AP1

Cojocaru Adelin-Iulian

January 5, 2025

1 Introducere

Scopul acestui proiect este de a prezice soldul energetic total (*Sold[MW]*) în Sistemul Energetic Național (SEN) din România pentru luna decembrie 2024. Acest raport documentează abordarea adoptată pentru a adapta algoritmul arborelui de decizie ID3 și clasificarea bayesiană la o problemă de regresie, evaluarea performanței acestor modele și concluziile rezultate.

2 Descrierea Problemei

Predicția soldului energetic este esențială pentru gestionarea eficientă a rețelei electrice. Setul de date furnizat de Transelectrica include diferite metrici de producție și consum de energie, iar provocarea constă în prezicerea *Soldului* (diferența dintre producție și consum) pentru luna decembrie 2024. Pentru antrenare, s-au folosit ultimii 5 ani (2020-2024) fără luna decembrie, iar pentru validare s-a folosit doar luna decembrie 2024.

2.1 Prezentarea Setului de Date

- **Dată**: Timpul înregistrării
- **Consum** [**MW**]: Consum total de energie
- **Producție** [**MW**]: Producție totală de energie
- **Producție pe Surse**: Cărbune, Hidrocarburi, Hidro, Nuclear, Eolian, Solar, Biomasă
- **Sold** [**MW**]: Diferența dintre producție și consum

3 Metodologie

Secțiunea de metodologie descrie modul în care au fost adaptați algoritmii ID3 și Bayesian pentru a gestiona problema de regresie a prezicerii valorii *Soldului*.

3.1 Adaptarea Algoritmului ID3

Algoritmul ID3, pentru a fi adaptat la regresie, a fost modificat prin discretizarea variabilei tintă (*Sold[MW]*) în intervale predefinite (sau *buckets*) pentru a transforma variabila continuă într-una categorială. Pentru antrenare și validare, parametrul (*criterion*) a fost setat la entropy iar parametrul (*max_depth*) a fost setat la mai multe valori dinamice pentru a găsi cazul cu cea mai bună acuratețe.

3.2 Adaptarea Clasificării Bayesiene

Algoritmul de clasificare bayesiană a fost adaptat pentru a prezice valori continue prin discretizarea caracteristicilor în intervale cu scopul de a simplifică calculul probabilităților. Pentru fiecare interval al variabilei *Sold*, am estimat probabilitățile condiționate ale fiecărui interval de caracteristici date de acel interval.

3.3 Preprocesarea Datelor

- Setul de date a fost împărțit în seturi de antrenament și testare, asigurându-se că datele din decembrie 2024 au fost excluse din procesul de antrenare.
- Proprietățile coloanelor și formatul dății au fost modificate pentru o gestionare mai ușoară.
- Valorile lipsă sau incomplete au fost șterse sau modificate pentru a păstra consistența.

3.4 Abordări Utilizate

Pentru a observa cum se descurcă algoritmii pe acest set de date, au fost folosite următoarele metode:

- **Metoda 1:** Prezicerea **Soldului**[**MW**] direct fără a utiliza alte coloane, ci doar cele extrase din dată.

- **Metoda 2:** Similar cu metoda 1, doar că antrenăm modelul pe lunile decembrie din anii precedenți.
- **Metoda 3:** Prezicerea tuturor coloanelor individual, ca eventual să calculăm **Soldul**[**MW**] ca diferența dintre producție și consum.
- **Metoda 4:** Similar cu metoda 3, doar că antrenăm modelul pe lunile decembrie din anii precedenți.

4 Analiză. Observații. Performanță.

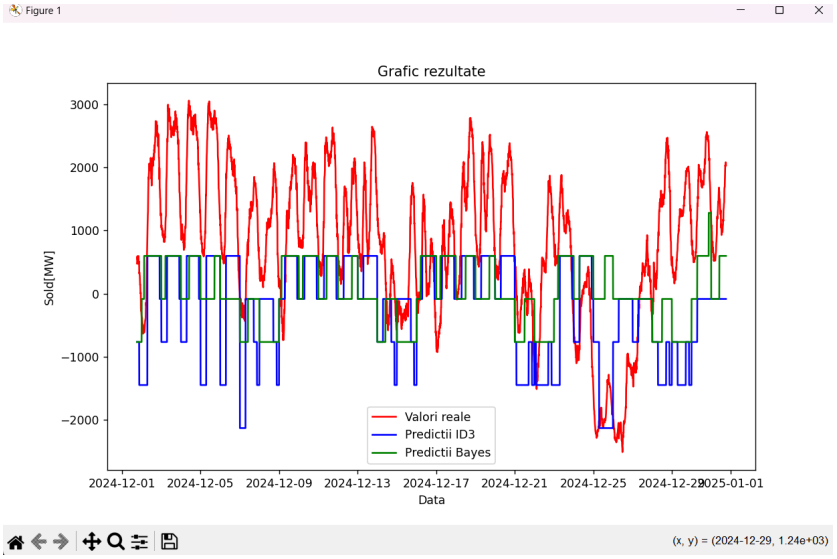

Figure 1: Grafic Sold Only pe 2 ani

Model	RMSE	MAE
Arbore de Decizie ID3	1467.23	1246.12
Clasificare Bayesiană	1417.49	1216.68

Table 1: Tabel Sold Only pe 2 ani

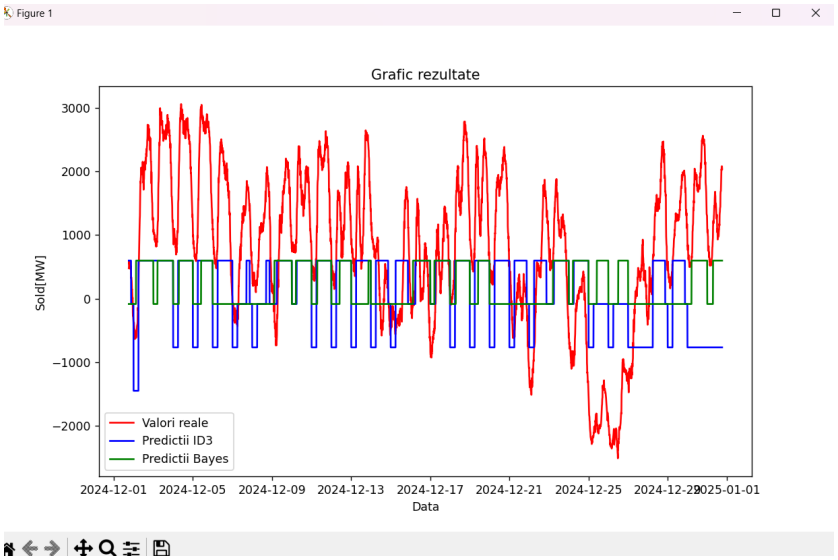

Figure 2: Grafic Sold Only pe 4 ani

Model	**RMSE**	**MAE**
Arbore de Decizie ID3	1278.67	1109.04
Clasificare Bayesiană	1299.87	1094.51

Table 2: Tabel Sold Only pe 4 ani

Se observă o creștere în acuratețe dacă datasetul nostru e mai mare (conține mai mulți ani). Putem trage concluzia că, pentru rezultate mai bune, trebuie crescut datasetul.

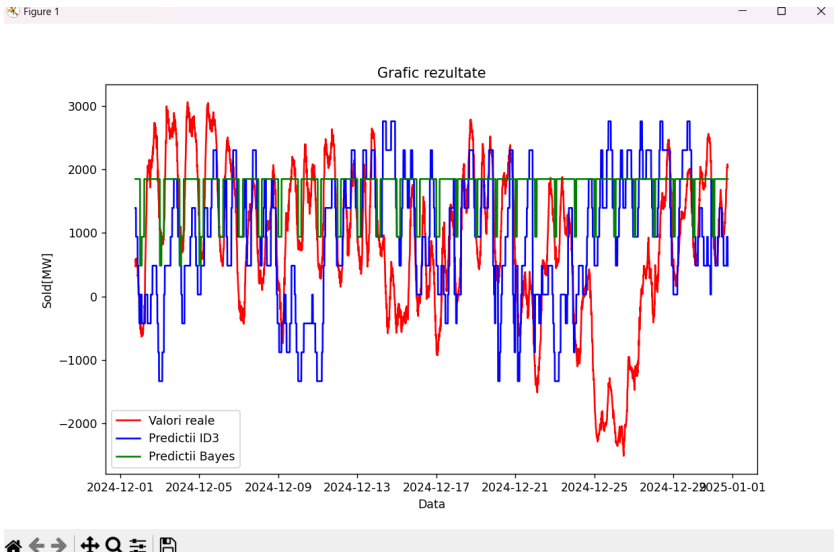

Figure 3: Grafic Sold Only anul inclus

Model	**RMSE**	**MAE**
Arbore de Decizie ID3	1256.43	1043.89
Clasificare Bayesiană	1279.12	1104.04

Table 3: Tabel Sold Only anul inclus

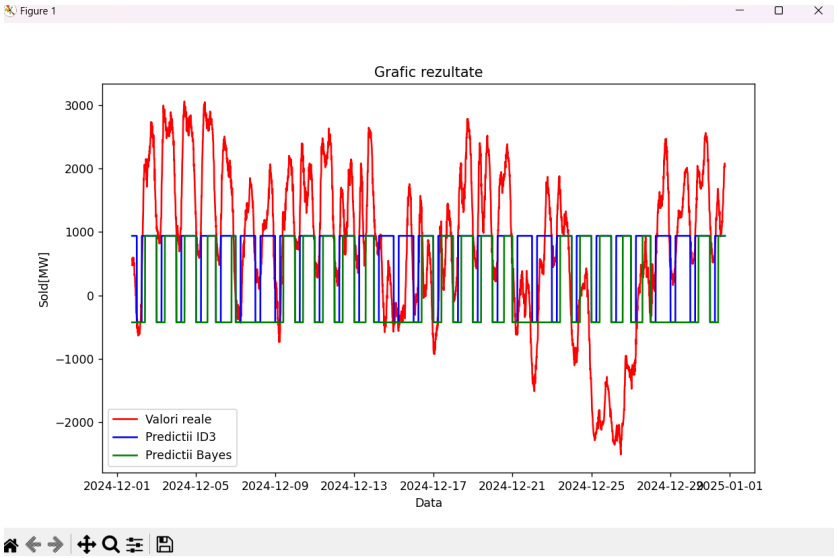

Figure 4: Grafic Sold Only ziua saptamanii inclusa + parametrii mai mici

Model	RMSE	MAE
Arbore de Decizie ID3	1129.53	929.35
Clasificare Bayesiană	1399.72	1171.49

Table 4: Tabel Sold Only ziua saptamanii inclusa + parametrii mai mici

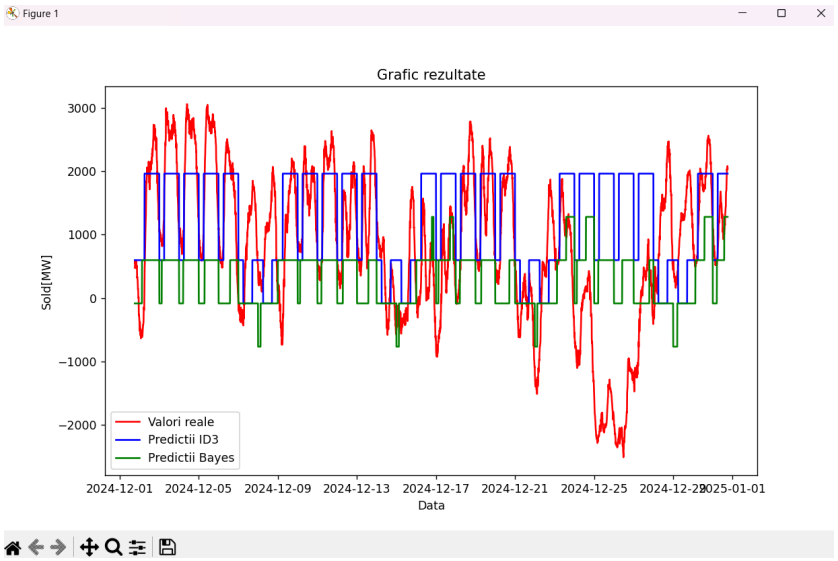

Figure 5: Grafic Sold Only parametrii normali

Model	RMSE	MAE
Arbore de Decizie ID3	1278.60	902.85
Clasificare Bayesiană	1274.18	1068.21

Table 5: Tabel Sold Only parametrii normali

Se observă că în cadrul metodei 1, cu cât adăugăm mai multe features acuratețea crește, cât timp nu deviem prea departe de la parametrii stabili (considerați prin convenție default).

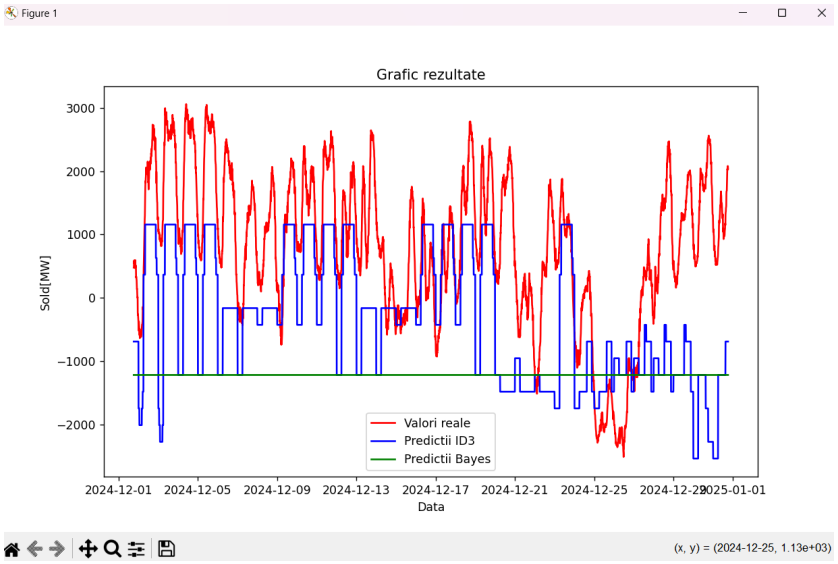

Figure 6: Grafic Sold Only doar luna decembrie ani precedenti

Model	**RMSE**	**MAE**
Arbore de Decizie ID3	1574.19	1317.72
Clasificare Bayesiană	2261.42	2027.29

Table 6: Tabel Sold Only doar luna decembrie ani precedenti

Se observă că în cazul metodei 2 acuratețea devine mult mai slabă, cel putin pentru clasificarea Bayesiana. Putem trage concluzia că dacă doar Soldul este luat în calcul, luna decembrie din ani precedenți nu va îmbunătăți performanță algoritmilor, ci o va diminua din cauza consistenței Soldului.

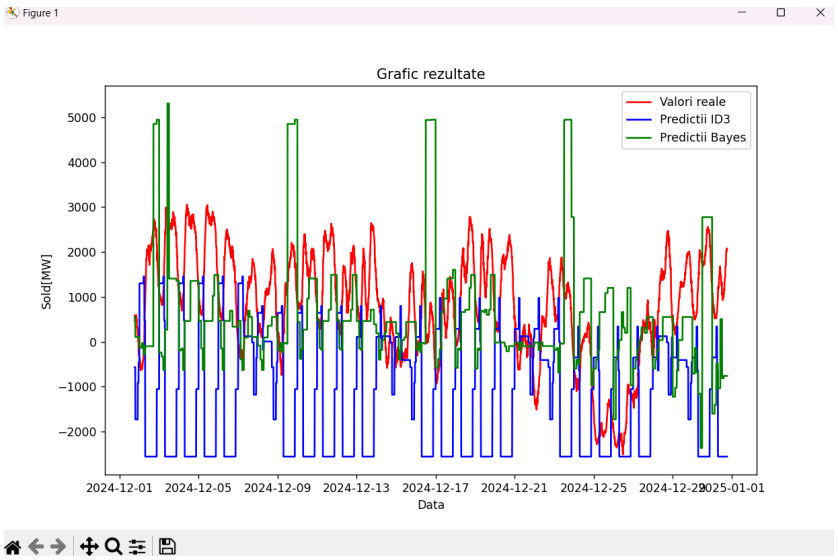

Figure 7: Grafic All Features parametrii normali

Model	RMSE	MAE
Arbore de Decizie ID3	1434.61	1076.49
Clasificare Bayesiană	1511.55	1145.23

Table 7: Tabel All Features paremtrii normali

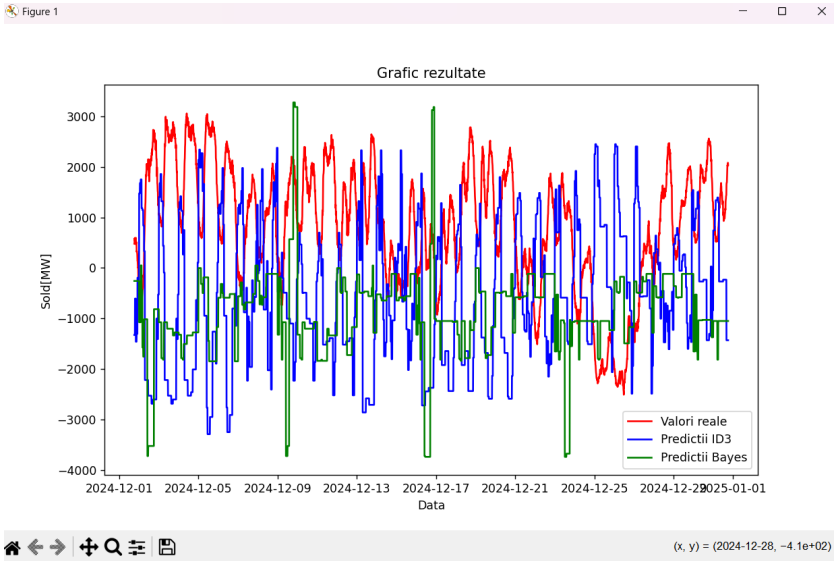

Figure 8: Grafic All Features parametrii mai mari

Model	RMSE	MAE
Arbore de Decizie ID3	2679.10	2196.47
Clasificare Bayesiană	2283.04	1894.41

Table 8: Tabel All Features paremtrii mai mari

Se observă că în cadrul metodei 3, algoritmii sunt mult mai sensibili la schimbare de parametrii ce duce la o acuratețe drastic de slabă din cauza că sunt luate în calcul mult mai multe coloane ce poate duce la zgomot și risc mai mare de erori.

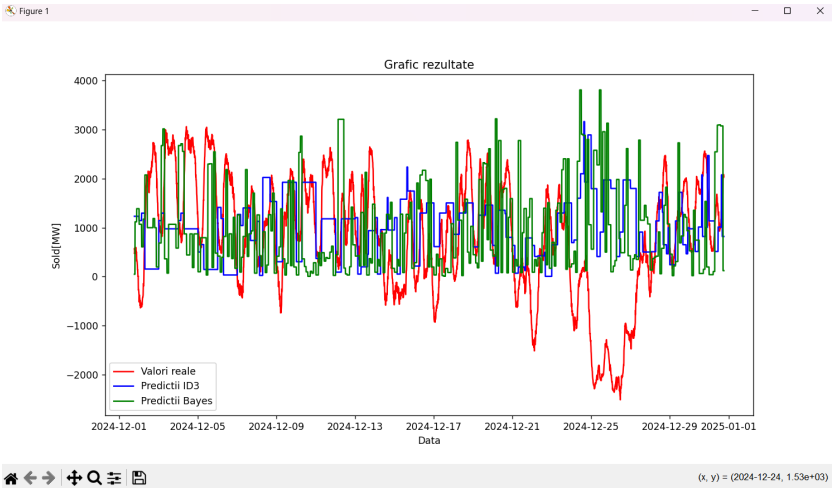

Figure 9: Grafic All Features doar luna decembrie ani precedenti

Model	RMSE	MAE
Arbore de Decizie ID3	1407.34	1047.55
Clasificare Bayesiană	1513.51	1174.51

Table 9: Tabel All Features doar luna decembrie ani precedenti

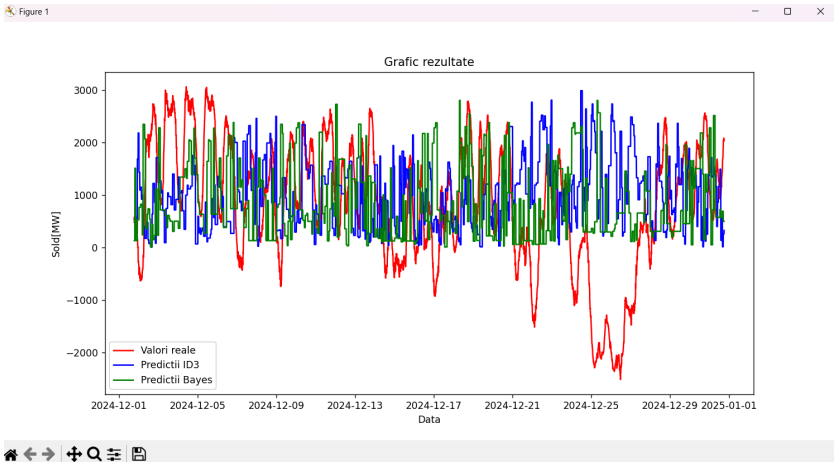

Figure 10: Grafic All Features doar luna decembrie ani precedenti + parametrii mai mari

Model	RMSE	MAE
Arbore de Decizie ID3	1476.20	1142.02
Clasificare Bayesiană	1321.70	1040.65

Table 10: Tabel All Features doar luna decembrie ani precedenti + parametrii mai mari

Se observă că în cazul metodei 4, acuratețea rămâne la fel că și în metodă 3 cu parametrii default. Schimbarea parametrilor în cadrul acestei metode pare să îmbunătățească performanță clasificării Bayesiene și să diminueze algoritmul ID3. Creșterea parametrilor în cazul metodei 4 nu pare să aibă un impact la fel de drastic că în metodă 2.

5 Concluzie

În acest proiect, am explorat adaptarea algoritmilor ID3 și clasificării Bayesiene pentru o sarcină de regresie cu scopul de a prezice soldul energetic în SEN. Rezultatele noastre au arătat că metodă 1, unde prezicem Soldul fără a folosi alte componente, duce la cea mai bună acuratețe overall. Din urmă acestor experimente, putem trage următoarele concluzii:

- cu cât crește datasetul, cu atât avem o acuratețe mai bună
- cu cât folosim mai multe componente, asumand că sunt utilizate cu grijă, vom avea rezultate mai clare și o acuratețe mai bună

- devierea semnificativă de la parametrii default rareori duce la o acuratețe mai bună, deci nu este optim să facem asta
- luând în considerare doar luna decembrie din ani precedenți duce la o acuratețe bună doar dacă datasetul nu are informații sau componente volatile care sunt sensibile la zgomot/erori
- calculând prezicerea unei clase din mai multe features devine din ce în ce mai sensibilă la schimbarea parametrilor
- când avem mai multe componente, limitarea la instanțe similare duce la o mică îmbunătățire în performanță (metodă 4 are un mic avantaj comparativ cu metodă 3 în cadrul parametrilor default)
- acuratețea devine mult mai slabă dacă avem prea multe instanțe cu preziceri calculate din mai multe componente (metodă 4 e mai puțîn afectată decât metodă 2)

Îmbunătățiri asupra performanței prezicerilor ar putea include:

- metode de a elimina outliers în mod consistent
- metode de a scade sensibilitatea la zgomot
- metode care reduc probabilitatea de a avea overfitting
- metode ce găsesc mediul cel mai optim pentru cele mai bune rezultate

Pe lângă asta, am putea adopta algoritmi care sunt specific făcuți pentru probleme de regresie.