

# Raport: Aplicatie Practica 1

Lacatus Stefania

January 4, 2025

## Descrierea Problemei

Scopul acestei aplicatii este sa prezica "soldul total" din Sistemul Energetic National (SEN) pentru luna decembrie 2024. Soldul total reprezinta diferenta dintre productia totala si consumul total de energie electrica. Pentru a rezolva aceasta problema, am implementat un program care foloseste algoritmi ID3 (arbore de decizie) si Bayes Naiv, adaptati pentru aceasta problema de regresie.

## Justificarea Abordarii

Pentru dezvoltarea unei solutii utilizand algoritmul ID3 am folosit 2 abordari diferite:

### ID3 cu DecisionTreeRegressor

In aceasta abordare, am folosit o interpretare a unui arbore de decizie, care simula comportamentul algoritmului ID3. In aceasta abordare, am utilizat un arbore de decizie care este potrivit pentru probleme de regresie, folosind `DecisionTreeRegressor` din biblioteca scikit-learn.

In cadrul acestei implementari, arborele de decizie a fost antrenat pe attributele `Day`, `Month` (extrase din coloana 'Data'), `Consum[MW]` si `Productie[MW]`, iar in loc sa aplic bucketing pentru `Sold[MW]`, am utilizat valorile continue. Acest model functioneaza prin impartirea repetata a datelor in subseturi mai mici pe baza celor mai bune attribute, astfel incat sa reduca erorile de predictie.

## ID3 cu bucketing

În această abordare, am adaptat algoritmul ID3 la problema de regresie data prin discretizarea valorilor coloanei `Sold[MW]` în intervale, prin metoda bucketing, folosind *KBinsDiscretizer*. Practic, fiecare valoare a `Sold[MW]` a fost asociată unui interval, iar predicția finală a algoritmului returnează mijlocul intervalului respectiv. Structura arborelui de decizie a fost construită folosind câștigul de informație, iar pentru frunze am folosit valoarea medie a `Sold[MW]` pentru a face predicția. Am folosit toate atributele din setul de date pentru antrenarea modelului.

## Bayes Naiv

Pentru modelul Bayes Naiv, am folosit aceleași atribute ca și în cazul ID3: `Day`, `Month`, `Consum[MW]` și `Productie[MW]` pentru a prezice valoarea discretizată a `Sold[MW]`. Am folosit algoritmul `GaussianNB()` din biblioteca `scikit-learn`, care presupune că valorile atributelor urmează o distribuție normală (Gaussiană) și calculează probabilitățile pe baza acestei presupuneri. La final, predicțiile au fost transformate în valori continue, prin alegerea mijlocului intervalului corespunzător.

## Prezentarea Rezultatelor

Pentru a evalua performanța acestor modele am folosit următoarele metrice:

- Eroarea Medie Absolută (MAE)
- Eroarea Medie Patrată (MSE)
- Radacina Eroarei Medii Patratice (RMSE)

### Rezultatele obținute:

**Abordarea ID3 cu `DecisionTreeRegressor`:**

- **Soldul Total Real:** 3427875 MW
- **Predicția ID3:** Soldul Total= 3437265 MW, Acuratete= 99.73%  
MAE: 170.33, MSE: 39200.59, RMSE: 197.99
- **Predicția Bayes:** Soldul Total= 2706843 MW, Acuratete= 78.97%  
MAE: 535.11, MSE: 497055.12, RMSE: 705.02

```

Algoritmul ID3 - MAE: 170.33, MSE: 39200.59, RMSE: 197.99, Acuratețe: 99.73%
Algoritmul Bayes Naiv - MAE: 535.11, MSE: 497055.12, RMSE: 705.02, Acuratețe: 78.97%
Sold total real in decembrie 2024: 3427875.00 MW
Sold total prezis de ID3: 3437265.00 MW
Sold total prezis de Bayes Naiv: 2706843.00 MW

```

Figure 1: Terminalul cu rezultatele programului

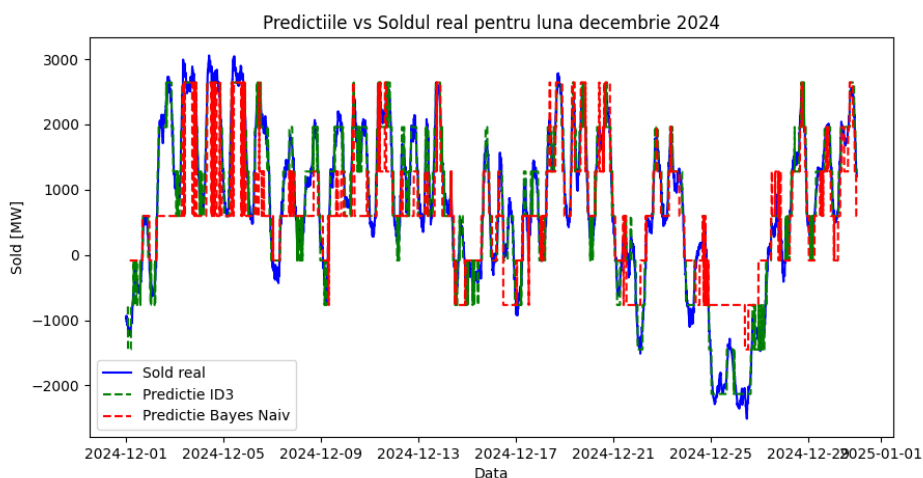


Figure 2: Graficul de comparatie

#### Abordarea ID3 cu bucketing:

- **Soldul Total Real:** 3427875 MW
- **Predictia ID3:** Soldul Total= 415465.99 MW, Acuratete= 12.12%  
MAE: 1119.26, MSE: 1905622.41, RMSE: 1380.44
- **Predictia Bayes:** Soldul Total= 2706843 MW, Acuratete= 78.97%  
MAE: 535.11, MSE: 497055.12, RMSE: 705.02

```

Algoritmul ID3 - MAE: 1119.26, MSE: 1905622.41, RMSE: 1380.44, Acuratete: 12.12%
Algoritmul Bayes Naiv - MAE: 535.11, MSE: 497055.12, RMSE: 705.02, Acuratete: 78.97%
Sold total real in decembrie 2024: 3427875.00 MW
Sold total prezis de ID3: 415465.99 MW
Sold total prezis de Bayes Naiv: 2706843.00 MW

```

Figure 3: Terminalul cu rezultatele programului

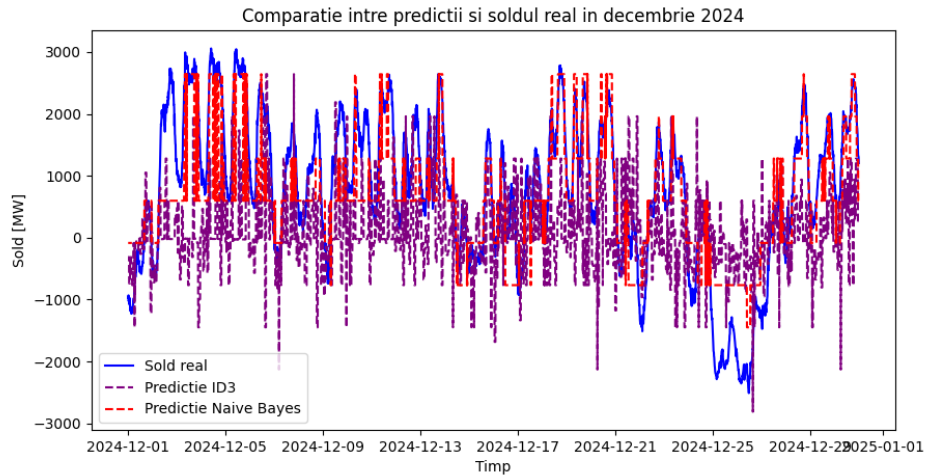


Figure 4: Graficul de comparatie

## Analiza Comparativa

Comparand cele doua abordari, putem observa ca abordarea care foloseste `DecisionTreeRegressor` din biblioteca `scikit-learn` a dat cele mai bune rezultate. Acuratetea predictiilor obtinute cu acest model a fost mult mai mare (99.73%) comparativ cu abordarea in care am folosit ID3 cu bucketing (12.12%). Modelul ID3 cu bucketing a avut rezultate destul de slabe, probabil datorate modului in care am discretizat datele si am folosit intervale largi pentru valorile din `Sold[MW]`.

Algoritmul Bayes Naiv a avut o performanta mai slaba in comparatie cu ID3 cu `DecisionTreeRegressor`, avand o acuratete de 78.97%.

## Concluzii

In concluzie, algoritmi ID3 si Bayes Naiv pot fi adaptati pentru probleme de regresie, desi sunt de obicei utilizati pentru clasificare. Discretizarea valorilor si inversarea intervalelor au fost esentiale pentru a obtine aceste rezultate.

## Imbunatatiri pentru rezultate mai bune:

- Alte tehnici de preprocesare a datelor, cum ar fi normalizarea.

- Folosirea unor modele mai complexe de regresie, cum ar fi Random Forest.
- Imbunatatirea evaluarii modelului prin folosirea cross validarii.