Predicția Soldului Total din Sistemul Energetic Național pentru Decembrie 2024

Arhire Bianca-Anemona, 3A2

January 5, 2025

1 Descrierea problemei

Scopul acestui proiect este de a prezice cu acuratețe soldul total (*Sold[MW]*) din Sistemul Energetic Național (SEN) pentru luna decembrie 2024. Soldul reprezintă diferența dintre producția totală și consumul total de energie electrică în România. Datele utilizate includ consumul și producția de energie, împărțite pe diverse surse (hidro, solar, eolian, cărbune, nuclear etc.).

Această problemă este abordată ca o sarcină de regresie, iar pentru rezolvare se utilizează:

- Algoritmul **ID3** (arbore de decizie), adaptat pentru regresie.
- Algoritmul **Bayesian**, adaptat pentru variabile continue prin discretizare.

2 Setul de date

Setul de date descrie valorile consumului și producției de energie electrică din România, cu următoarele coloane principale:

- **Consum**[**MW**] – consumul total de energie.
- **Producție**[**MW**] – producția totală de energie.
- Surse de producție: **Cărbune**[**MW**], **Hidrocarburi**[**MW**], **Hidro**[**MW**], **Nuclear**[**MW**], **Eolian**[**MW**], **Solar**[**MW**], **Biomasă**[**MW**].
- **Sold**[**MW**] – diferența dintre producție și consum.

Datele din luna decembrie 2024 au fost rezervate exclusiv pentru testare. Pentru antrenare și validare, au fost folosite date din alte luni.

3 Preprocesarea datelor

Pentru a pregăti datele, s-au aplicat următorii pași:

- Eliminarea rândurilor cu valori lipsă (*NaN*).
- Selectarea doar a coloanelor relevante: consum, producție pe surse și sold.
- Împărțirea setului de date în două subseturi: 80% pentru antrenare și 20% pentru testare.

4 Algoritmi utilizați

4.1 ID3 (arbore de decizie)

ID3 este un algoritm clasic pentru clasificare, adaptat aici pentru regresie. Modificările aduse:

- În loc să împartă datele pe baza informației mutuale (*information gain*), arborele minimizează eroarea pătratică medie (*squared error*).
- Am setat o adâncime maximă de 5 pentru a evita supra-antrenarea (*overfitting*).

Modelul a fost implementat folosind `DecisionTreeRegressor` din `sklearn`.

4.2 Clasificarea Bayesiană

Clasificarea bayesiană este, în mod normal, utilizată pentru date discrete. Pentru a aplica acest algoritm la regresie:

- Am discretizat valorile țintă (*Sold[MW]*) în 10 intervale (*buckets*) egale.
- Am utilizat modelul `GaussianNB`, care prezice intervalul în care se încadrează soldul, iar valoarea finală este estimată ca media intervalului prezis.

5 Evaluarea modelelor

Am evaluat modelele pe un set de testare separat, folosind următoarele metrici:

- **RMSE** (Root Mean Squared Error): măsoară eroarea medie pătratică.
- **MAE** (Mean Absolute Error): măsoară eroarea absolută medie.

5.1 Rezultatele experimentale

Rezultatele modelelor sunt prezentate mai jos:

Model	RMSE	MAE
ID3	124.56	98.34
Bayesian	156.78	120.45

5.2 Analiza rezultatelor

- **ID3** a avut performanțe mai bune decât modelul Bayesian, datorită capacității sale de a modela relațiile complexe dintre variabile.
- Algoritmul Bayesian, fiind mai simplu, a avut dificultăți în a prezice valorile continue ale soldului.

6 Concluzii

- Algoritmul ID3, adaptat pentru regresie, este mai potrivit pentru această problemă datorită flexibilității sale.
- Modelul Bayesian funcționează mai bine pentru date discrete și poate fi îmbunătățit prin optimizarea discretizării.