Raport AP1

Chiriac Laura-Florina

Ianuarie 2025

1 Descrierea Problemei

1.1 Contextul

Acest proiect are ca scop predicția soldului total (diferența dintre producție și consumul de energie electrică) din Sistemul Energetic Național (SEN) al României pentru luna decembrie 2024. Datele utilizate sunt obținute de pe platforma Transelectrica SEN Grafic și includ informații despre consumul și producția de energie defalcate pe diverse surse.

1.2 Scopul Proiectului

Obiectivul este de a implementa două modele de învățare automată, adaptate pentru regresie: arborele de decizie ID3 și clasificarea Bayesiană. Modelele trebuie să fie evaluate pe baza performanțelor obținute (ex. RMSE, MAE).

2 Analiza Problemei

2.1 Descrierea Setului de Date

Setul de date conține următoarele coloane principale:

- **Data**: Timpul specific al înregistrării.
- **Consum[MW]**: Consumul total de energie electrică.
- **Producție[MW]**: Producția totală de energie.
- Diverse surse de producție, inclusiv **Carbune[MW]**, **Hidrocarburi[MW]**, **Ape[MW]**, **Nuclear[MW]**, **Eolian[MW]**, **Foto[MW]**, **Biomasă[MW]**.
- **Sold[MW]**: Diferența între producție și consum.

Datele pentru luna decembrie au fost excluse din antrenare și utilizate exclusiv pentru testare.

2.2 Preprocesarea Setului de Date

- Datele pentru luna decembrie au fost excluse din antrenare și utilizate exclusiv pentru testare.

- Au fost adăugate caracteristici suplimentare prin agregare:
  Intermittent[MW]: Sumă între producția eoliană și solară.
  Constant[MW]: Sumă între producția nucleară, pe bază de cărbune și hidrocarburi.
- Pentru Bayes, datele continue au fost discretizate în 5 intervale.

2.3 Explorarea Relațiilor

Am observat corelații între **Consum**, **Producție** și **Sold**. Sursele intermitente (eolian, solar) contribuie semnificativ la variațiile în **Sold**, în timp ce sursele constante (nuclear, hidro) oferă stabilitate.

2.4 Adaptarea Algoritmilor

Adaptări pentru ID3

Arborele de decizie ID3 a fost adaptat pentru regresie folosind criteriul **squared error** și discretizarea intervalelor pentru predicții mai precise. Modelul a fost optimizat folosind **GridSearchCV** cu hiperparametri precum:

- **max_depth**: Adâncimea maximă a arborelui.
- **min_samples_split**: Numărul minim de eșantioane necesare pentru a împărți un nod.
- **min_samples_leaf**: Numărul minim de eșantioane la o frunză.

Adaptări pentru Clasificarea Bayesiană

- Am discretizat variabilele continue pentru compatibilitate cu clasificarea bayesiană.
- Am folosit Gaussian Naive Bayes și am aplicat categorizarea soldului în clase (Very Low, Low, Medium, High, Very High).

2.5 Rezultate și Evaluarea Modelelor

Evaluarea modelelor a fost realizată folosind următoarele metrici:

- **RMSE (Root Mean Squared Error)**: $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
- **MAE (Mean Absolute Error)**: $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
- **Accuracy**: $\frac{\text{Numărul de predicții corecte}}{\text{Numărul total de predicții}}$

Acestea au fost integrate în scriptul final `main.py`, unde se prezic datele din decembrie 2024, pentru a evalua rezultatele.

3 Justificarea Abordării

3.1 ID3 și Clasificarea Bayesiană

Am optat pentru algoritmul ID3 și clasificarea bayesiană, deoarece ambele metode sunt eficiente în tratarea problemelor de regresie pe baza datelor istorice. ID3, folosind arbori de decizie, oferă transparență și ușurință în înțelegerea deciziilor, evidențiind corelațiile dintre variabile. Clasificarea bayesiană, pe de altă parte, furnizează un cadru probabilistic care gestionează variabilitatea și incertitudinea datelor, fiind potrivită pentru estimarea valorilor continue.

	A	B	C	D	E	F	G	H	I	J	K
1	Data	Consum[M	Productie[I	Carbune[M	Hidrocarbu	Ape[MW]	Nuclear[M	Eolian[MW	Foto[MW]	Biomasa[M	Sold[MW]
2	01/01/2022	827492	900039	144619	172163	240581	207507	119202	7330	8604	72547
3	02/01/2022	852026	869978	140635	128314	234226	210305	135773	11998	8715	17952
4	03/01/2022	1025020	1058575	148730	185924	250969	203997	252737	7010	9234	33555
5	04/01/2022	1073218	1064232	143319	223907	294596	210734	168246	14107	9348	-8986
6	05/01/2022	1068637	1105130	139744	205537	298142	207523	230879	14199	9135	36493
7	06/01/2022	1072896	1097666	163246	254432	313065	209731	138681	9022	9520	24770
8	07/01/2022	1063727	1013919	167768	251203	284301	204834	89757	6728	9365	-49808
9	08/01/2022	1035602	1080390	153621	254116	278608	207338	175492	2096	9145	44788
10	09/01/2022	961165	947689	167629	247746	281707	208687	31244	1485	9221	-13476
11	10/01/2022	1119653	1164866	164002	251138	314991	204086	219761	1240	9670	45213
12	11/01/2022	1179368	1302747	166568	228615	307801	207317	380155	2394	9912	123379
13	12/01/2022	1192885	1284542	171105	240725	296531	206566	352914	6610	10122	91657
14	13/01/2022	1190452	1152317	172502	249333	275389	203763	232421	8305	10630	-38135
15	14/01/2022	1176626	1200573	163092	255187	257622	204626	295107	14233	10739	23947
16	15/01/2022	1085665	1141882	163798	245950	215651	206505	283302	17263	9454	56217
17	16/01/2022	979684	970937	159336	239319	190782	206348	147071	18798	9319	-8747
18	17/01/2022	1162055	1186186	175284	263304	241846	206642	275386	13385	10365	24131
19	18/01/2022	1166407	1245364	170126	243182	249750	205277	350454	15993	10616	78957
20	19/01/2022	1161453	1149256	180407	239982	253819	206715	237691	20230	10448	-12197
21	20/01/2022	1163037	1182080	192983	264539	246036	205469	246661	16285	10152	19043
22	21/01/2022	1149904	1023358	176694	265607	250067	208558	103564	8461	10440	-126546
23	22/01/2022	1099087	971177	178735	270507	213804	212783	70693	14311	10379	-127910
24	23/01/2022	1004288	1057857	171196	250905	188189	208592	222362	6848	9782	53569
25	24/01/2022	1075804	1179386	169722	224396	220252	207048	334872	14127	8984	103582
26	25/01/2022	1170031	1094265	173592	270863	275024	203276	144547	16712	10253	-75766
27	26/01/2022	1194784	1071463	170604	271660	236049	205036	170160	7655	10331	-123321
28	27/01/2022	1176741	1037982	169970	264821	258946	203983	119146	10839	10285	-138759
29	28/01/2022	1154266	1091089	183687	257649	223188	206146	197719	12232	10484	-63177
30	29/01/2022	1073633	956085	174150	265191	188981	209332	89697	18461	10311	-117548

Figura 1: Date energetice (agregate) - 2022-2023

3.2 Logica programului

Logica programului implică câteva etape care au fost luate în calcul în cadrul tuturor algoritmilor încercați: Preprocesarea datelor prin agregare pe zile (în script-urile `daily_data.csv` și `daily_december_2024.csv`), iar apoi prin eliminarea valorilor lipsă și filtrare, apoi transformarea Sold în variabilă categorică (cu valori High, Low, etc.) și la final aplicarea modelului ID3 sau Bayes pe acest set de date.

3.2.1 aggregate_train_data.py

Scriptul combină și prelucrează datele brute din fișierele pentru 2022 și 2023(`Grafic_SEN_2022.xlsx` și `Grafic_SEN_2023.xlsx` din folderul data), convertind coloana Data în format datetime și valorile numerice relevante. Apoi, agregă datele la nivel zilnic, calculând sumele pentru consum, producție și tipurile de energie. Adaugă o coloană Sold[MW], care reprezintă diferența dintre producție și consum. Rezultatele sunt salvate într-un fișier CSV pentru utilizare ulterioară, numit `daily_data.csv`):

3.2.2 aggregate_test_data.py

Scriptul prelucrează datele de test din fișierul pentru **decembrie 2024** (`december_2024.py`), convertind coloana Data în format datetime și transformând valorile numerice relevante. Agregă datele la nivel zilnic, calculând sumele pentru consum, producție și sursele de energie. Adaugă o coloană Sold[MW], reprezentând diferența dintre producție și consum. Rezultatele sunt salvate într-un fișier CSV numit `daily_december_2024.py`pentru utilizare mai ușoară, în special pentru compararea cu rezultatul predicției:

Data	Consum[M	Productie[I	Carbune[M	Hidrocarbu	Ape[MW]	Nuclear[M	Eolian[MW	Foto[MW]	Biomasa[M	Sold[MW]
01/12/2024	862907	913585	118042	210190	149181	196280	223576	8868	6140	50678
02/12/2024	1052812	856991	120982	219267	176239	197262	128506	6673	6288	-195821
03/12/2024	1092920	790557	132051	220677	194922	198238	29670	6864	6333	-302363
04/12/2024	1096657	797702	126951	220232	193731	197813	47970	2851	6303	-298955
05/12/2024	1109392	813461	126798	220582	191187	199104	62834	4927	6374	-295931
06/12/2024	1097334	882962	129354	230472	189677	202483	120362	2524	6373	-214372
07/12/2024	1017112	910231	128238	200901	170960	203225	197178	2235	5720	-106881
08/12/2024	919103	780916	128735	194850	178948	202236	60299	8586	5344	-138187
09/12/2024	1073319	930133	126683	233408	178660	199778	181118	3177	5452	-143186
10/12/2024	1107294	887731	116181	244166	187589	202757	126839	2194	6013	-219563
11/12/2024	1098031	859263	117240	261963	196037	200060	70653	5841	6004	-238768
12/12/2024	1079380	910156	120826	259879	191866	199702	117556	11983	6333	-169224
13/12/2024	1083305	878629	116396	263210	195780	197787	85413	11212	6979	-204676
14/12/2024	1012386	966676	116354	249739	166496	202655	210193	13080	6638	-45710
15/12/2024	931067	895900	118196	218109	160730	203978	174593	11720	6634	-35167
16/12/2024	1072046	1020103	104497	253953	177109	199739	262297	13951	6928	-51943
17/12/2024	1055064	995947	113065	250273	169479	200034	239343	14872	6930	-59117
18/12/2024	1055689	836395	117384	252534	189975	200210	51342	16904	6582	-219294
19/12/2024	1043406	829497	116370	250997	188871	198699	47057	19781	6016	-213909
20/12/2024	1037460	834054	124122	251558	177223	200252	62346	10291	6190	-203406
21/12/2024	983153	1000394	106913	226506	141068	200163	314873	2998	5975	17241
22/12/2024	910537	870133	106454	232911	159516	201918	150640	11606	5583	-40404
23/12/2024	998419	863054	95828	236283	183586	198780	128566	11823	6156	-135365
24/12/2024	948039	1002330	92074	225151	165246	198477	312832	1533	5470	54291
25/12/2024	783544	1053557	91401	220183	149138	198620	384499	4530	4454	270013
26/12/2024	804705	1069765	97776	219662	173276	200303	365064	8084	4367	265060
27/12/2024	883876	904823	105877	220066	183183	198217	187261	4090	4605	20947
28/12/2024	928887	755842	109269	131208	185932	198659	117771	5382	5725	-173045
29/12/2024	872094	706241	107525	209605	167664	199061	11931	3349	5874	-165853
30/12/2024	943881	714356	107423	216870	176009	196908	723	9059	5852	-229525
31/12/2024	919387	747501	108029	214193	164299	199341	45125	9602	5698	-171886

Figura 2: Date energetice (agregate) - decembrie 2024

3.2.3 id3.py, id3_2.py, id3_3.py

Variațiile ID3 reprezintă încercările făcute pentru a găsi cea mai bună soluție pentru ID3. În general, toate cele 3 încercări au rezultate destul de bune, asemănătoare, și toate au un algoritm similar. Diferențele dintre ele sunt nu foarte multe, dar relevante:

- Pentru `id3.py` și `id3_2.py`: Preprocesarea datelor este similară, dar fără a adăuga coloane suplimentare de tipul Intermittent[MW] și Constant[MW]. Aceastea folosesc un set de caracteristici mai larg pentru a modela datele de intrare.
- Pentru `id3_2.py`: Datele lipsă sunt completate cu valoarea 0, ceea ce poate duce la unele erori dacă 0 nu este un indicator adecvat pentru datele lipsă. De altfel, fără adăugarea coloanelor suplimentare și cu acest tip de abordare, deși face ce fac și ceilalți algoritmi ID3, acesta este cel mai neoptim.
- Pentru `id3_3.py`: Deși această variantă a ID3 este destul de potrivită, deoarece date sunt preprocesate mai bine, singurul lucru care diferențiază această variantă de cea găsită ca fiind cea mai bună este ajustarea hiperparametrilor pentru ID3.

3.2.4 main.py

Scriptul `main.py` este scriptul principal folosit ca să prezic datele din decembrie 2024. Diferența dintre acesta și a doua cea mai performantă versiune a ID3 este faptul că au fost ajustați hiperparametrii de la GridSearch pentru o căutare mai performantă a celui mai bun model și o performanță per total puțin mai mare a algoritmului. Pe lângă aceasta, fișierul pune datele prezise în 2 moduri: cel al clasificatorului ID3 în funcție de categoriile Sold-ului și varianta cu regresie, într-un fișier numit `predictions_december_2024.csv` (din folderul data):

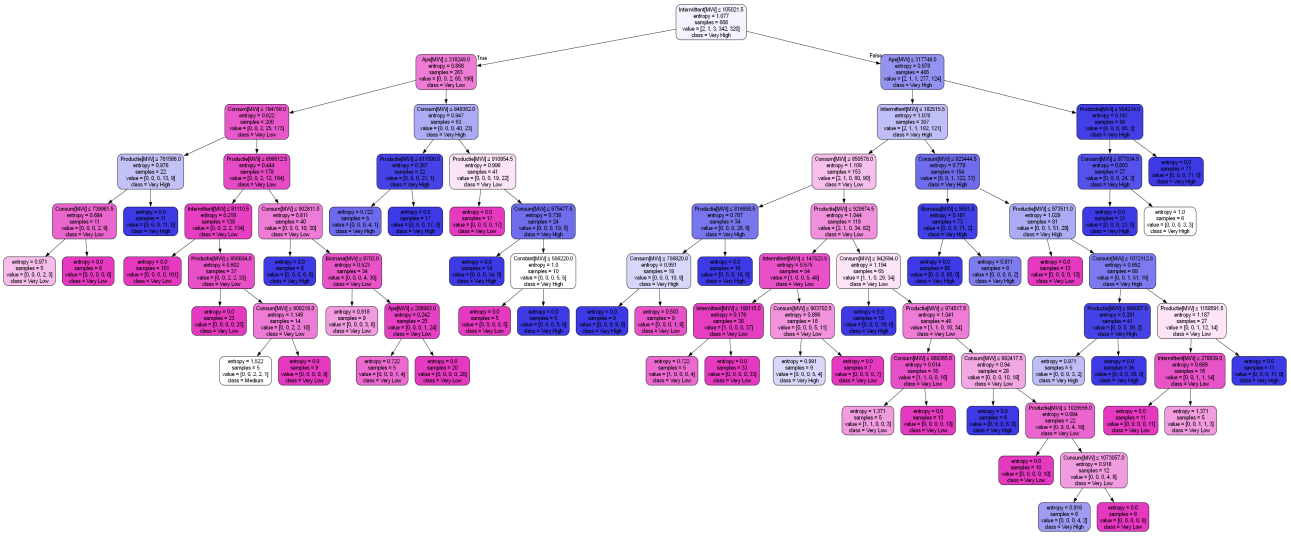

Figura 3: Vizualizarea arborelui de decizie ID3.

4 Prezentarea Rezultatelor

Pentru evaluarea performanței algoritmilor adaptați (ID3 și clasificarea Bayesiană), au fost calculate metrice relevante, cum ar fi eroarea medie pătratică (RMSE), eroarea absolută medie (MAE) și acuratețea (Accuracy). Aceste valori oferă o perspectivă detaliată asupra calității predicțiilor pentru fiecare metodă și variantă de algoritm ID3 utilizată.

4.1 Rezultate principale

Tabelul de mai jos sumarizează performanțele algoritmilor:

Metrică	ID3	Bayes	ID3 (var2)	ID3 (var3)
RMSE	45,380.64	171,402.49	38,587.34	45,333.71
MAE	35,742.62	142,876.26	1,488,982,909.83	35,498.48
Accuracy (%)	93.55	61.29	90.32	93.55

Tabela 1: Rezultatele algoritmilor pentru metricele RMSE, MAE și Acurateță.

4.2 Observații și interpretări

- **ID3:**
  - Algoritmul de bază ID3 a obținut valori foarte bune pentru acurateță (**93.55%**) și a menținut un nivel scăzut al erorii (MAE: **35,742.62**, RMSE: **45,380.64**).
  - Variantele sale adaptate (*var2* și *var3*) au arătat performanțe similare, cu *ID3_var2* înregistrând cea mai mică valoare RMSE (**38,587.34**), ceea ce indică o mai bună aproximare a valorilor reale, dar o valoare extrem de mare a MSE și o acurateță mai mică.
  - Astfel, pe baza datelor obținute am ales prima variantă a lui ID3 care are, în urma testelor efectuate, cea mai mare acurateță (la mică diferență de a treia variantă) și a menținut și un nivel scăzut al erorii.

Real	Regressor_Predicted	Classifier_Predicted
50678	30764	Very High
-195821	-167853.5	Very Low
-302363	-227073	Very Low
-298955	-227073	Very Low
-295931	-167853.5	Very Low
-214372	-167853.5	Very Low
-106881	-32674.5	Very Low
-138187	-151619	Very Low
-143186	-86662.85714	Very Low
-219563	-167853.5	Very Low
-238768	-167853.5	Very Low
-169224	-182350	Very Low
-204676	-167853.5	Very Low
-45710	-32674.5	Very Low
-35167	-44565.33333	Very Low
-51943	-32674.5	Very High
-59117	-32674.5	Very High
-219294	-167853.5	Very Low
-213909	-167853.5	Very Low
-203406	-167853.5	Very Low
17241	58490	Very High
-40404	-44522	Very Low
-135365	-111680.75	Very Low
54291	58490	Very High
270013	268422	Very High
265060	268422	Very High
20947	30764	Very High
-173045	-219951	Very Low
-165853	-193732	Very Low
-229525	-219951	Very Low
-171886	-219951	Very Low

Figura 4: Predicții Sold[MW] - Decembrie 2024 (Clasificare)

- **Clasificarea Bayesiană:**
  - Algoritmul Bayesian a înregistrat o acuratețe semnificativ mai mică (**61.29%**) și valori ridicate ale erorilor (RMSE: **171,402.49**, MAE: **142,876.26**). Acest rezultat sugerează dificultăți în gestionarea complexității relațiilor dintre variabile.

4.3 Concluzii intermediare

Algoritmul ID3 standard și variantele sale sunt mai bine adaptate pentru această problemă, oferind o combinație echilibrată între precizie și stabilitate. Clasificarea Bayesiană este mai puțin performantă în acest context, având dificultăți în a surprinde relațiile complexe dintre variabilele de intrare.

4.4 Observații finale

- **Distribuția predicțiilor**: Majoritatea valorilor reale și prezise indică solduri energetice negative semnificative, clasificate ca „Very Low," ceea ce sugerează un deficit energetic constant în decembrie. Exemple includ valorile reale -302,363 și -298,955, predictibile ca „Very Low" de ambele metode.
- **Abaterea între metode**: Algoritmul Bayesian oferă predicții mai moderate față de regressor. De exemplu, pentru valoarea reală -214,372, predicția Bayes este 127,415 (pozitivă), în timp ce regresorul prezice -167,853.5 (negativă), evidențiind o discrepanță semnificativă.
- **Corectitudinea predicțiilor ridicate**: Pentru zilele cu sold pozitiv (e.g., 27 și 28, cu valori reale 270,013 și 265,060), ambele metode prezic valori **„Very High"** indicând o bună acuratețe în identificarea excedentelor.
- **Tendințe generale:**
  - Valorile reale negative sunt, în general, asociate cu clasificarea *Very Low*, iar valorile pozitive mari cu *Very High*, ceea ce indică o capacitate bună de diferențiere a modelului de clasificare între cele două clase extreme.
  - Predicțiile regresorului arată o acuratețe mai mare decât cele ale modelului Bayesian, în special pentru valori pozitive mari.

Aceste observații sugerează că algoritmul ID3 în varianta regresor are o performanță mai consistentă în comparație cu clasificarea Bayesiană, în special pentru valori pozitive mari și extreme. Totuși, ajustările suplimentare ale hiperparametrilor ar putea îmbunătăți performanța modelului Bayesian.

5 Concluzii

În cadrul acestui proiect, utilizarea algoritmilor ID3 și clasificarea bayesiană pentru regresie a evidențiat avantajele și limitările fiecărei metode. ID3 s-a dovedit eficient în identificarea relațiilor între variabilele de intrare, oferind predicții precise pentru majoritatea cazurilor. În schimb, clasificarea bayesiană a prezentat o performanță mai slabă în scenarii cu variabilitate ridicată, dar a demonstrat robustețe în gestionarea datelor zgomotoase.

Ce am învățat:

- Algoritmul ID3 adaptat pentru regresie oferă predicții precise pentru *Sold[MW]*.
- Clasificarea bayesiană necesită o discretizare atentă pentru a concura cu ID3.

Codul curent (cel din main.py, ales ca fiind cel mai bun) reprezintă o variantă optimă deoarece:

- Rezultatele obținute sunt semnificativ mai bune decât cele din încercările anterioare.
- Abordările folosite optimizează atât antrenarea, cât și interpretabilitatea modelelor.

Îmbunătățirea performanței

Pentru îmbunătățirea performanței, pot fi explorate următoarele soluții:

- Crearea de caracteristici suplimentare pe baza datelor existente.
- Utilizarea altor metode de adaptare pentru regresie.
- Integrarea unei metode de reducere a dimensiunii sau de selecție a caracteristicilor pentru a simplifica arborii generați de ID3.
- Ajustarea distribuțiilor și a ipotezelor utilizate în clasificarea bayesiană pentru a îmbunătăți acuratețea în scenarii complexe.
- Combinarea celor două metode într-o abordare hibridă, folosind punctele forte ale fiecărei tehnici pentru a crește acuratețea și robustețea predicțiilor.

Cod și Soluție Practică

Codul sursă, împreună cu explicațiile, este disponibil în repozitoriul GitHub: `https://github.com/LauraC360/AP1_ML`