Raport: Predictia soldului total in sistemul energetic national (SEN)

Bocăneț Raluca-Andreea

02.01.2025

1. Introducere

Scopul acestui proiect este de a prezice soldul total din Sistemul Energetic National (SEN) pentru luna decembrie 2024 folosind doua tehnici de invatare automata: **Decision Tree Regressor (ID3)** si **Bayes Naiv**. Datele utilizate pentru antrenarea si testarea modelelor provin dintr-un fisier Excel care contine informasiti despre consumul si productia de energie pe diferite surse de energie (carbune, hidrocarburi, eolian, etc.) si soldul de energie.

2. Procesarea datelor

2.1 Incarcarea si explorarea datelor

Datele au fost incarcate din fisierul Excel `Grafic_SEN.xlsx`. La inceput, au fost verificate informatiile generale despre date, precum tipurile de date si statistici descriptive, folosind urmatoarele comenzi:

```
print(data.info())
print(data.describe())
```

2.2 Curatarea datelor

Coloanele numerice care contin datele de consum si productie au fost curatate pentru a elimina caracterele non-numerice si a transforma valorile intr-un tip numeric adecvat (`float`). Procesul a fost realizat cu ajutorul urmatorului cod:

```
for col in columns_to_clean:
    data[col] = data[col].replace(r'[^\d.-]', '', regex=True).astype(float)
```

2.3 Conversia datelor

Coloana `Data` a fost transformata intr-un format de tip data (`datetime`), iar valorile lipsa au fost eliminate pentru a asigura o baza de date curata si completa. Codul folosit pentru aceasta conversie este:

```
data['Data'] = pd.to_datetime(
                data['Data'],
                format='%d-%m-%Y %H:%M:%S',
                errors='coerce'
            )
data = data.dropna()
```

2.4 Divizarea datelor in seturi de antrenament si testare

Datele au fost impartite pe baza lunii. Setul de antrenament include datele din lunile diferite de decembrie, iar setul de testare include datele din luna decembrie 2024. Daca setul de date de antrenament este gol, s-a folosit o divizare aleatorie a datelor. Codul utilizat este:

```
train_data = data[data['Data'].dt.month != 12]
test_data = data[data['Data'].dt.month == 12]
```

2.5 Selectarea caracteristicilor si a tintei

Caracteristicile utilizate pentru a prezice soldul sunt variabilele legate de consumul si productia de energie. Tinta este soldul de energie (`Sold[MW]`). Codul pentru definirea caracteristicilor si a tintei:

```
features = [
            'Consum[MW]', 'Medie Consum[MW]',
            'Productie[MW]', 'Carbune[MW]', 'Hidrocarburi[MW]', 'Ape[MW]',
            'Nuclear[MW]', 'Eolian[MW]', 'Foto[MW]', 'Biomasa[MW]'
           ]
target = 'Sold[MW]'
```

3. Modele de invatare automata

3.1 Modelul Decision Tree Regressor (ID3)

Modelul ID3 a fost antrenat cu datele de antrenament si a fost folosit pentru a prezice soldul de energie. Performanta a fost evaluata folosind doua metrici: **Root Mean Squared Error (RMSE)** si **Mean Absolute Error (MAE)**.

```
id3_model = DecisionTreeRegressor(max_depth=5, random_state=42)
id3_model.fit(X_train, y_train)
y_pred_id3 = id3_model.predict(X_test)
```

Rezultatele obtinute pentru modelul ID3 au fost:

RMSE: 109.52

MAE: 85.41

3.2 Modelul Bayes Naiv

Pentru modelul Bayes Naiv, datele de antrenament si testare au fost discretizate in intervale de valori (binning) pentru a se potrivi cu natura algoritmului.

```
bins = np.linspace(X_train.min().min(), X_train.max().max(), 10)
X_train_binned = np.digitize(X_train, bins=bins)
X_test_binned = np.digitize(X_test, bins=bins)
bayes_model = GaussianNB()
bayes_model.fit(X_train_binned, y_train)
y_pred_bayes = bayes_model.predict(X_test_binned)
```

Performanta modelului Bayes Naiv a fost evaluata cu urmatoarele rezultate:

RMSE: 295.08

MAE: 228.37

4. Compararea performantei

O comparatie vizuala intre predictiile celor doua modele a fost realizata printr-un grafic de dispersie, unde valorile reale ale soldului au fost comparate cu valorile prezise de ambele modele. Graficul a fost generat folosind urmatorul cod:

```
plt.figure(
            figsize=(10, 6))sns.scatterplot(x=y_test,
            y=y_pred_id3, label='ID3', color='blue', s=100, marker='o',
            alpha=0.6)sns.scatterplot(x=y_test, y=y_pred_bayes,
            label='Bayesian', color='red', s=100,
```

```
            marker='^', alpha=0.6)
for i in range(len(y_test)):
    plt.text(y_test.iloc[i], y_pred_id3[i],
    f'{data["Data"].iloc[i].strftime("%d-%m-%Y")}',
    fontsize=8, alpha=0.6)
plt.xlabel('Valoare reala')
plt.ylabel('Valoare prezisa')
plt.title('Compararea predictiilor cu Sold si Data')
plt.legend()
plt.show()
```

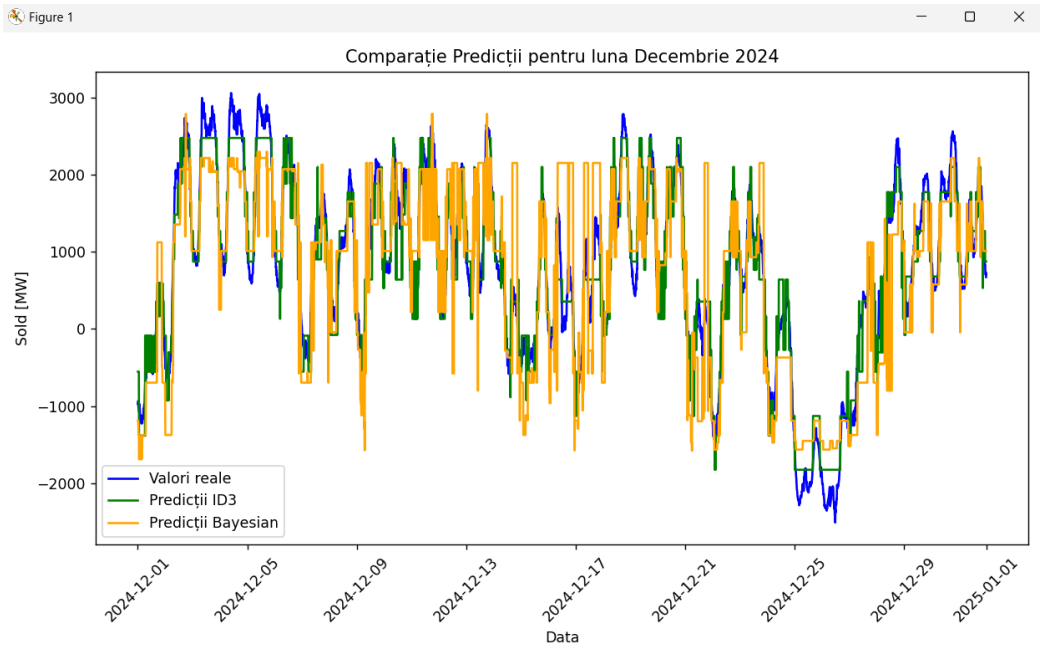

Figure 1: Compararea predictiilor intre modelele ID3 si Bayes Naiv

5. Cel mai mare sold total din decembrie 2024

Am calculat soldul total pe luna decembrie 2024 folosind valorile reale si predictiile realizate de cele doua modele implementate. Rezultatele sunt urmatoarele:

Sold Total Real: 3600389.00 MW

Sold Total Prezis de ID3: 3391099.62 MW

Sold Total Prezis de Bayesian: 3195007.00 MW

Aceste valori reflecta performanta fiecarui model si contribuie la evaluarea diferentelor dintre valorile reale si cele estimate.

Pentru a obtine aceste informatii, s-a folosit urmatorul cod:

```
total_sold_real = y_test.sum()
total_sold_id3 = y_pred_id3.sum()
total_sold_bayes = y_pred_bayes.sum()

print(f"Soldul total real pe decembrie 2024: {total_sold_real:.2f} MW")
print(f"Soldul total prezis de ID3 pe decembrie 2024
                {total_sold_id3:.2f} MW")
print(f"Soldul total prezis de modelul Bayesian pe decembrie 2024
                {total_sold_bayes:.2f} MW")
```

6. Concluzii

Modelul **Decision Tree Regressor (ID3)** a oferit o performanta mai buna decat modelul Bayes Naiv, avand valori RMSE si MAE mai mici. In ciuda acestui fapt, ambele modele au avut performante rezonabile, avand in vedere complexitatea datelor si natura lor.

Este recomandat sa se testeze si alte tehnici de invatare automata, cum ar fi **Random Forest**, pentru a observa daca o performanta mai buna poate fi obtinuta.

De asemenea, pentru a imbunatati performanta generala a modelelor, este recomandata extinderea setului de date utilizat pentru antrenament. Aceasta poate include:

- **Date istorice suplimentare:** Adaugarea datelor din alte perioade (luni sau ani) pentru a oferi modelelor o diversitate mai mare si o mai buna intelegere a sezonalitatii si a altor tipare.
- **Date externe:** Integrarea de variabile externe, precum date meteorologice, preturi ale energiei sau alte indicatori economici relevanti, care ar putea influenta consumul si productia de energie.