Predicția Soldului Total în Sistemul Energetic Național (SEN) pentru Decembrie 2024

Motelică Sandu, grupa A2

December 21, 2024

1 Descrierea Problemei

Scopul acestui proiect este de a prezice soldul total al Sistemului Energetic Național (SEN) din România pentru luna decembrie 2024. Datasetul utilizat descrie consumul și producția de energie electrică, segmentate în funcție de surse precum hidro, solar, eolian, cărbune, și altele.

Principala provocare constă în adaptarea algoritmilor ID3 și clasificarea bayesiană, proiectați pentru probleme de clasificare, pentru a rezolva o problemă de regresie. De asemenea, datele din decembrie nu pot fi utilizate pentru antrenare, fiind destinate exclusiv testării.

2 Justificarea Abordării

Pentru a rezolva problema regresiei folosind algoritmii menționați, au fost realizate următoarele adaptări:

2.1 Adaptarea Algoritmului ID3

Algoritmul ID3 a fost modificat pentru a se potrivi problemei de regresie prin:

- **Bucketing pentru variabila țintă**: Valorile soldului (Sold[MW]) au fost împărțite în intervale discrete, utilizând binning uniform. Fiecărui interval i s-a atribuit o valoare reprezentativă, corespunzând mijlocului intervalului.
- **Predicția componentelor individuale**: În loc să se prezică direct soldul, algoritmul a fost utilizat pentru a prezice separat Producția[MW] și Consumul[MW]. Soldul a fost calculat ca diferența între aceste două valori.

2.2 Adaptarea Clasificării Bayesiene

Clasificarea bayesiană a fost adaptată astfel:

- Toate variabilele continue au fost discretizate utilizând binning uniform.
- Probabilitățile condiționate au fost calculate pentru fiecare combinație de caracteristici, iar predicția finală a fost realizată prin media valorilor țintă corespunzătoare.

2.3 Abordări Multiple de Predicție

Au fost testate mai multe metode:

- Predicția directă a soldului utilizând ID3.
- Predicția componentelor (producție și consum) cu calculul soldului ulterior.
- Predicția soldului utilizând bucătirea valorilor (bucket prediction).
- Regresia bayesiană pe baza variabilelor discretizate.

3 Prezentarea Rezultatelor

Rezultatele obținute pentru fiecare abordare au fost evaluate utilizând metrici standard: Eroare Medie Absolută (MAE), Eroare Pătratică Medie (RMSE) și coeficientul R-squared (R^2). În continuare, este prezentată o analiză detaliată pentru fiecare set de caracteristici utilizat.

3.1 Performanța Abordărilor pe Seturi de Caracteristici

Au fost analizate patru seturi de caracteristici principale, descrise în Tabelul 1, împreună cu rezultatele obținute.

3.2 Analiza Performanțelor

Setul de Caracteristici 1: [Day_of_Week, Hour, Consum[MW], Intermittent_Production, Constant_Production]

Predicția directă a oferit cele mai bune rezultate (MAE=249.31, RMSE=309.36, R^2=0.8832), cu producția intermitentă și consumul având cea mai mare importanță. Modelul bazat pe componente a avut performanțe extrem de slabe, indicând o inadecvare a acestei abordări pentru acest set de caracteristici.

Setul de Caracteristici 2: [Day_of_Week, Hour, Consum[MW], Productie[MW]]

Acest set a produs cele mai bune rezultate generale, cu predicția directă oferind MAE=93.19 și R^2=0.9835. Consum și producție sunt principalele caracteristici predictive. Performanțele abordării bazate pe componente rămân scăzute.

Setul de Caracteristici 3: [Consum[MW], Productie[MW]]

Simplificarea setului la două caracteristici principale nu a afectat negativ predicția

directă, obținând aceleași rezultate ca setul anterior. Abordările alternative (componente și bucket) au avut rezultate moderate.

Setul de Caracteristici 4: [Year, Day_of_Week, Hour]
Performanțele generale au fost cele mai slabe, toate modelele având valori R^2 scăzute sau negative. Acest set este inadecvat pentru predicția soldului total.

Set de Caracteristici	**Metodă**	**MAE**	**RMSE**	**R^2**
['Day_of_Week', 'Hour', 'Consum[MW]', 'Intermittent_ Production', 'Constant_ Production']	Bayesian	412.75	535.67	0.6498
	Direct Prediction	**249.31**	**309.36**	**0.8832**
	Component Prediction	1450.25	1739.57	-2.6928
	Bucket Prediction	403.30	502.50	0.6919
['Day_of_Week', 'Hour', 'Consum[MW]', 'Productie[MW]']	Bayesian	402.83	497.89	0.6975
	Direct Prediction	**93.19**	**116.23**	**0.9835**
	Component Prediction	888.89	1074.40	-0.4087
	Bucket Prediction	371.59	446.67	0.7565
['Consum[MW]', 'Productie[MW]']	Bayesian	397.72	500.47	0.6943
	Direct Prediction	**93.19**	**116.23**	**0.9835**
	Component Prediction	374.31	458.32	0.7437
	Bucket Prediction	371.44	446.43	0.7568
['Year', 'Day_of_Week', 'Hour']	Bayesian	702.40	841.66	0.1355
	Direct Prediction	699.93	842.03	0.1348
	Component Prediction	1232.24	1441.95	-1.5373
	Bucket Prediction	746.04	898.25	0.0154

Table 1: Rezultatele obținute pentru fiecare set de caracteristici.

4 Concluzii

Rezultatele acestui proiect evidențiază că metoda de predicție directă este cea mai eficientă abordare pentru problema predicției soldului energetic. Dintre toate seturile de caracteristici analizate, setul compus din ['Day_of_Week', 'Hour', 'Consum[MW]', 'Productie[MW]'] s-a remarcat prin performanțe

excepționale, atingând un coeficient R^2 de 0.9835 și valori scăzute pentru MAE și RMSE. Acest lucru demonstrează că includerea consumului și producției ca variabile esențiale contribuie decisiv la acuratețea predicțiilor, iar variabile precum ziua săptămânii și ora adaugă un context valoros pentru model.

Simplificarea caracteristicilor la un set minimal, precum ['Consum[MW]', 'Productie[MW]'], nu a compromis performanța modelelor directe, ceea ce sugerează că redundanța variabilelor poate fi redusă fără a afecta rezultatele. Totuși, abordările bazate pe componente, care separă predicția în estimări individuale pentru producție și consum, au arătat limitări evidente. Aceste metode au generat valori inconsistente și adesea negative ale coeficientului R^2, indicând o necesitate clară de optimizare structurală.

În ceea ce privește algoritmii utilizați, adaptarea ID3 și Bayes la probleme de regresie a fost o soluție viabilă, mai ales prin integrarea discretizării variabilelor

continue. Această flexibilitate a permis modelelor să învețe relațiile dintre variabile și să genereze predicții relevante. Totuși, succesul acestor metode a fost strâns legat de calitatea preprocesării datelor, ceea ce a evidențiat importanța discretizării și selecției atente a caracteristicilor.

Lecțiile învățate din acest proiect subliniază rolul esențial al preprocesării și calibrării modelelor în succesul predicțiilor. De exemplu, discretizarea variabilelor și ajustarea parametrilor modelelor, cum ar fi adâncimea arborilor de decizie, au contribuit semnificativ la îmbunătățirea performanței.

Privind spre viitor, îmbunătățirea metodei poate include explorarea unor tehnici avansate de discretizare, adaptate mai bine distribuției datelor. De asemenea, utilizarea metodelor de ansamblu, precum bagging sau boosting, poate crește robustețea și reduce erorile modelelor. Ajustarea mai fină a parametrilor, cum ar fi numărul de intervale pentru discretizare și adâncimea arborilor, ar putea aduce un plus de precizie și stabilitate.

În concluzie, proiectul a demonstrat că predicția directă, bazată pe variabile relevante și suportată de algoritmi bine adaptați, reprezintă o soluție promițătoare pentru predicția soldului energetic. Aceste rezultate oferă o bază solidă pentru explorări ulterioare și deschid calea către metode mai avansate și robuste, capabile să răspundă cerințelor practice ale domeniului energetic.