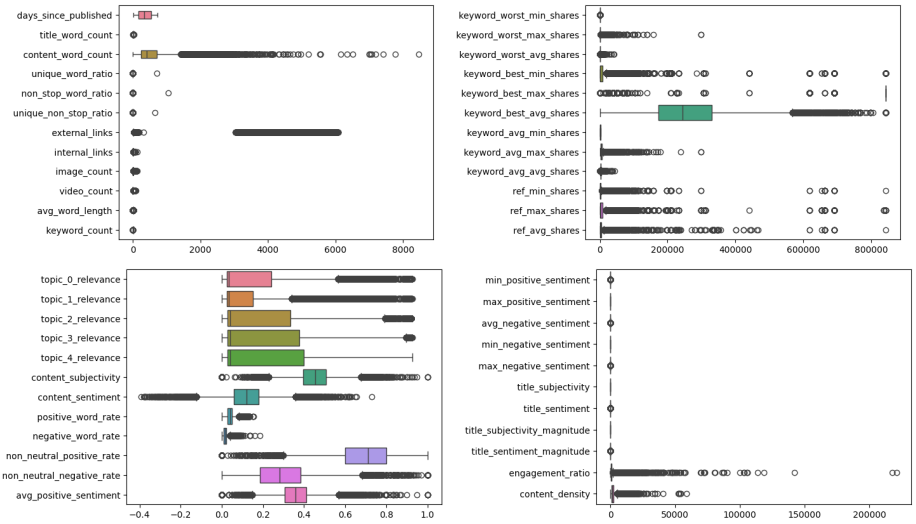

Figura 20: Boxplot pentru atributele numerice ale setului de date

	count	unique	top	freq
url	39644	39644	http://mashable.com/2014/12/27/youtube-channel...	1
channel_lifestyle	35680	2	N	33777
hannel_entertainment	39644	2	N	32587
channel_business	39644	2	N	33386
channel_social_media	39644	2	N	37321
channel_tech	39644	2	N	32298
channel_world	39644	2	N	31217
day_monday	39644	2	N	32983
day_tuesday	39644	2	N	32254
day_wednesday	39644	2	N	32209
day_thursday	39644	2	N	32377
day_friday	39644	2	N	33943
day_saturday	39644	2	N	37191
day_sunday	39644	2	N	36907
is_weekend	39644	2	N	34454
publication_period	39644	2	Weekday	34454

Figura 21: Statistici despre atributele categorice ale setului de date

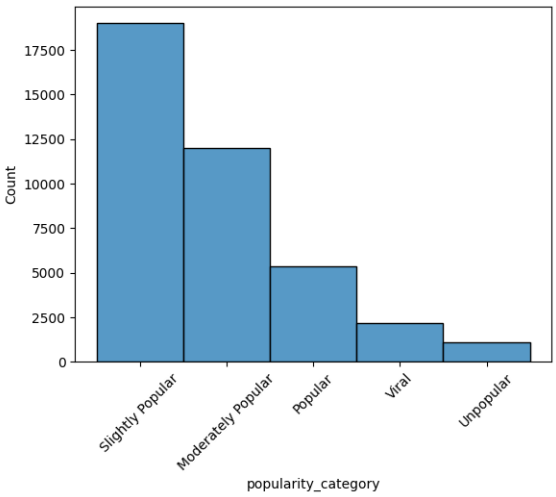

Figura 22: Statistici despre atributul țintă al setului de date

Figura 23: Histograme pentru atributele categorice ale setului de date (incluzând atributul țintă). *url* este ignorat. (1)

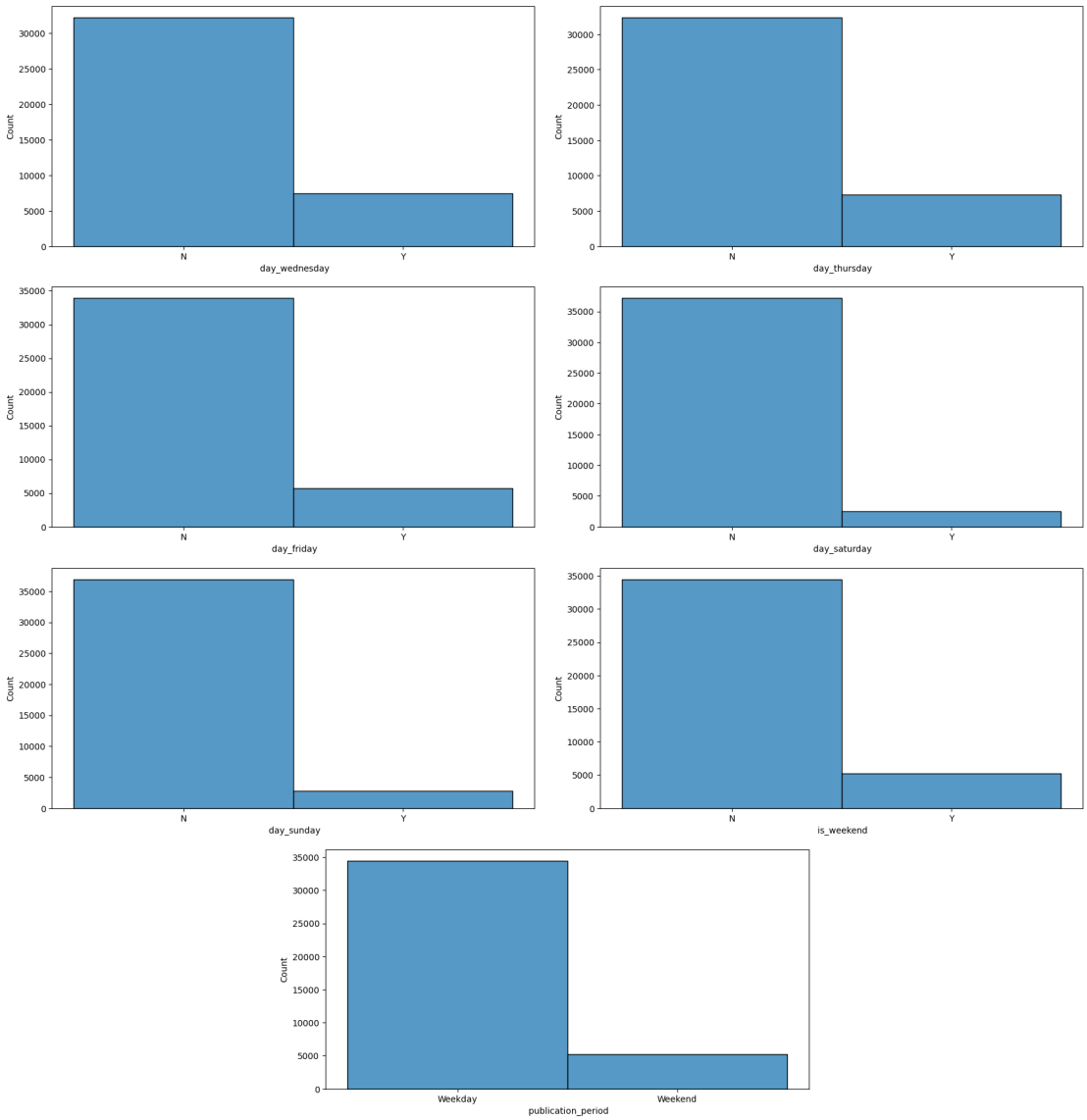

Figura 24: Histograme pentru atributele categorice ale setului de date (2)

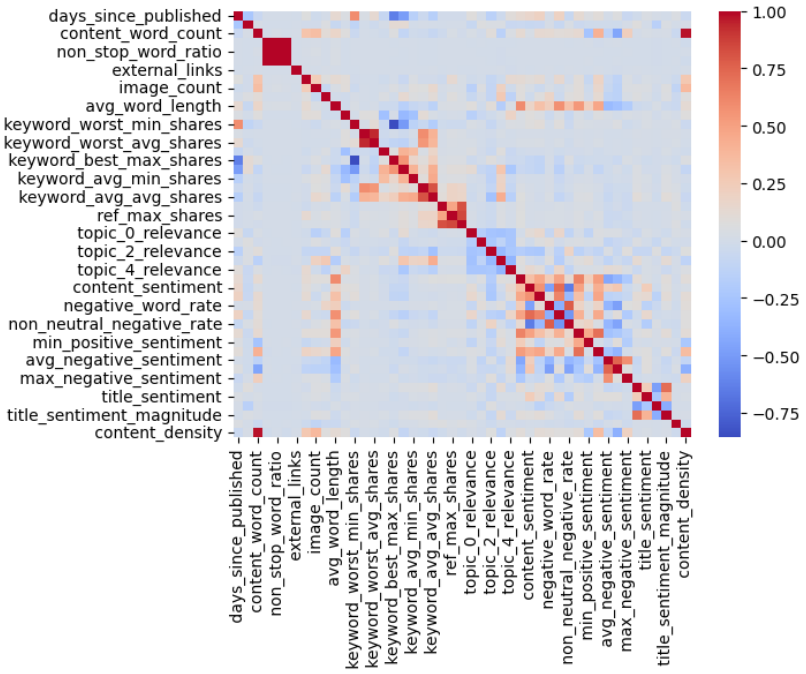

Figura 25: Corelația dintre atributele numerice ale setului de date, folosind coeficientul Pearson

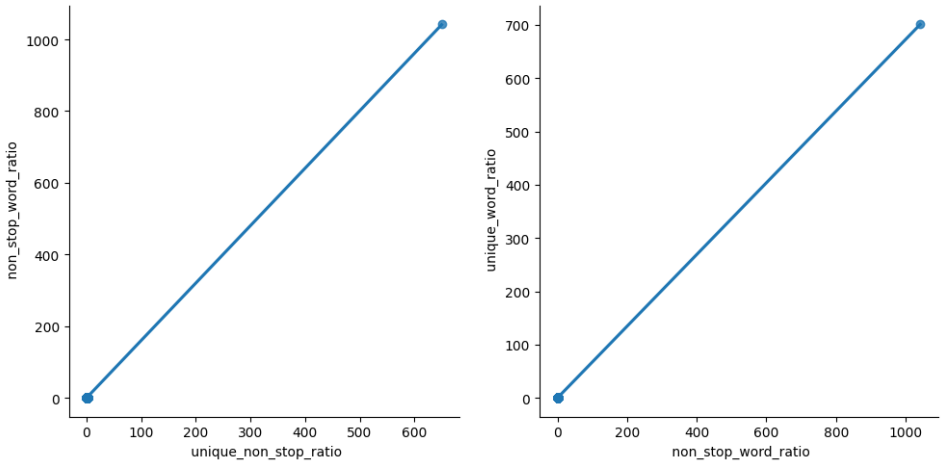

Figura 26: Corelații eronate între atributele numerice ale setului de date datorate outlierilor

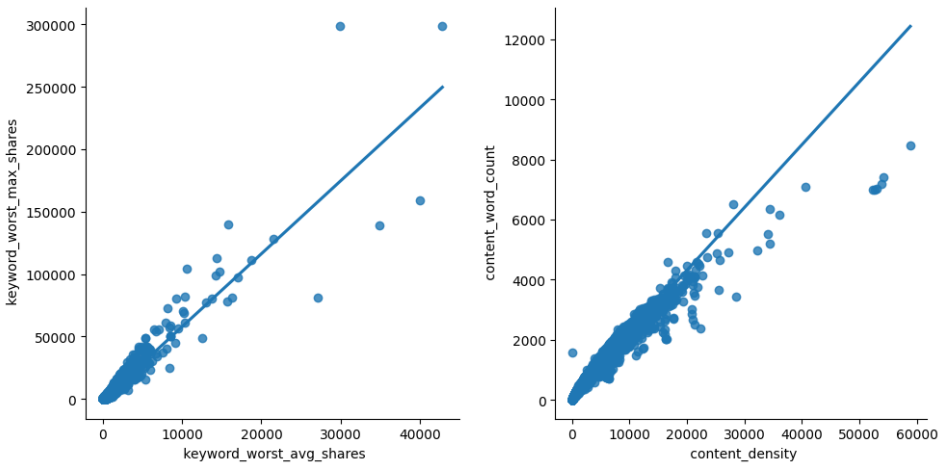

Figura 27: Corelații puternice între atributele numerice ale setului de date

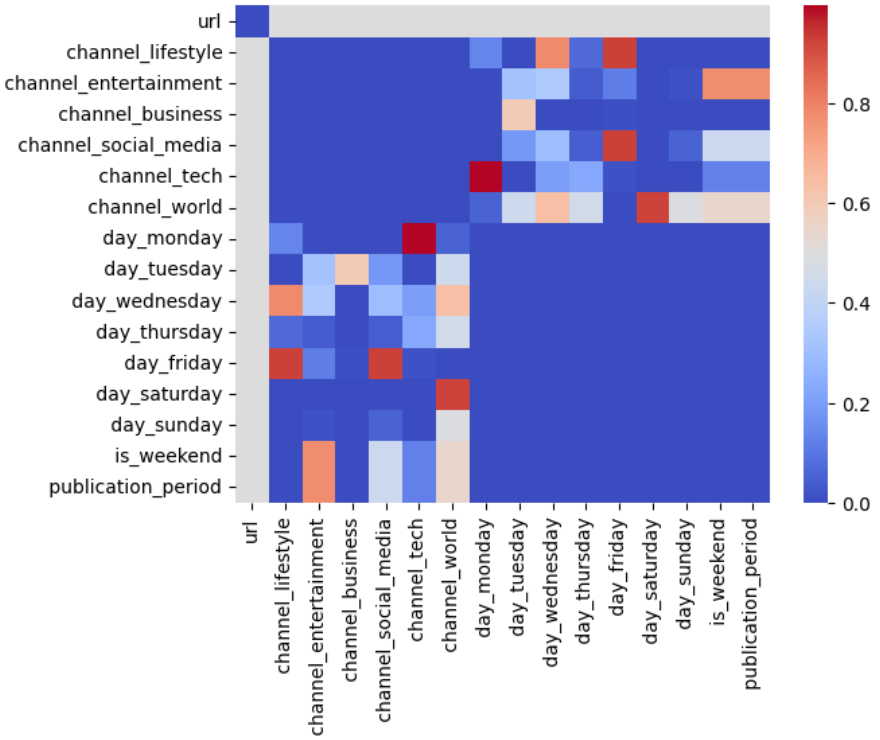

Figura 28: Testul χ^2 pentru atributele categorice ale setului de date

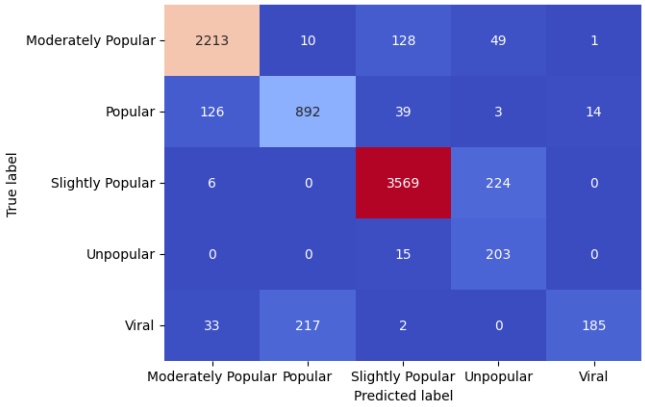

Figura 29: Matricea de confuzie a modelului de tip arbore de decizie

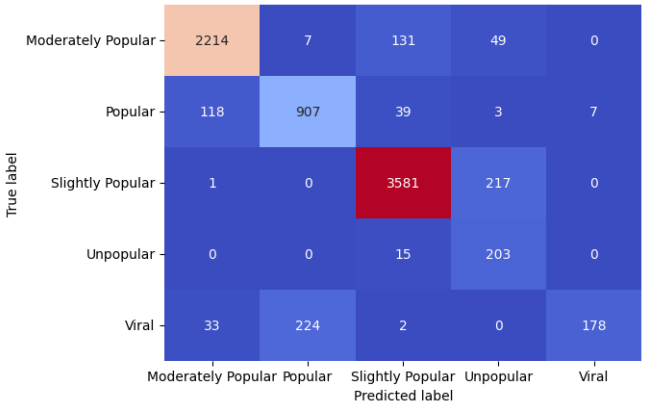

Figura 30: Matricea de confuzie a modelului de tip pădure aleatoare

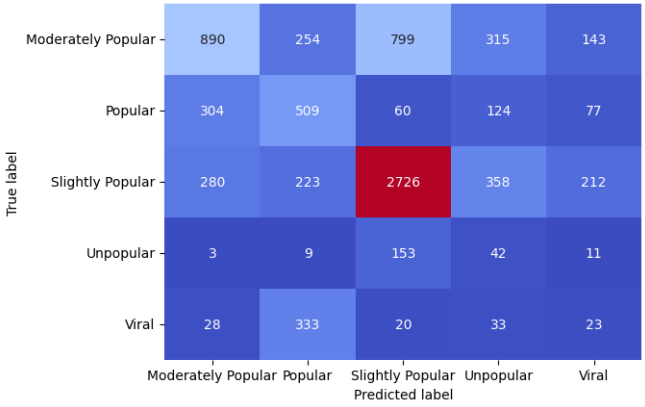

Figura 31: Matricea de confuzie a modelului de regresie logistică

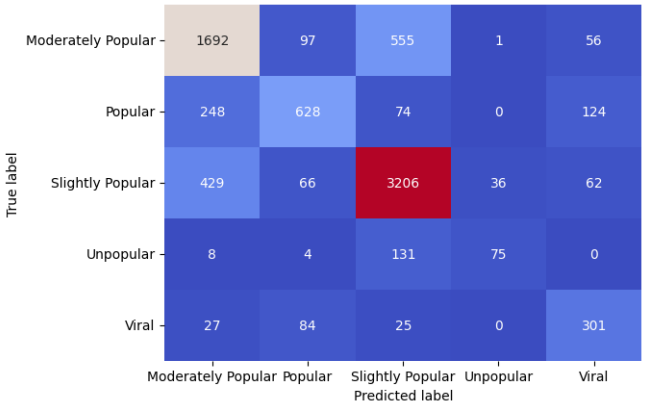

Figura 32: Matricea de confuzie a modelului de tip rețea neurală adâncă

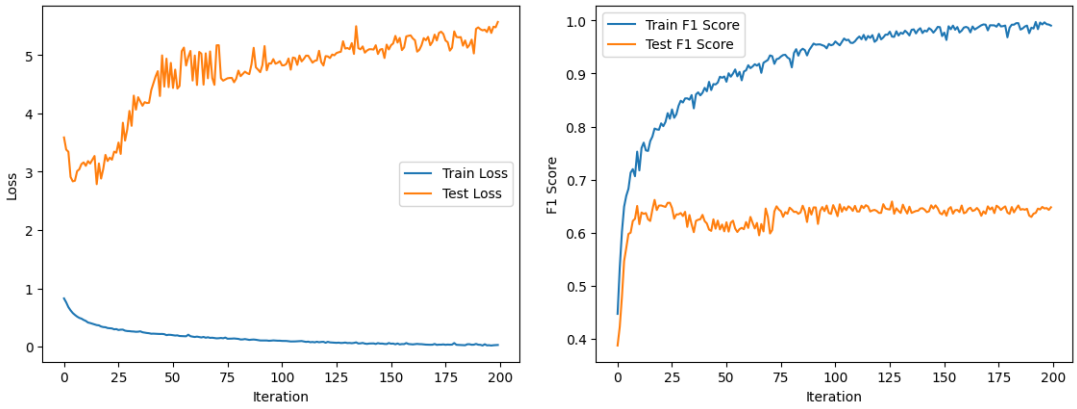

Figura 33: Curbe de învățare și de acuratețe pentru rețea

3 Concluzii

În urma analizei și a învățării automate, s-au obținut rezultate foarte bune pentru ambele seturi de date, cu o acuratețe de peste 90% în cazul poluării și de peste 70% în cazul popularității știrilor. Cele mai bune predicții au fost realizate folosind păduri aleatoare sau arbori de decizie, diferența dintre cele 2 modele fiind că arborii de decizie sunt mai rapizi, dar pădurile aleatoarea scalează mai bine la seturi de date mari, putând fi mărit numărul de arbori. Rețelele neurale au avut rezultate decente, care ar fi putut fi îmbunătățite cu costul timpului de antrenament. Acestea s-au comportat mai slab decât modelele bazate pe arbori de decizie posibil datorită atributelor seturilor de date (fiind mai facil clasificate prin decizii). Regresia liniara s-a comportat cel mai slab, aceasta fiind folosită în principal pentru clasificări binare.