Aplicatie Practica 1

Iacobuț Denisa-Delia

December 2024

1 Analiza problemei si descrierea problemei:

Scopul acestui proiect este de a dezvolta și compara mai multe modele de regresie pentru a prezice soldul total de energie consumată, pe baza unui set de date care conține mai multe caracteristici relevante, cum ar fi consumul de energie din diferite surse. Datele sunt extrase din surse reale și vor fi folosite pentru a evalua performanța modelelor de predicție. Proiectul se concentrează pe utilizarea algoritmilor de regresie Naive Bayes și ID3, comparându-le pentru a evalua care model are cea mai bună performanță în termeni de precizie și erori de predicție.

2 Justificarea Abordării

Pentru a aborda problema predicției soldului total de energie, am ales două metode de regresie populare: **Naive Bayes** și **ID3**.

În cadrul acestei probleme, pentru predicția "soldului total" din Sistemul Energetic Național (SEN) pentru luna decembrie 2024, am observat mai multe legături intre valorile atributelor și corelațiile dintre acestea.

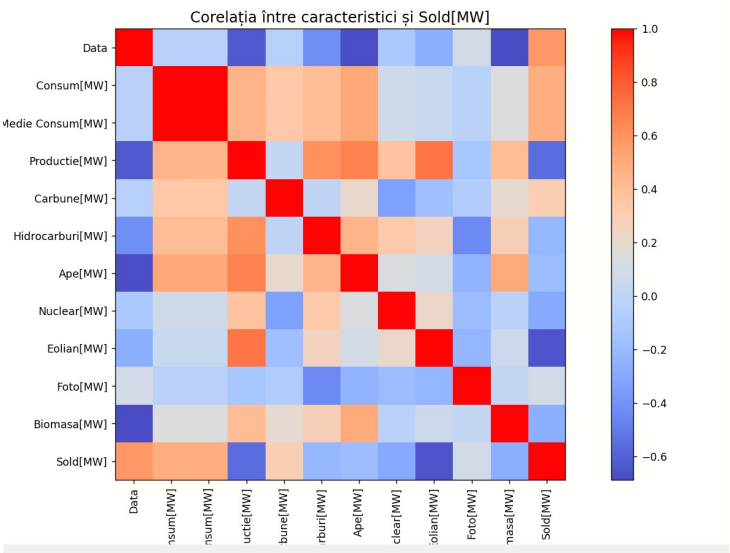

Figure 1: Corelațiile dintre atribute pentru datasetul nemodificat

În această diagramă, se poate vedea că atributele Consum și Consum Mediu sunt foarte corelate între ele, așa că le-am unit într-un singur atribut, păstrând doar valoarea

maximă dintre ele. Astfel, am obținut un nou dataset care generează erori mai puține și are o corelație mai mică între atribute, așa cum se poate observa mai jos.

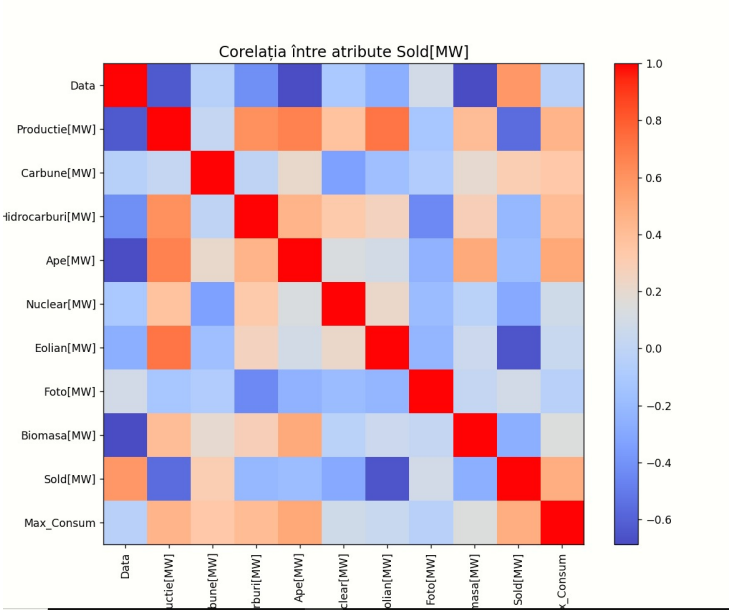

Figure 2: Corelațiile dintre atribute pentru dataset-ul modificat

Într-adevăr, mai sunt atribute ce au o corelație medie, însă dacă le-am uni și pe acestea, s-ar pierde prea multe informații.

De asemenea, am făcut comparația valorilor între datasetul corespunzător întregului an (din ianuarie până în noiembrie) și al datasetului ce conține un mixt creat pe baza caracteristicilor sezoniere ale producției și consumului de energie (luna ianuarie, iulie și noiembrie). Am observat că numărul erorilor a scăzut considerabil pentru datasetul corespunzător doar lunilor sezoniere. Valorile pentru RMSE și MAE pentru datasetul nemodificat sunt:

```
ID3 RMSE: 401.73870813293127, MAE: 321.6570858149823
Naive Bayes RMSE: 655.0689518781696, MAE: 465.4366015884637
```

Figure 3: RMSE și MAE pentru lunile sezoniere

Iar pentru datasetul modificat sunt:

```
Bayes RMSE: 711.3560178321121, MAE: 524.0463576158941
ID3 RMSE: 199.3981167446919, MAE: 153.84669692332926
Naive Bayes RMSE: 813.7850431834515, MAE: 531.3005186908752
Naive Bayes (uniform bins) RMSE: 892.0250763007866, MAE: 680.4571692033138
Naive Bayes (KMeans bins) RMSE: 808.5476576496166, MAE: 518.560499245248
```

Figure 4: RMSE și MAE pentru toate lunile

De asemenea, am abordat problema și în funcție de intervalul orar crezând că în cursul orelor de lucru crește consumul de energie însă nu a fost de ajutor. Pentru informațiile din cursul serii (după ora 16) corelația dintre atribute a crescut foarte mult crescând în

același timp și eroarea. Așadar am rămas la a analiza datasetul corespunzător lunilor sezoniere dar nu în funcție de intervalul orar.

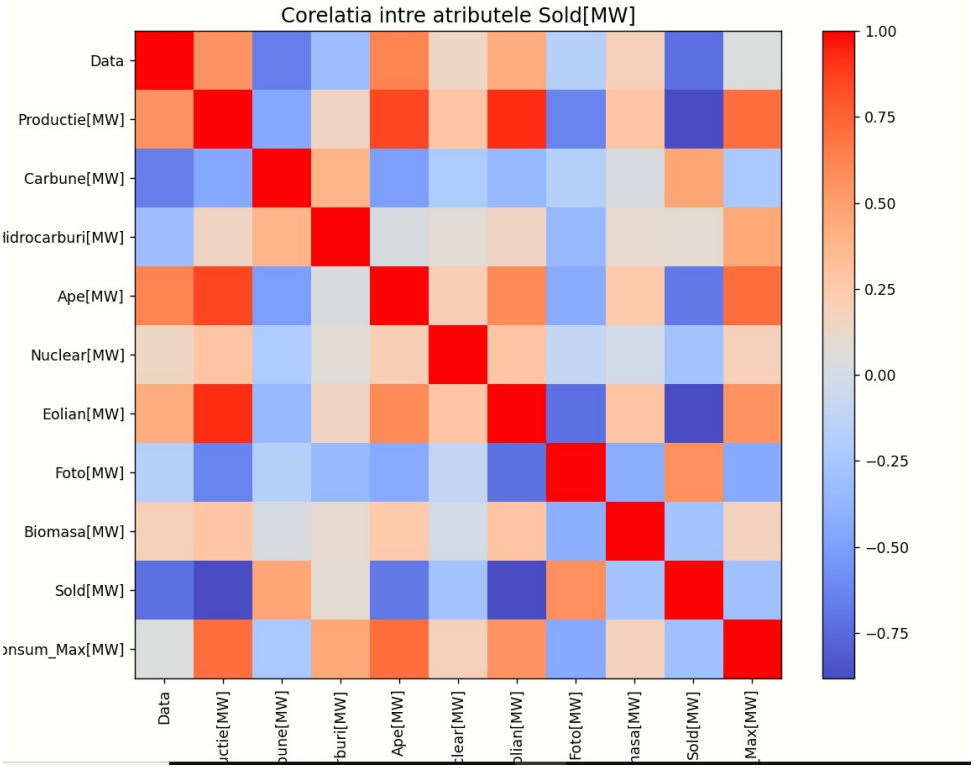

Figure 5: Grafic pentru intervalul orar 24-7

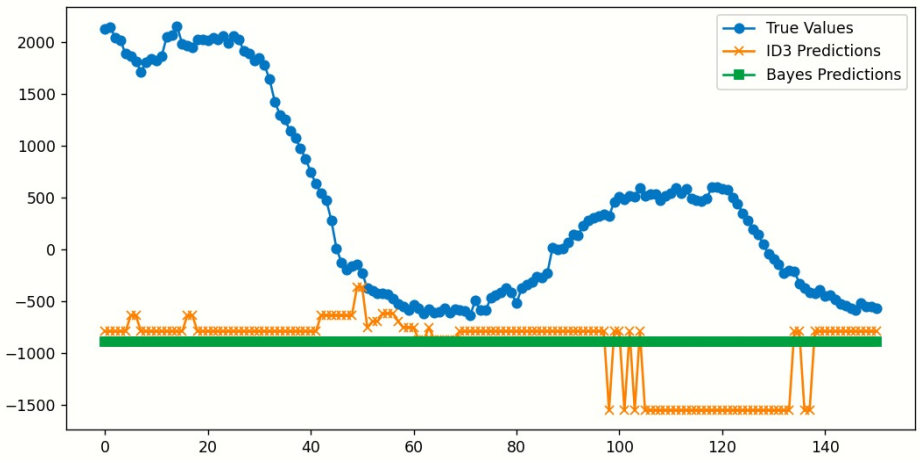

Figure 6: Precizie pentru intervalul orar 24-7

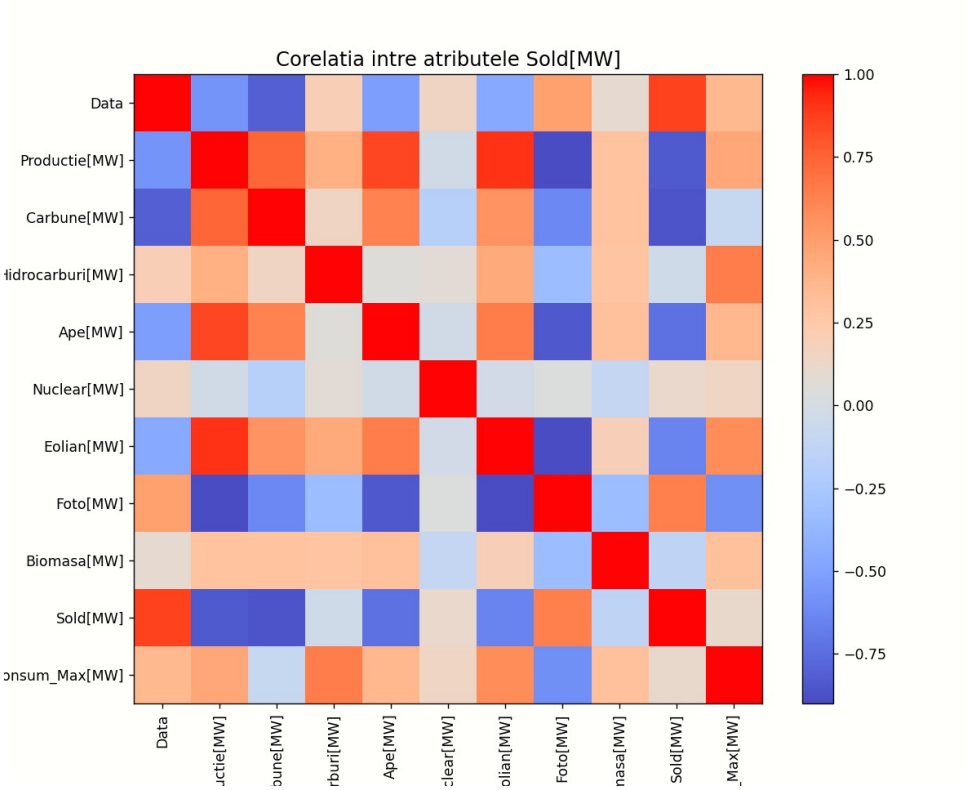

Figure 7: Grafic pentru intervalul orar 7-16

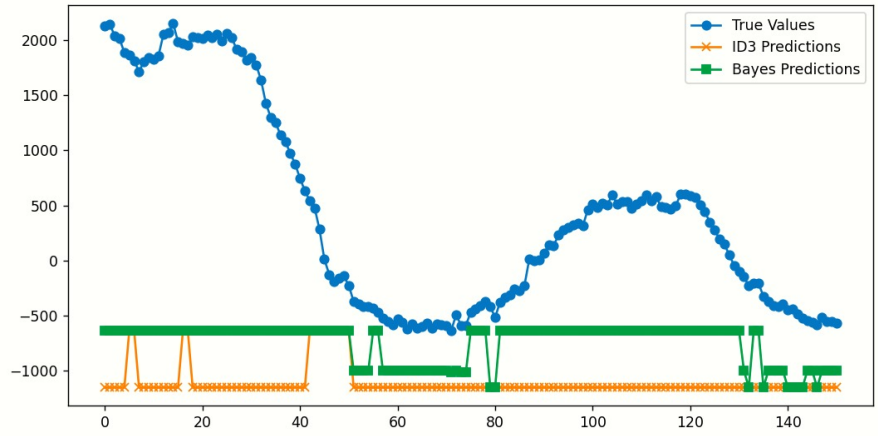

Figure 8: Precizie pentru intervalul orar 7-16

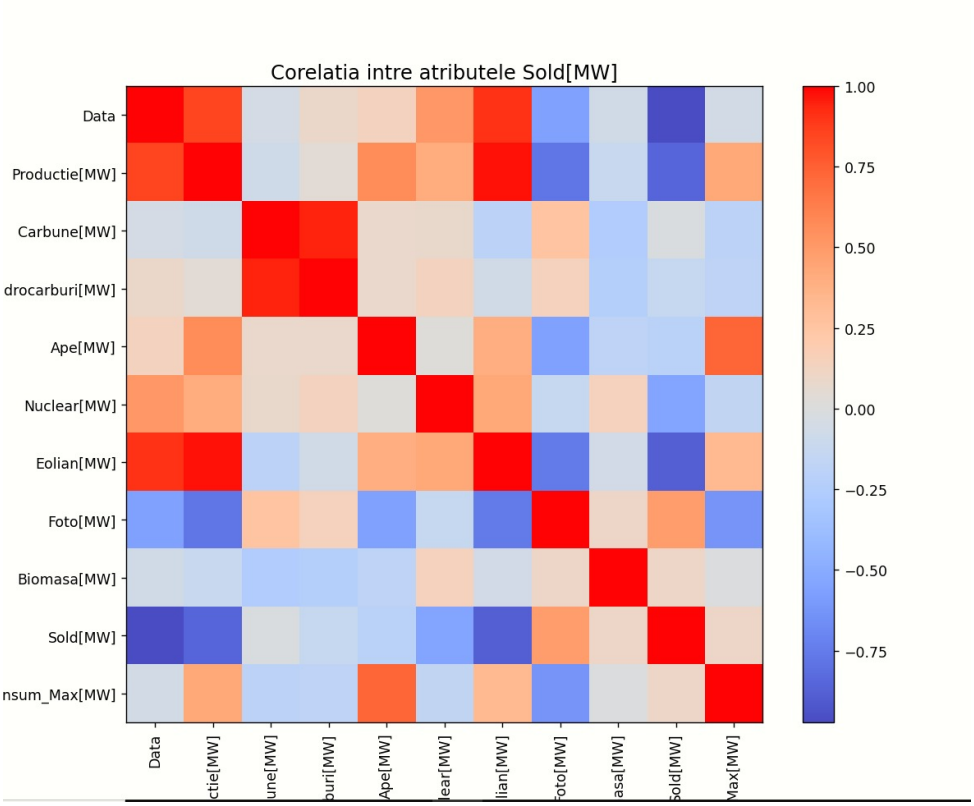

Figure 9: Grafic pentru intervalul orar 16-23

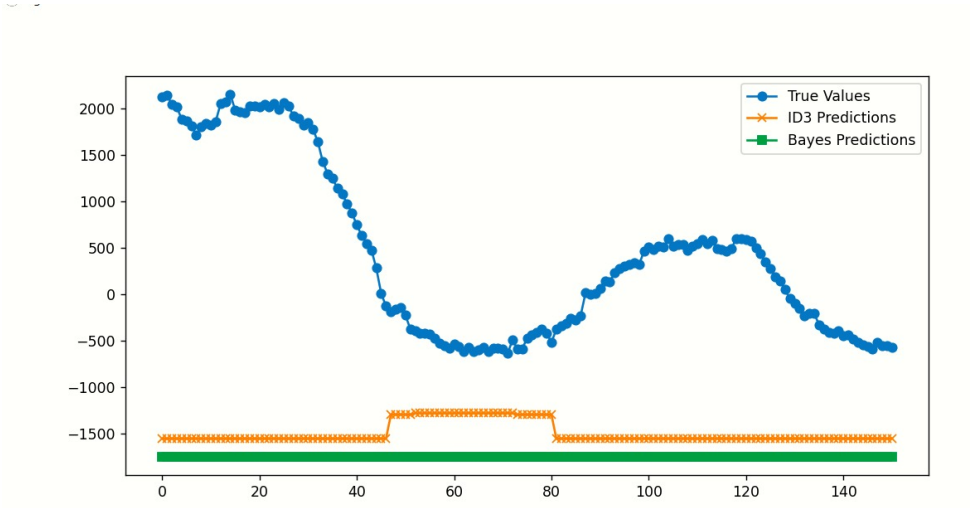

Figure 10: Precizie pentru intervalul orar 16-23

Deși am combinat deja atributele ce aveau o corelație mare, se poate observa că acum s-au creat altele, ceea ce influențează negativ dataset-ul.

3 Implementarea algoritmilor:

În ceea ce privește adaptarea pentru regresie, am folosit mai multe metode pentru a putea face comparația dintre ele. Aceste metode sunt:

* **Naive Bayes**: Am folosit o variantă de Naive Bayes pentru a prezice soldul total. Menționez că este o variantă, deoarece am discretizat intervalele pentru variabilele continue folosind binuri, atât uniform, cât și prin metode de KMeans. Această metodă poate fi eficientă în situațiile unde distribuțiile datelor sunt aproximativ normale și nu există corelații puternice între caracteristici. Însă, cum în datasetul inițial existau astfel de atribute, după combinarea acestora, Bayes Naiv a început să genereze erori mai putine.

* **Algoritmul ID3** : A fost utilizat pentru a construi un arbore de decizie care prezice soldul total pe baza variabilelor de intrare. Algoritmul ID3 împarte datele în funcție de cele mai bune caracteristici, reducând astfel complexitatea și îmbunătățind performanța predicțiilor. Am limitat adâncimea arborelui de decizie pentru a preveni overfitting-ul și am utilizat varianta de ID3 pentru regresie, care optimizează varianta în fiecare nod al arborelui.

* **Gaussian Naive Bayes**: Am folosit Gaussian Naive Bayes pentru a prezice soldul total, iar caracteristicile (precum consumul și producția de energie) au fost tratate ca fiind independente.

* **KMeans pentru Bins (Discretizare prin KMeans)**: KMeans a fost folosit pentru a crea binuri pe baza valorilor țintei și pentru a antrena modelul Naive Bayes pe aceste binuri, îmbunătățind astfel adaptabilitatea modelului.

În contextul evaluării performanței modelelor, am utilizat două metrici comune: RMSE (Root Mean Squared Error) și MAE (Mean Absolute Error). RMSE reflectă cât de departe sunt predicțiile soldului total de valorile reale. Un RMSE mic indică faptul că modelul este capabil să prezică cu acuratețe valorile țintă, iar MAE va indica, în mod similar cu RMSE, cât de precise sunt predicțiile modelului, dar fiind mai puțin sensibil la erorile extreme.

Pentru algoritmul Bayes Naiv folosind binuri, am generat folosind valorile 3, 5, 7 și respectiv 10. Mai jos sunt graficele care ilustrează aceste binuri:

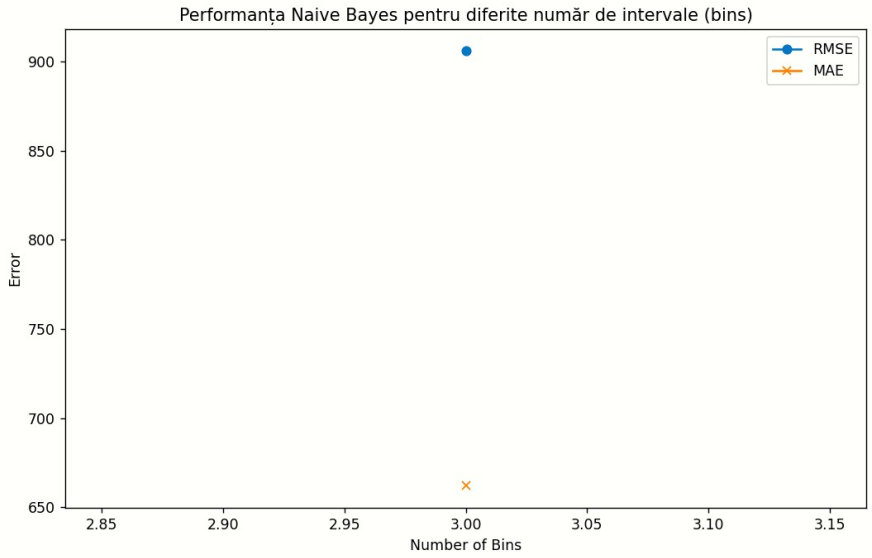

Figure 11: Bin 3

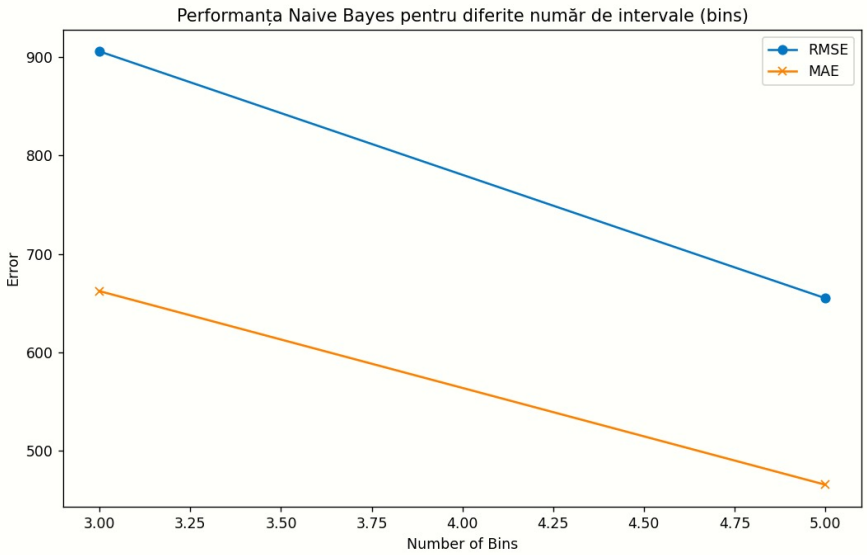

Figure 12: Bin 5

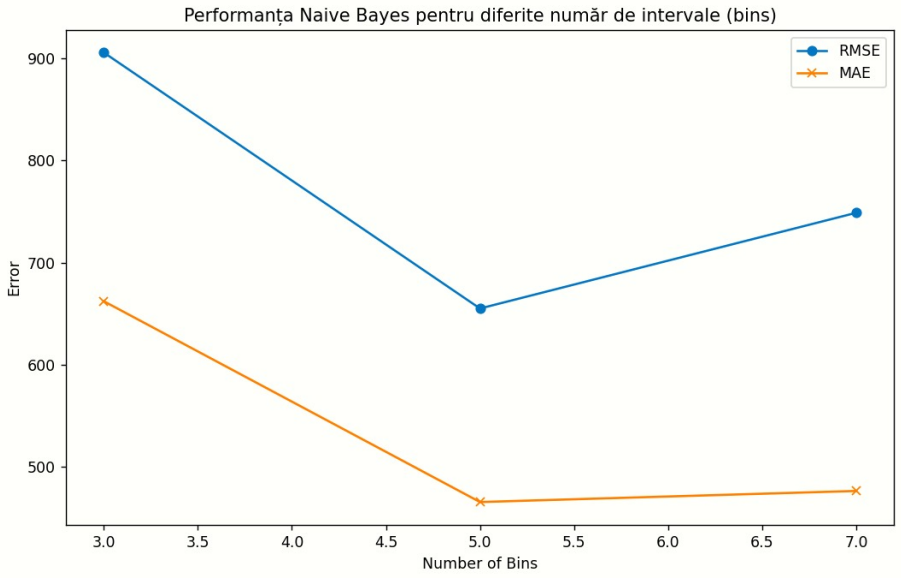

Figure 13: Bin 7

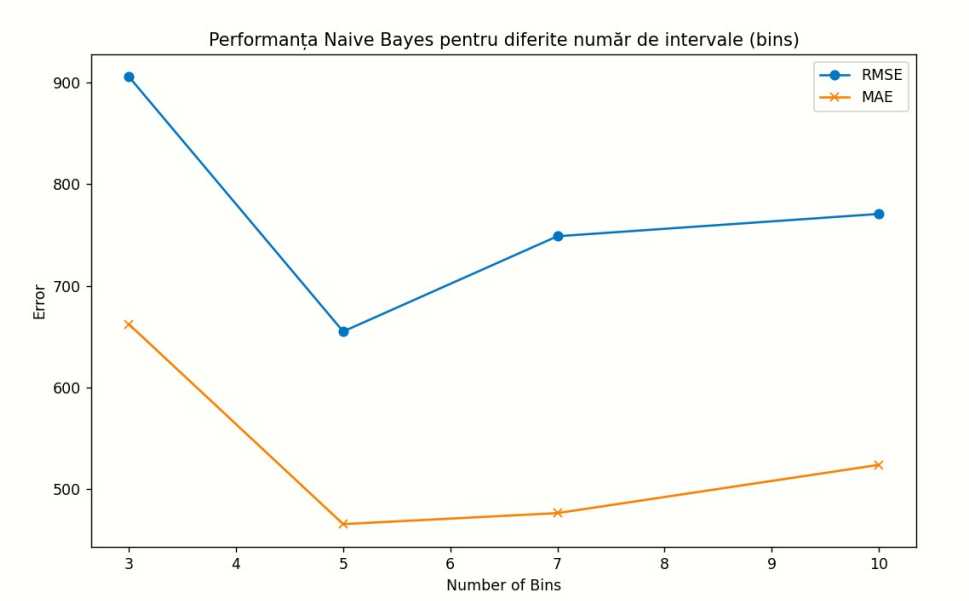

Figure 14: Bin 10

Pentru toate acestea valorile pentru RMSE și MAE sunt:

```
ID3 RMSE: 401.73870813293127, MAE: 321.6570858149823
Naive Bayes RMSE: 655.0689518781696, MAE: 465.4366015884637
ID3 Hybrid RMSE: 927.883048642792, MAE: 768.2394728143171
Naive Bayes (uniform bins) RMSE: 806.054952822618, MAE: 634.136189986358
Naive Bayes (percentile bins) RMSE: 655.0689518781696, MAE: 465.4366015884637
Naive Bayes (KMeans bins) RMSE: 854.1852274454434, MAE: 581.5761667705567
Naive Bayes bins=3: RMSE=905.8499349467209, MAE=662.1809973929522
Naive Bayes bins=5: RMSE=655.0689518781696, MAE=465.4366015884637
Naive Bayes bins=7: RMSE=748.8442147426944, MAE=476.2872511139583
Naive Bayes bins=10: RMSE=770.7847145054535, MAE=523.753688547643
```

Figure 15: RMSE și MAE pentru bin

Putem observa ca valorile sunt foarte mari deci ar trebui sa ne orientam spre alti algoritmi.

Asemănător am făcut pentru algoritmul ID3 pentru a putea face comparație în funcție de max depth. Tot pentru valorile 3, 5, 7 și respectiv 10 avem următoarele valori:

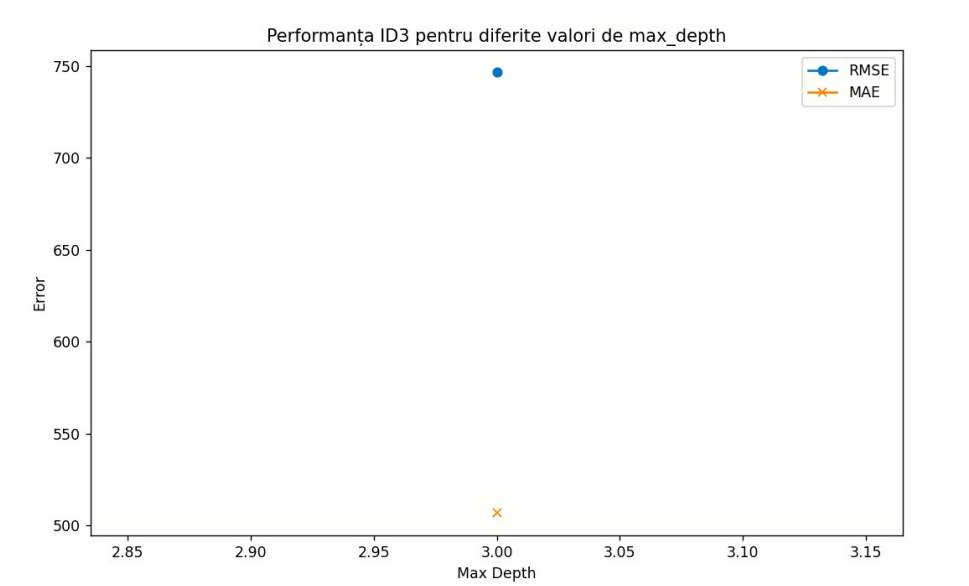

Figure 16: ID3 cu max depth = 3

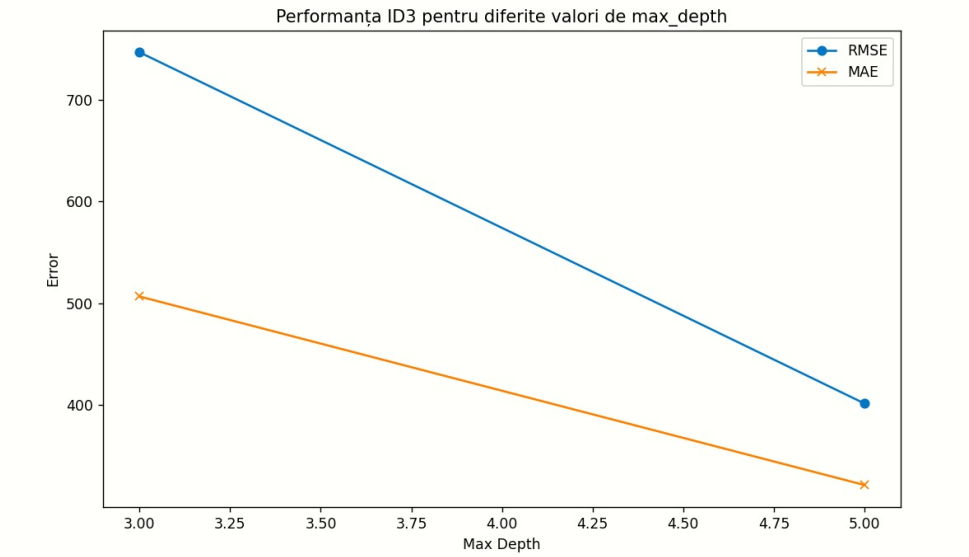

Figure 17: ID3 cu max depth = 5

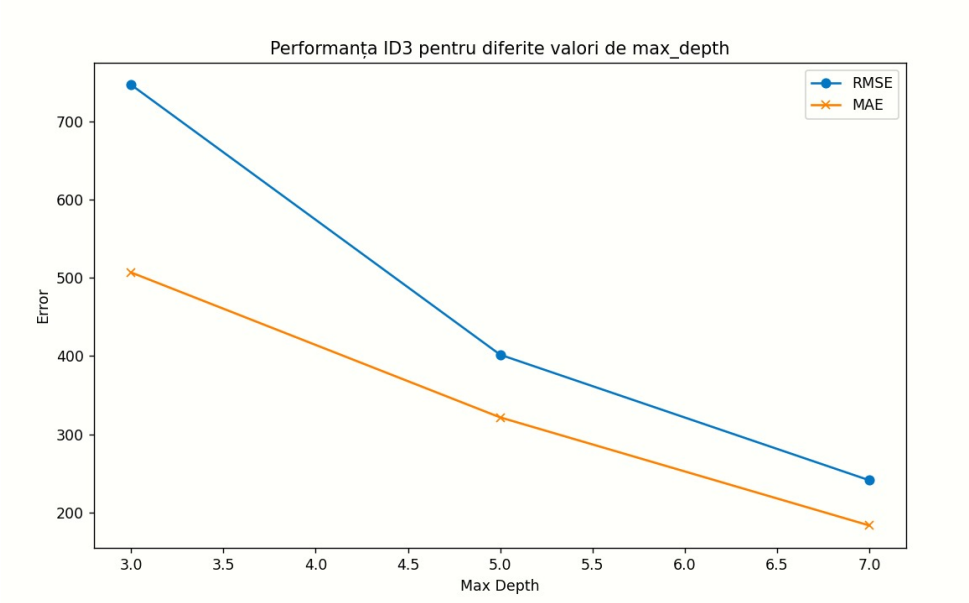

Figure 18: ID3 cu max depth = 7

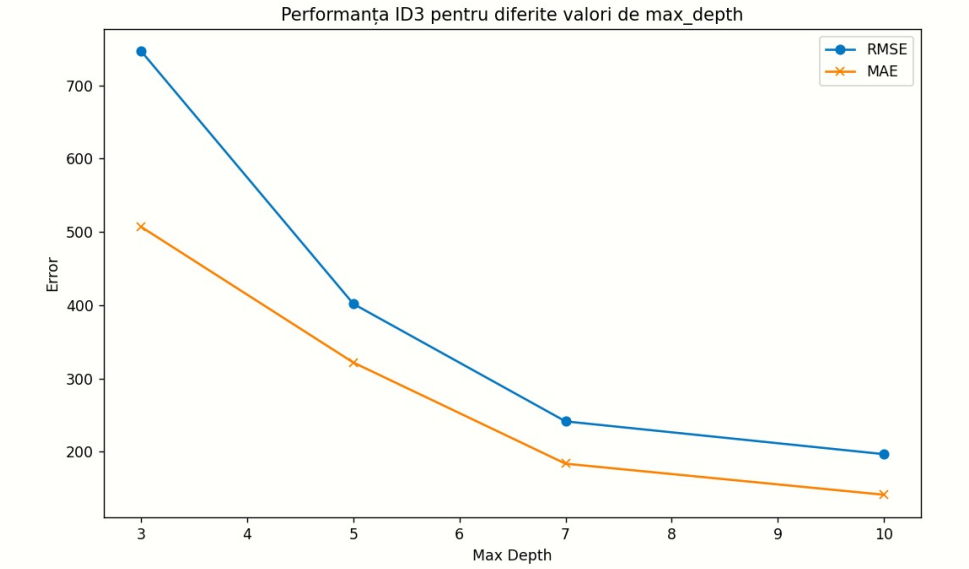

Figure 19: ID3 cu max depth = 10

Pentru toate acestea, valorile RMSE și MAE sunt:

```
ID3 max_depth=3: RMSE=746.6338846955159, MAE=506.8275324065047
ID3 max_depth=5: RMSE=401.73870813293127, MAE=321.6570858149823
ID3 max_depth=7: RMSE=241.34595384918535, MAE=183.58232394470232
ID3 max_depth=10: RMSE=196.61010460072652, MAE=141.19222465200147
```

Figure 20: RMSE și MAE pentru ID3

De aici putem observa că valoare 10 pentru max depth oferă cea mai mică valoare atât pentru RMSE cât și pentru MAE.

Prezentarea Rezultatelor

Am evaluat modelele de regresie folosind două metrici comune: RMSE (Root Mean Squared Error) și MAE (Mean Absolute Error). Rezultatele obținute pentru fiecare model sunt prezentate mai jos:

Modelul Naive Bayes a fost testat cu binuri uniforme și binuri generate prin KMeans. Performanța acestui model, măsurată prin RMSE și MAE, este următoarea:

- RMSE: 711.3560178321121
- MAE: 524.0463576158941

Rezultate pentru ID3

Modelul ID3 a oferit următoarele rezultate:

- RMSE: 199.3981167446919
- MAE: 153.84669692332926

Rezultate pentru Bayes cu uniform bins

Modelul Bayes uniform bins a oferit următoarele rezultate:

- RMSE: 892.0250763007866
- MAE: 680.4571692033138

Rezultate pentru Bayes KMeans bins

Modelul Bayes KMeans bins a oferit următoarele rezultate:

- RMSE: 808.5476576496166
- MAE: 518.560499245248

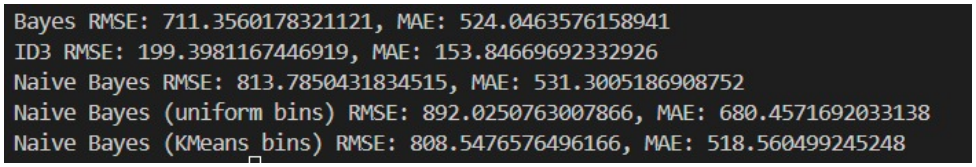

Figure 21: RMSE și MAE final

De mentionat ca valorile pentru ID3 sunt pentru noul dataset si pentru max depth = 10.

Analiza Comparativă

Comparând rezultatele pentru fiecare model, putem observa că ID3 cu max depth = 10 generează cel mai mic RMSE, ceea ce sugerează că acest model a avut cea mai bună performanță.

Grafice

În continuare, vom vizualiza predicțiile fiecărui model comparativ cu valorile reale pentru a observa diferențele între modelele de regresie.

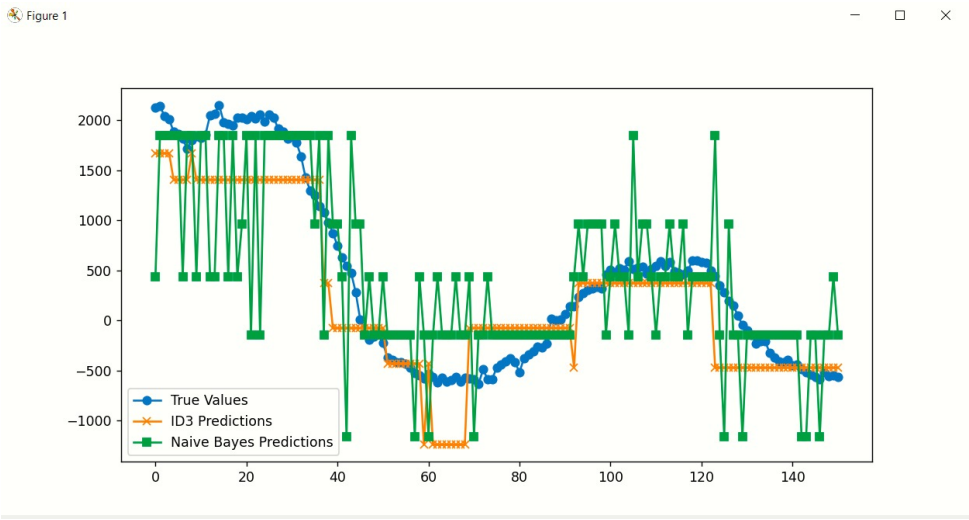

Figure 22: Comparație între valorile reale și predicțiile modelelor ID3 și Naive Bayes pentru lunile sezoniere

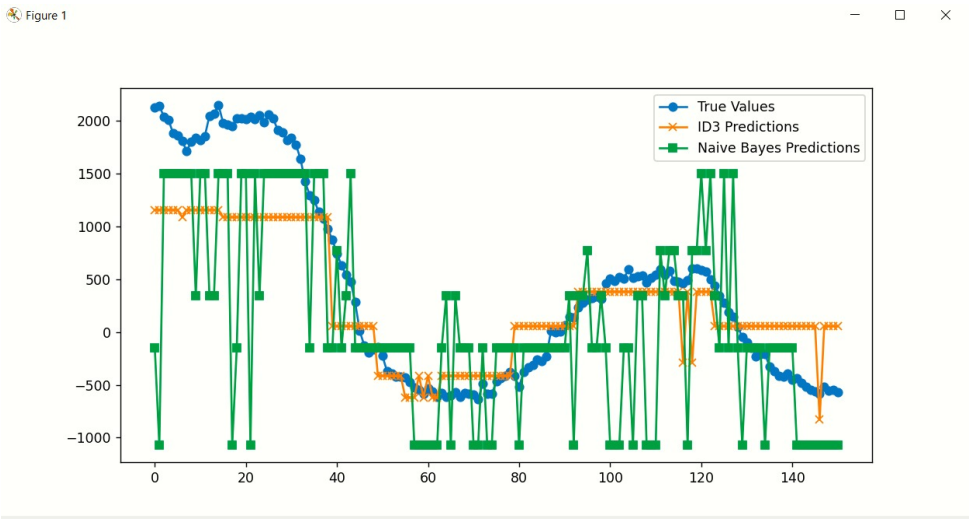

Figure 23: Comparație între valorile reale și predicțiile modelelor ID3 și Naive Bayes pentru tot anul

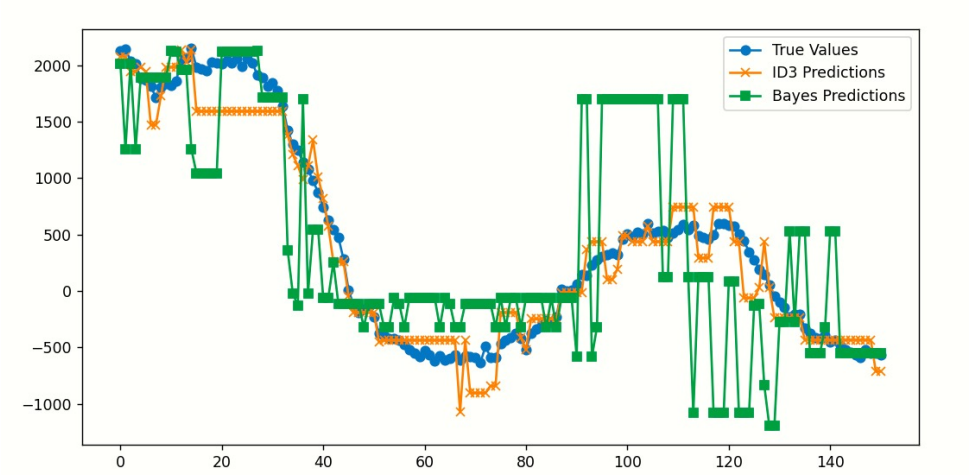

Figure 24: Comparație între valorile reale și predicțiile modelelor ID3 și Naive Bayes

Se poate observa foarte bine ca algoritmul ID3 cu max depth = 10 si cu datasetul luat pentru lunile ianuarie, iulie si noiembrie impreuna cu combinarea valorilor ce au corelatie foarte mare ofera o predictibilitate mult mai exacta. Desi unele puncte sunt clasificate gresit, acesta are cea mai mica eroare.

Concluzii

În urma testării și evaluării modelelor de regresie Naive Bayes și ID3, am observat că ambii algoritmi sunt capabili să prezică soldul total de energie cu o precizie rezonabilă. Totuși, modelul ID3 a avut performanțe mai bune în ceea ce privește reducerea erorilor, obținând valori mai mici pentru RMSE și MAE.

Pentru îmbunătățirea metodelor, am putea explora opțiuni suplimentare pentru discretizarea datelor, utilizând metode avansate de preprocesare, cum ar fi tehnici de scaling sau transformări non-liniare ale variabilelor. De asemenea, s-ar putea încerca optimizarea hiperparametrilor pentru ID3 și Naive Bayes pentru a obține o performanță mai bună.