Aplicație Practică 1

1. Descrierea Problemei

În acest proiect, scopul principal a fost implementarea și evaluarea a două tehnici de învățare automată: ID3 și clasificare Bayesiana, pentru predicția consumului de energie electrică pe baza unui set de date care include diverse surse de producție de energie. Cei doi algoritmi au trebuit sa fie adaptați pentru o problemă de regresie, ei fiind folosiți la probleme de clasificare.

Tema proiectului a fost prezicerea soldului total de energie din luna decembrie 2024, fără a folosi date din luna decembrie.

2. Justificarea Abordării

Pentru rezolvarea problemei am ales sa folosesc date din ultimii 3 ani, iar pentru partea de preprocesare a datelor luate de pe site-ul Transelectrica au fost făcute următoarele:

- împărțirea setului de date în set de antrenament (toate lunile mai puțin decembrie) si de test (luna decembrie 2024, până la data de 28);
- selectarea targetului și a atributelor;

2.1. ID3

Pentru a rezolva acest task folosind ID3 am abordat 2 metode diferite care au dus la o modificare a algoritmului, atât pentru a crește acuratețea, cât și pentru a putea prezice valori numerice continue cat mai apropiate de target.

2.1.1. Metoda 1

Am implementat o funcție pentru a calcula entropia pe baza distribuției datelor într-un set. Entropia măsoară incertitudinea în date și este mai frecvent folosită în clasificare, dar a fost adaptată aici pentru variabile continue prin utilizarea unor histograme.

Reducerea entropiei a fost utilizată pentru a decide cel mai bun punct de divizare. Divizarea maximizează reducerea entropiei totale, ceea ce ajută arborele să separe mai bine datele.

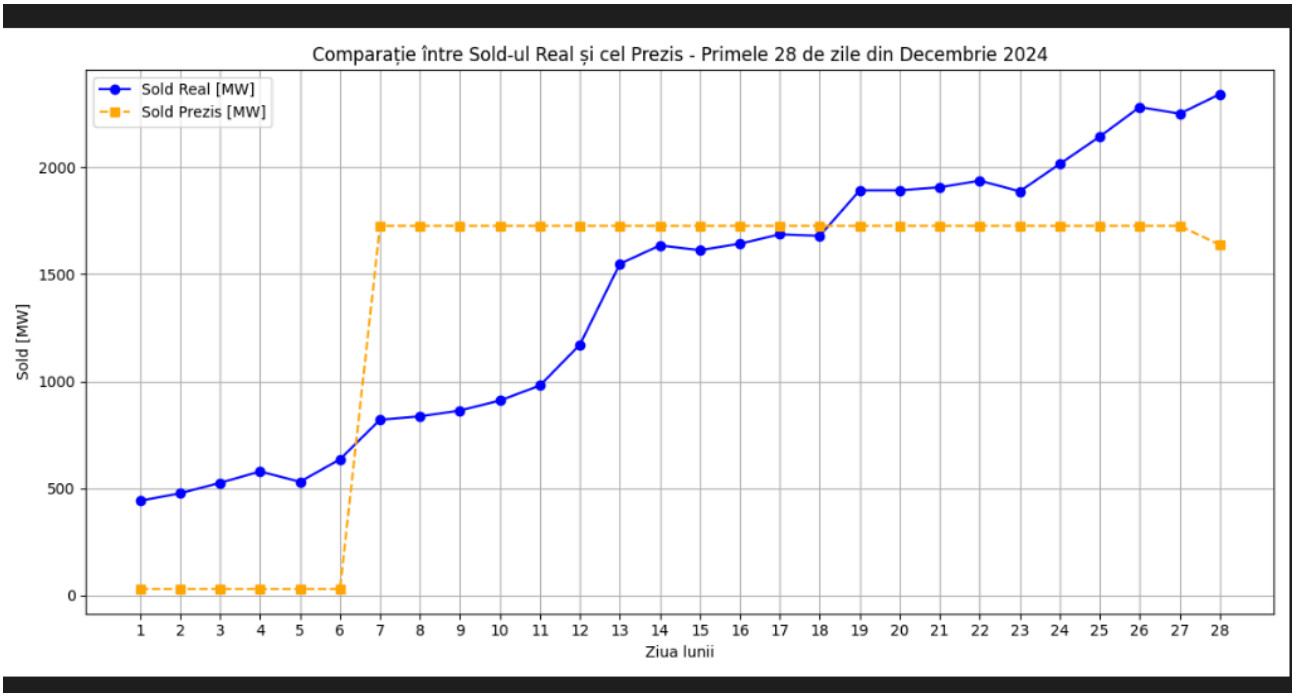
Algoritmul construiește arborele iterativ, până când sunt atinse condițiile de oprire: adâncime maximă sau un număr minim de eșantioane pentru divizare.

Arborii au fost antrenați pentru fiecare caracteristică individuală (Consum, Biomasă, etc.). Reducerea entropiei a fost utilizată pentru a identifica cea mai buna caracteristică și valoare de divizare la fiecare pas. Arborii au generat predicții pentru setul de test, iar soldul a fost calculat ca diferența dintre consumul prezis și producția prezisă. Performanța predicției a fost evaluată folosind media pătratică(MSE).

Metoda are anumite limitari în cazul variabilelor continue. Entropia poate să nu fie întotdeauna intuitivă și necesită discretizarea datelor(histograme), ceea ce poate introduce incertitudini.

Rezultatele obținute folosind entropia ca și criteriu de divizare au fost:

- Soldul total real pe luna decembrie 2024: 3028349 MW
- Soldul total prezis: 1011076 MW
- MSE: 884391.4594528578



2.1.2. Metoda 2

Metoda 2 se bazează pe înlocuirea entropiei cu principiul reducerii erorii mediei pătratice(MSE), care este un indicator al variației dintr-un set de date.

La fel ca la metoda anterioară, a fost construit un arbore de decizie pentru fiecare caracteristică. Algoritmul determină cel mai bun punct de divizare pe baza reducerii MSE la fiecare pas:

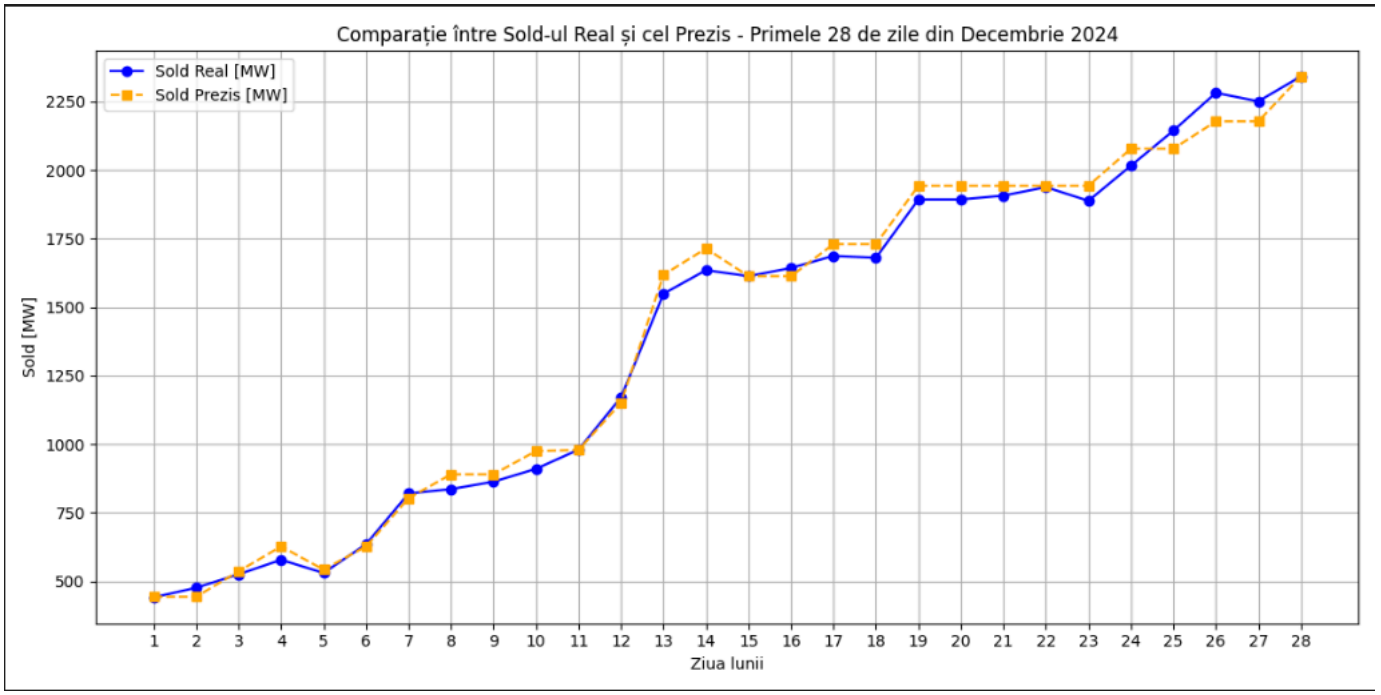
- Calculul MSE înainte de divizare: este calculată varianța target-ului pentru întregul set de date curent.
- Calculul MSE pentru divizări: datele sunt împărțite în două subseturi pe baza medianei unei caracteristici, iar erorile din fiecare subset sunt combinate pentru a evalua reducerea totala a MSE.

Soldul și performanța au fost calculate la fel ca la metoda anterioară.

Această metodă rezolvă problemele întâlnite anterior fiind ușor de interpretat, mai ales pentru variabile continue. Algoritmul este mai rapid deoarece MSE nu necesită discretizarea datelor, iar prin utilizarea MSE arborii tind să ofere predicții cu erori mai mici pentru problemele de regresie.

Rezultatele obținute folosind reducerea MSE ca și criteriu de divizare au fost:

- Soldul total real pentru luna decembrie 2024: 3028349 MW
- Soldul total prezis: 3067995 MW
- MSE: 8164.687347337567



2.2. Bayes

Pentru bayes, variabila Sold a fost transformată în variabilă categorială cu doua clase: Pozitiv(Sold>=0) si Negativ(Sold<0). Am implementat un model de Naive Bayes cu MLE astfel:

- Estimarea parametrilor: pentru fiecare clasa am estimat media și varianța caracteristicilor, iar priorul a fost calculat ca proporția fiecărei clase în setul de date.
- Calculul probabilităților: pentru fiecare observație din setul de test, am calculat probabilitățile condiționate folosind distribuția normală pentru fiecare caracteristică. Verosimilitatea a fost calculată ca produsul acestor probabilități, iar predicția finală a fost data de clasa cu probabilitate maximă.

Nu am reușit să prezic valori numerice, doar în ce clasa s-ar încadra soldul. Modelul de Bayes Naiv a avut acurateţe de 80% pe setul de date de test, însă limitarea sa la nivel de predicție m-a facut sa nu îl aleg ca algoritm final.

Acuratețea modelului Naive Bayes cu MLE: 80.2%

Matricea de confuzie:

[[1024 16]

[795 2259]], de aici observăm că Bayes a clasificat greșit 811 intrări.

3. Prezentarea Rezultatelor

Mă voi axa pe o comparare a rezultatelor celor doua metode cu ID3, fiind cele care au avut rezultate mai bune.

Mertica	Metoda Entropiei	Metoda MSE
Sold total real	3,028,349	3,028,349
Sold total prezis	1,011,076	3,067,995
Diferența	-2,017,273	39,646
MSE	884,391.46	8,164.69

3.2. Observații și interpretări

Metoda 2, care utiliza MSE, a oferit un rezultat mult mai apropiat de valoarea reală a soldului total pentru luna decembrie 2024, cu o eroare de ~1.31%.

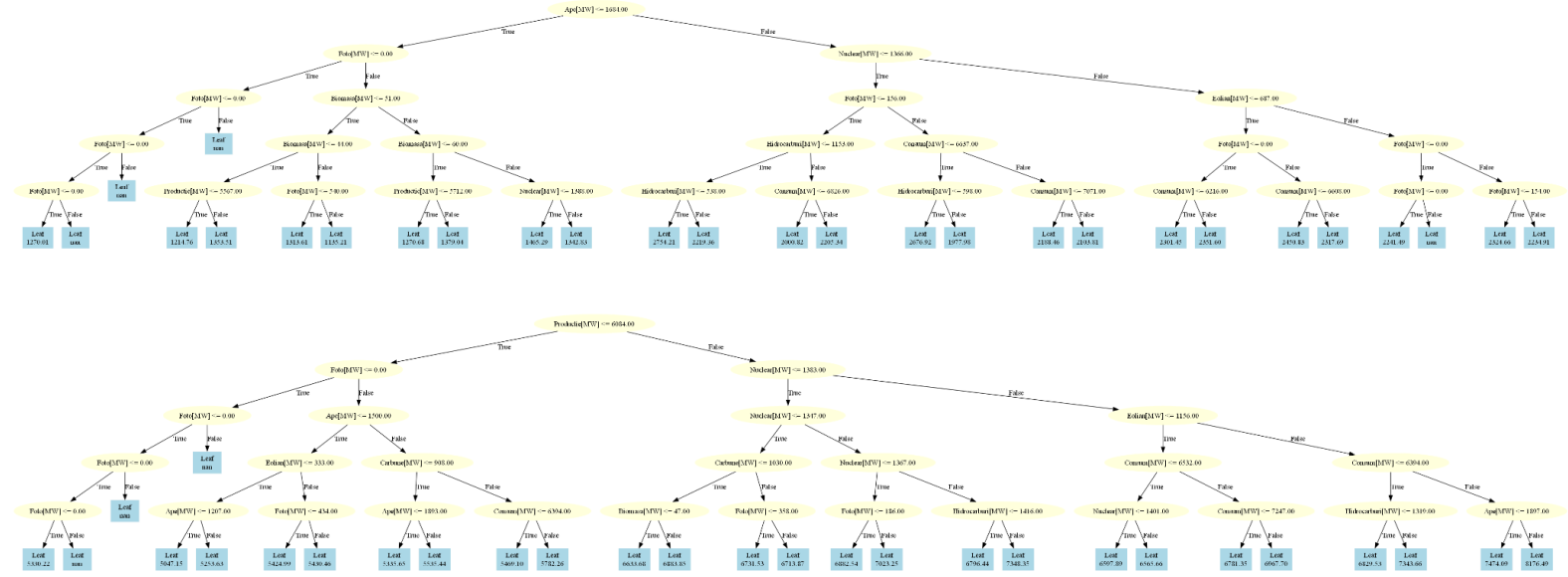
Metoda 1, care utiliza entropia, a subestimat foarte mult valoarea soldului total, generând o diferență semnificativă, cu o eroare de ~66.63%.

Metoda entropiei pare a fi mai puțin potrivită pentru acest tip de probleme, măsurând incertitudinea, aceasta poate duce la divizari care nu optimizează direct variația numerică a datelor. În schimb, MSE a fost mult mai eficientă în gestionarea variabilelor numerice continue, fiind concepută pentru a minimiza variația.

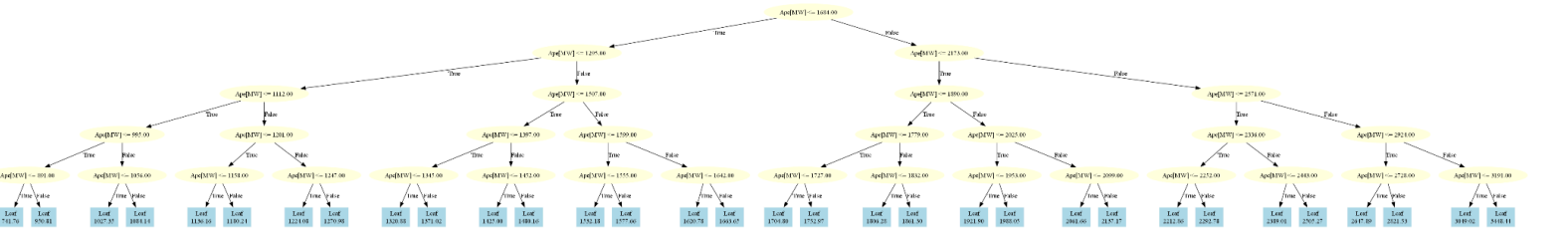
3.3. Imagini Arbori

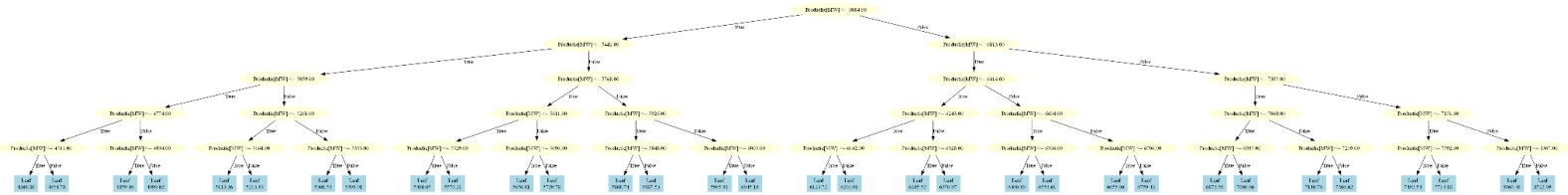
Ilustrație a arborilor pentru Ape și Producție cu ambele metode.

3.3.1. ID3 cu entropie



3.3.2. ID3 cu MSE





4. Concluzii

4.1. Ce am învățat

Pe parcursul rezolvării acestui proiect am avut șansa de a experimenta lucrul cu algoritmii învățați la curs pe o problemă reală, cu un set de date foarte mare, lucru care mi-a arătat în mod practic care sunt limitările celor doi algoritmi, dar și punctele forte. Am învățat că o simplă modificare la o regulă de decizie poate influența drastic comportamentul și puterea unui model atât de simplu cum este ID3.

4.2. Îmbunătățiri

O posibilă îmbunătățire ar fi folosirea unui model mai complex, cum ar fi Gradient Boosting sau alte tehnici de ensemble, pentru a îmbunătăți performanța. De asemenea, testarea pe un set mai mare de date sau utilizarea unor funcții de regularizare ar putea stabiliza și mai mult predicția.