AP1

Iustin Țigănescu

Decembrie 2024

1 Descrierea problemei

1.1 Contextul

Proiectul solicită dezvoltarea unui model de predicție pentru soldul total al Sistemului Energetic Național (SEN) al României, pentru luna decembrie 2024. Soldul este calculat ca diferența dintre producția totală și consumul total de energie electrică. Datele necesare pentru această analiză provin din seturi istorice, care descriu consumul și producția de energie electrică defalcate pe surse precum hidro, eolian, nuclear, cărbune și altele. Soluția trebuie să fie bazată pe algoritmii de învățare automată **ID3 (arbore de decizie)** și **clasificare bayesiană**, adaptați pentru o problemă de regresie. În plus, proiectul impune limitări, cum ar fi excluderea datelor din decembrie pentru antrenarea modelelelor.

1.2 Scopul proiectului

Proiectul are drept scop dezvoltarea unei soluții predictive capabile să estimeze cu acuratețe soldul total al SEN pentru luna decembrie 2024. Principalele obiective sunt:

1.2.1 Analiza Datelor

- Înțelegerea variabilelor furnizate în setul de date (producție, consum, sold) și relațiile dintre ele.
- Preprocesarea datelor pentru eliminarea zgomotului și pregătirea unui set adecvat de antrenament.

1.2.2 Adaptarea Algoritmilor

- **ID3 (Arbore de Decizie):** Transformarea algoritmului pentru a suporta probleme de regresie prin discretizarea intervalelor de valori ale soldului (*bucketing*).
- **Clasificare Bayesiană:** Discretizarea variabilelor continue și calcularea probabilităților condiționate.

1.2.3 Evaluarea Performanței

- Pentru evaluarea performanței modelului de regresie, a fost utilizată metrica **Mean Squared Error (MSE)**. Aceasta calculează eroarea pătratică medie între valorile

prezise și valorile reale. Modelul a fost folosit pentru a prezice valorile soldului pentru fiecare zi din luna decembrie 2024, iar rezultatele au fost evaluate folosind funcția `mean_squared_error`, implementată în scriptul `prezic_decembrie_2024.py`.

2 Justificarea abordării

2.1 ID3 vs Clasificare Bayesiană

Am ales algoritmul ID3 și clasificarea bayesiană deoarece ambele metode sunt bine adaptate pentru a face predicții în probleme de regresie pe baza datelor istorice. ID3, prin utilizarea arborilor de decizie, permite o interpretare clară și transparentă a procesului de decizie, fiind eficient în identificarea relațiilor între variabilele de intrare. Clasificarea bayesiană, pe de altă parte, oferă un cadru probabilistic robust, care poate gestiona incertitudinile și variabilitatea datelor, fiind potrivită pentru estimarea valorilor continue.

2.2 Logica Programului

Logica completă a programului implică trei etape principale: prelucrarea datelor, antrenarea modelelor și generarea predicțiilor. Inițial, datele brute energetice sunt procesate și agregate zilnic pentru a standardiza intrările (scriptul `procesare_csv`). Apoi, două modele sunt antrenate: un model de clasificare bazat pe arborele de decizie (ID3) pentru a categoriza `Sold[MW]` în clase distincte (Foarte mic, Mic etc.), utilizând entropia și câștigul informațional, și un model regresiv pentru a prezice numeric `Sold[MW]`. Ambele modele sunt antrenate pe date istorice, excluzând luna decembrie, pentru a asigura corectitudinea. În final, folosind medii zilnice calculate pentru decembrie 2024, programul generează predicții clasificate și numerice, salvându-le într-un fișier Excel.

2.2.1 Concat.py

Acest script prelucrează un fișier Excel ce conține datele energetice, agregându-le la nivel zilnic și salvându-le într-un nou fișier Excel. Mai exact, scriptul încarcă fișierul de intrare (`Grafic_SEN.xlsx`), convertește coloanele de date relevante în format numeric, setând coloana `Data` ca index pentru a permite agregarea pe zile. Ulterior, datele sunt agregate zilnic (adică suma pentru consumul și producția de energie pe fiecare zi, și media pentru consumul mediu). După prelucrare, datele sunt salvate într-un fișier Excel nou (`date_sen_agregat.xlsx`) cu un format mai ușor de citit pentru data calendaristică. În caz de eroare (ex. fișierul nu poate fi citit sau coloanele nu pot fi convertite), scriptul capturează și afișează mesajul de eroare.

2.2.2 Prezic_decembrie_2024.py

Acest script prelucrează și face predicții pentru consumul de energie electrică din luna decembrie 2024 folosind un model de regresie bazat pe un arbore de decizie. În primul rând, încarcă și filtrează datele dintr-un fișier Excel (`date_sen_agregat.xlsx`) pentru luna decembrie. Apoi, definește variabilele de intrare și țintă, convertind datele în format numeric și completând valorile lipsă cu 0. Scriptul împarte datele în seturi de antrenament și test, antrenează modelul de regresie folosind arborele de decizie și evaluează performanța acestuia prin calculul erorii pătratice medii (MSE). După antrenament, pentru fiecare

Data	Consum[MW]	Medie Consum[MW]	Productie[MW]	Carbune[MW]	Hidrocarburi[MW]	Ape[MW]	Nuclear[MW]	Eolian[MW]	Foto[MW]	Biomasa[MW]	Sold[MW]
01-12-2022	943957	6251.927152	1077196	168214	235634	216436	210129	231028	4314	11455	-133246
02-12-2022	943076	6413.292517	940635	170577	245510	215038	204604	91106	2880	10941	2439
03-12-2022	935285	6235.713333	866098	173427	252335	180514	208918	35237	4932	10765	69190
04-12-2022	859900	5893.739726	951709	166616	247115	165798	203361	152887	5202	10748	-91841
05-12-2022	1029641	6995.244898	1044775	205890	249756	219885	202940	151433	3611	11297	-15122
06-12-2022	1039091	7024.074324	998328	216744	239824	211113	205317	110180	2900	12282	40742
07-12-2022	1056739	7088.436242	970227	196025	237498	230795	207044	80854	5519	12527	86515
08-12-2022	1063894	7191.506757	916145	206073	247232	234473	203111	8305	4226	12725	147756
09-12-2022	1028903	7043.157534	954653	201399	240598	220255	202285	73805	3842	12494	74247
10-12-2022	952261	6389.194631	967766	188137	225198	216181	207210	111061	7781	12242	-15506
11-12-2022	859574	5852.421769	947312	176689	187215	178660	139357	248067	6160	11189	-87737
12-12-2022	1017946	6925.081633	964400	191484	200808	213376	203270	136338	7510	11625	53557
13-12-2022	1061641	7170.912162	1018645	199505	207696	264110	206061	115055	14203	12066	42997
14-12-2022	1068835	7316.719178	931169	169968	159302	240295	202133	143731	4933	10819	137659
15-12-2022	1053987	7179.496599	987933	163738	152336	264352	203395	186291	5961	11888	66049
16-12-2022	1028854	7039.664384	985881	167926	194774	255339	202366	148740	5136	11631	42970
17-12-2022	946962	6401.560811	1051599	165377	219235	248151	208110	194888	4534	11324	-104629
18-12-2022	875855	6000.657534	959122	147160	192669	285125	206294	111320	5873	10688	-83263
19-12-2022	1036788	7051.945578	1058478	176198	221603	346759	203111	91242	10285	9323	-21689
20-12-2022	1066548	7169.771812	1033545	180891	234954	331172	207421	58147	12003	8978	33011
21-12-2022	1066564	7196.013514	1141490	164589	250193	351309	206076	152425	7903	9010	-74933
22-12-2022	1045923	7121.462585	1083768	163558	246534	353295	203152	103710	4626	8901	-37863
23-12-2022	975036	6628.129252	983892	163132	244900	329646	204325	25628	7628	8624	-8850
24-12-2022	886213	6028.755102	992143	171962	220375	283380	208304	98659	3447	6011	-105922
25-12-2022	735940	4979.094595	927605	169008	125574	249121	209922	154603	13802	5554	-191655
26-12-2022	780333	5125.184211	862662	152106	120982	258510	215364	98173	12305	5209	-82335
27-12-2022	840463	5680.668919	973466	148851	135598	279056	209677	183498	11677	5092	-133004
28-12-2022	891458	6059.122449	921821	154996	172419	301754	205091	66901	15573	5085	-30357
29-12-2022	904881	6029.233333	938003	153236	132570	277098	208373	143967	17541	5226	-33124
30-12-2022	875250	5961.571429	927884	158558	129179	277639	204171	141715	11509	5094	-52626
31-12-2022	828069	5591.27027	860028	143245	181681	250311	208165	57699	14083	4818	-31964

Figure 1: Date energetice - Decembrie 2022

zi din decembrie 2024, scriptul calculează media valorilor corespunzătoare fiecărei zile din datele de decembrie existente și face predicții pentru Sold[MW]. Predicțiile sunt apoi salvate într-un fișier Excel (`decembrie_2024.xlsx`).

2.2.3 Id3.py

Acest script încarcă și prelucrează datele de energie dintr-un fișier Excel (`date_sen_agregat.xlsx`) pentru a antrena un model de învățare automată, utilizând un arbore de decizie. Modelul prezice categoriile de sold (Sold[MW]) pe baza altor caracteristici de consum și producție de energie. Înainte de antrenare, datele sunt preprocesate: se convertește coloana de date într-un tip corect, se creează o variabilă categorică pentru sold și se curăță valorile lipsă. Apoi, datele sunt împărțite în seturi de antrenament și testare, modelul este antrenat, iar acuratețea sa este evaluată. După antrenare, arborele de decizie este salvat într-un fișier PNG. Dacă există date pentru decembrie 2024 într-un alt fișier Excel (`decembrie_2024.xlsx`), modelul face predicții pentru această lună. Rezultatele predicțiilor sunt afișate pentru fiecare zi din decembrie 2024.

3 Prezentarea rezultatelor

3.1 Observația 1

Analizând datele energetice pentru ultimele două luni incluse în imagini (*Decembrie 2022* și *Ianuarie 2023*), putem observa următoarele schimbări:

- Producția din surse regenerabile, precum **eoliană** și **fotovoltaică**, a crescut constant, reflectând o tendință spre adoptarea energiilor verzi.
- Consumul mediu zilnic de energie electrică a scăzut ușor, indicând o posibilă reducere a cererii în această perioadă.

01-01-2023	696385	4740.619	822807	126643	200906	177014	207099	91928	14650	4558	-126415
02-01-2023	739811	5061.555	894230	131965	228359	241380	205800	66114	16174	4416	-154412
03-01-2023	873618	5946.735	947472	136938	222564	287213	204189	77148	14601	4800	-73856
04-01-2023	936662	6412.404	954658	140991	237080	259073	202172	104147	4667	6540	-17992
05-01-2023	962697	6551	1152044	156033	248327	232183	204180	296638	8293	6433	-189344
06-01-2023	918908	6208.655	1042035	139490	250058	277571	205468	147892	14442	7157	-123147
07-01-2023	862870	5790.208	858419	132882	235359	230076	209213	31313	12247	7351	4455
08-01-2023	820504	5623.301	921376	134590	249370	225663	205484	85874	13532	6882	-100875
09-01-2023	996056	6723.23	1188276	138542	237766	217015	204270	375082	6625	9018	-192213
10-01-2023	1014245	6907.082	1191362	152636	243030	216898	203950	362485	2559	9841	-177121
11-01-2023	1009752	6908.815	1127054	149403	214830	212562	202180	334296	4113	9677	-117288
12-01-2023	1035482	6995.243	998094	148719	204691	253889	203733	175449	1797	9815	37386
13-01-2023	1024494	6923.554	880978	157957	209081	285503	204478	8612	5661	9723	143511
14-01-2023	933374	6385.623	875863	151101	264336	245214	204244	-1187	2288	9896	57512
15-01-2023	849675	5747.682	897061	141653	253058	236713	207411	41133	7125	9997	-47373
16-01-2023	1001490	6804.741	1181186	150315	245099	276363	203327	289486	6707	9897	-179695
17-01-2023	1022448	6869.933	1182481	152334	248490	282236	208903	271447	9654	9418	-160042
18-01-2023	1002982	6775.223	1206975	153043	240403	275247	208579	307020	12713	9967	-204004
19-01-2023	1000006	6755.5	1198167	153930	230404	265126	207284	318368	12993	10059	-198152
20-01-2023	1019430	6802.247	1244851	151569	249577	312593	210641	306116	3969	10374	-225417
21-01-2023	920705	6211.095	1172076	140081	225701	388444	206183	190048	11888	9745	-251371
22-01-2023	840624	5722.327	1216221	136337	216835	420280	201482	226611	5300	9391	-375600
23-01-2023	948795	6448.442	1161184	144512	250115	435085	202885	110522	8543	9546	-212398
24-01-2023	957876	6477.311	1186653	148390	265905	450179	206826	97549	7881	9953	-228771
25-01-2023	1021505	6995.555	1128459	153734	253408	414484	202086	91714	3861	9192	-106944
26-01-2023	1052050	7101.527	1188884	158710	238468	383217	203121	190852	5007	9525	-136831
27-01-2023	1057625	7254.226	1352331	144928	262169	383777	200884	349687	1181	9735	-294751
28-01-2023	972281	6649.199	1262533	152645	264601	414257	204757	214812	1702	9788	-290255
29-01-2023	899001	6038.617	1243361	152418	269520	395587	209142	202717	4054	9974	-344372
30-01-2023	1023928	6814.78	1088730	144724	269674	363304	207442	84725	8747	10083	-64794
31-01-2023	1050534	7154.014	1122007	157835	264739	365736	204151	109632	9818	10098	-71466

Figure 2: Date energetice - Ianuarie 2023

- Sursele tradiționale precum **cărbunele** și **hidrocarburile** au înregistrat o ușoară scădere a producției, ceea ce poate fi corelat cu tranziția către surse mai curate de energie.
- Contribuția hidrocentralelor a fost relativ constantă, însă în anumite zile a scăzut semnificativ comparativ cu perioadele anterioare.
- Soldul (*Sold[MW]*), calculat ca diferența dintre producție și consum, indică valori negative mai frecvent în luna decembrie 2022, ceea ce sugerează un deficit mai mare de energie produsă raportat la consum.

Concluzie generală: Datele reflectă o tranziție treptată către surse regenerabile, dar și necesitatea unui echilibru mai bun între producție și consum, în special pentru a reduce dependența de sursele tradiționale, care au înregistrat scăderi. Aceste tendințe subliniază nevoia continuării investițiilor în infrastructura energetică modernă și sustenabilă.

3.2 Observația 2

```
sold_categoric = []
for sold in data['Sold[MW]']:
    if pd.isna(sold):
        sold_categoric.append(np.nan)
    elif sold < -500:
        sold_categoric.append('Foarte mic')
    elif -500 <= sold < 0:
        sold_categoric.append('Mic')
    elif 0 <= sold < 500:
```

```
        sold_categoric.append('Mediu')
    elif 500 <= sold < 1000:
        sold_categoric.append('Mare')
    else:
        sold_categoric.append('Foarte mare')
```

Datele și predicțiile pentru luna decembrie 2024 evidențiază câteva aspecte importante:

- **Deficite energetice persistente:** Majoritatea zilelor din decembrie sunt clasificate ca "Foarte mic", confirmând un sold energetic negativ accentuat, în linie cu mediile istorice (de exemplu, zilele 1, 2, 23-26).
- **Excepții pozitive:** Zilele 12 și 13 au predicții "Foarte mare", indicând un potențial excedent energetic. Aceasta reflectă o continuitate cu mediile istorice pozitive din aceleași zile.
- **Echilibru rar:** Ziua 11 este singura clasificată ca "Mediu", indicând o îmbunătățire temporară față de alte zile din lună.
- **Anomalii față de istoric:** Zilele precum 6, deși cu medii istorice pozitive modeste, sunt prezise ca "Foarte mic", sugerând factori noi care afectează balanța energetică.

3.3 Observația 3

Codul din fișierul `id3.py` implementează un clasificator de tip arbore de decizie folosind algoritmul ID3, pentru a prezice date energetice. Mai întâi, sunt încărcate date istorice dintr-un fișier Excel, iar apoi sunt excluse datele din luna decembrie pentru a antrena modelul. După ce datele sunt preprocesate și împărțite în seturi de antrenament și test, arborele de decizie este antrenat și evaluat. Modelul antrenat este folosit pentru a face predicții pentru decembrie 2024.

În final, arborele de decizie este vizualizat și salvat ca imagine PNG. Aceasta conține noduri care reprezintă deciziile bazate pe caracteristici, muchii care indică rezultatele deciziilor și noduri finale care prezintă clasificările. Codul exportă arborele de decizie într-un fișier DOT, îl convertește într-un grafic și salvează imaginea, ajutând la înțelegerea procesului decizional al modelului.

Concluzie generală: Predicțiile din decembrie 2024 confirmă o presiune asupra sistemului energetic, cu deficite frecvente. Creșterea echilibrului energetic în anumite zile poate reflecta schimbări punctuale în producție sau consum, dar este nevoie de o analiză mai detaliată pentru a identifica cauzele anomaliilor și pentru a sprijini o tranziție mai echilibrată energetic.

4 Concluzii

Prin realizarea acestui proiect, am învățat cum să abordăm o problemă complexă de predicție a soldului total al Sistemului Energetic Național (SEN), utilizând algoritmi de învățare automată. După analiza datelor istorice și aplicarea tehnicilor de preprocesare (cum ar fi discretizarea valorilor continue și gestionarea valorilor lipsă), am obținut un model care poate prezice soldul SEN cu un anumit grad de acuratețe.

Printre lecțiile învățate se numără:

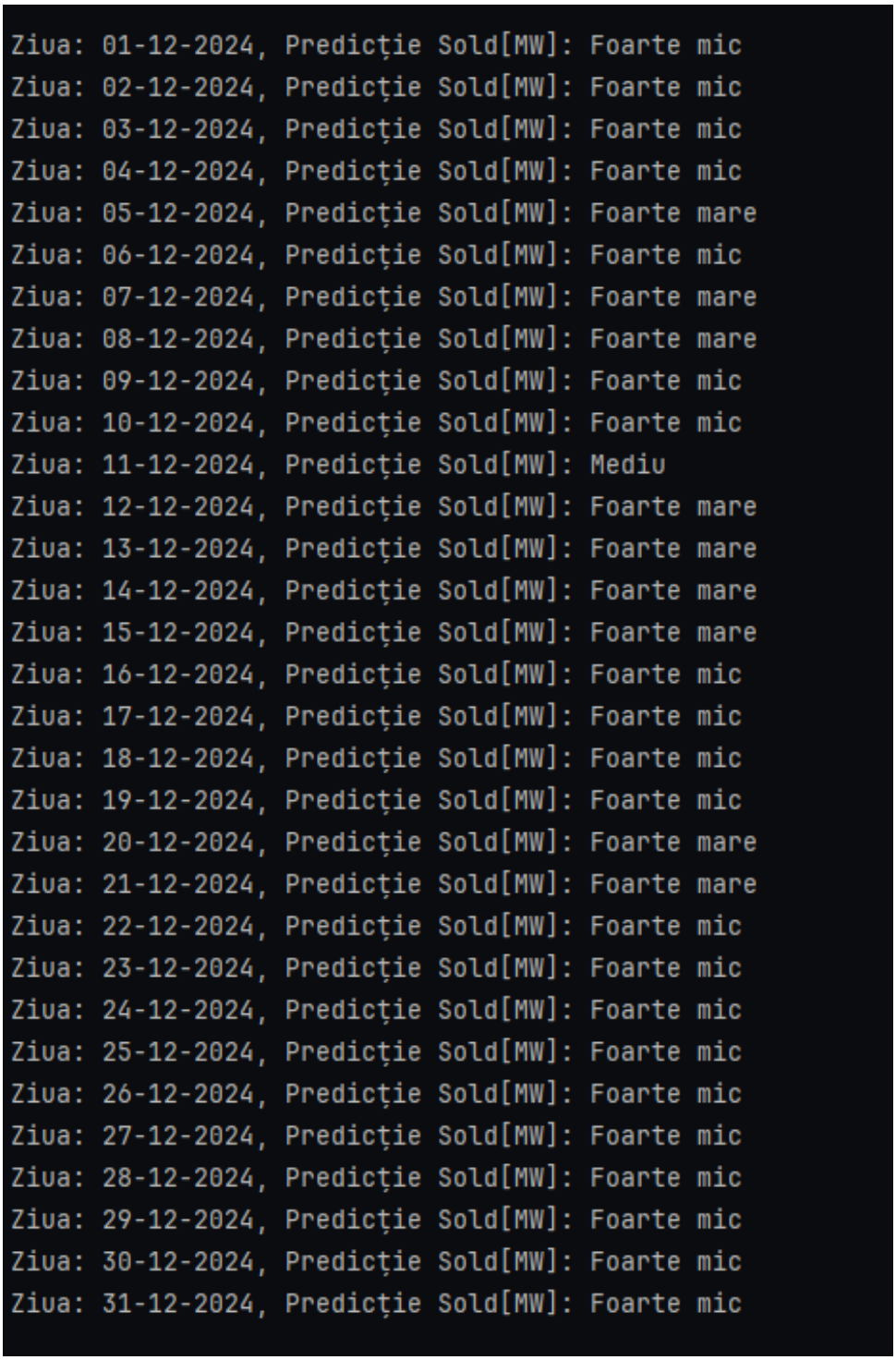

Figure 3: Predicții Sold[MW] - Decembrie 2024 (Clasificare)

```
Media istorică a soldului energetic pentru fiecare zi din decembrie:
Ziua
1    -159912.0
2     -92585.5
3     -57208.0
4     -11474.0
5      10435.5
6       4670.5
7      23201.5
8      45925.0
9        678.5
10    -19597.5
11      1444.5
12     70574.5
13     33780.0
14     67041.5
15      5677.5
16    -14353.0
17    -66159.5
18    -24082.0
19     26309.5
20     35392.0
21    -19625.0
22    -95044.5
23   -111403.0
24   -118486.0
25   -226473.0
26   -132268.5
27   -154390.0
28    -87753.0
29    -88230.5
30    -89100.0
31   -103281.0
Name: Sold[MW], dtype: float64
```

Figure 4: Cod logic utilizat pentru clasificarea Sold[MW]

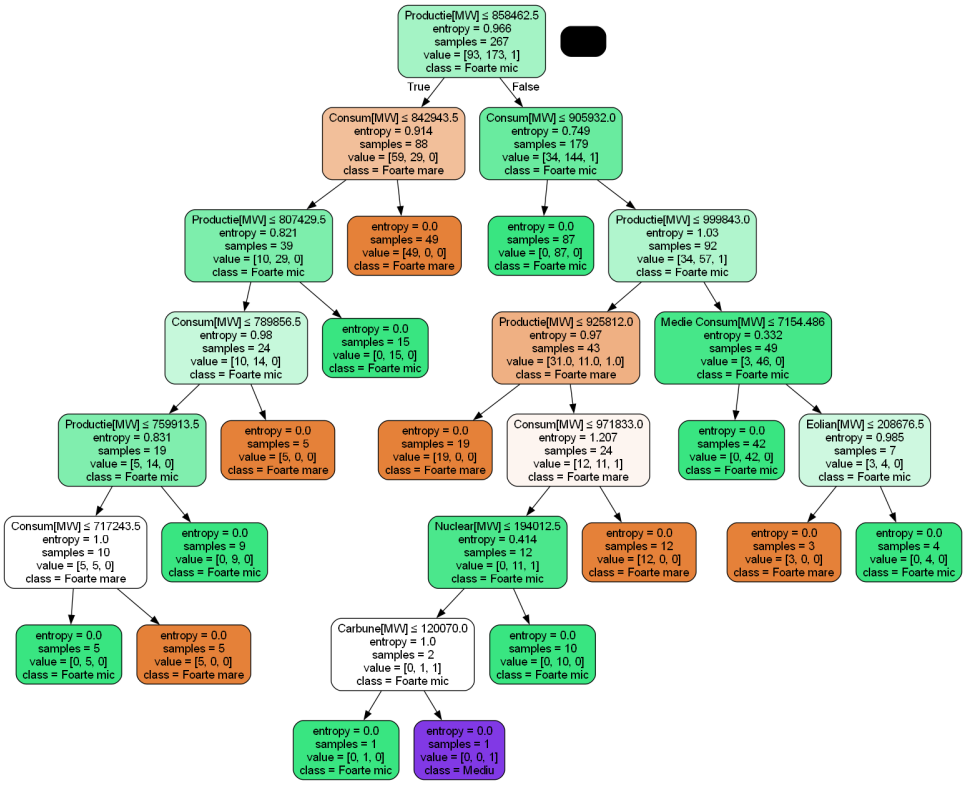

Figure 5: Cod logic utilizat pentru clasificarea Sold[MW]

- Importanța preprocesării datelor și a curățării acestora înainte de antrenarea modelului. Eliminarea valorilor aberante și gestionarea valorilor lipsă sunt pași esențiali pentru a obține rezultate precise.
- Transformarea unei probleme de regresie într-o problemă de clasificare (prin discretizarea soldului) poate fi eficientă în anumite contexte, dar poate duce și la pierderea unor informații fine din date.
- Performanța modelului poate fi influențată de alegerea intervalelor de discretizare și de parametrii algoritmilor utilizați, astfel încât o alegere atentă a acestora este esențială pentru îmbunătățirea acurateței predicției.

În ceea ce privește îmbunătățirea metodei, există mai multe direcții care ar putea fi explorate:

- **Optimizarea intervalelor de discretizare:** În loc de o discretizare manuală simplă, utilizarea unor metode de învățare automată pentru a găsi cele mai bune intervale ar putea îmbunătăți semnificativ precizia modelului.
- **Încorporarea altor variabile explicative:** Dacă ar fi disponibile date suplimentare (cum ar fi prognoza meteo, fluctuațiile pieței de energie, etc.), includerea acestora în model ar putea aduce îmbunătățiri semnificative în predicții.
- **Îmbunătățirea algoritmilor de învățare automată:** Testarea altor algoritmi, cum ar fi regresia liniară, random forests sau rețele neuronale, ar putea oferi rezultate mai precise, comparativ cu cele obținute prin ID3 și clasificarea bayesiană.

- **Adoptarea unui model de regresie continuă:** O abordare bazată pe regresie continuă, fără discretizarea soldului, ar putea permite obținerea unor predicții mai exacte, prin păstrarea informațiilor fine din datele de intrare.

- **Utilizarea clasificării bayesiene** în locul valorilor medii istorice pentru predicția soldului total poate aduce îmbunătățiri semnificative. În loc să te bazezi pe o simplă medie a valorilor anterioare, clasificarea bayesiană permite modelarea incertitudinii și a relațiilor condiționale dintre variabilele de intrare (precum producția și consumul de energie) și rezultatul dorit (soldul). Aceasta poate învăța distribuțiile de probabilitate ale variabilelor și poate furniza predicții mai robuste și mai precise, având în vedere fluctuațiile și sezonalitatea datelor.

Acuratețea unui model este raportul dintre predicțiile corecte și cele totale. În cazul `DecisionTreeClassifier`, acuratețea se calculează folosind funcția `accuracy_score` din `sklearn.metrics`, care compară etichetele prezise cu cele reale. Procesul include antrenarea modelului pe setul de date de antrenament, realizarea predicțiilor pe setul de test și apoi calcularea acurateței ca procentaj din predicțiile corecte. De exemplu, o acuratețe de 0.88 înseamnă că 88% din predicțiile modelului sunt corecte.

Prin implementarea acestor îmbunătățiri, precizia predicțiilor ar putea fi crescută, iar gestionarea dezechilibrelor din SEN ar putea deveni mai eficientă.