ML AP1

Cazacu Ion

Decembrie 2024

1 Descrierea problemei

1.1 Contextul

Proiectul solicită dezvoltarea unui model de predicție pentru soldul total al Sistemului Energetic Național (SEN) al României, pentru luna decembrie 2024. Soldul este calculat ca diferența dintre producția totală și consumul total de energie electrică. Datele necesare pentru această analiză provin din seturi istorice, care descriu consumul și producția de energie electrică defalcate pe surse precum hidro, eolian, nuclear, cărbune și altele. Soluția trebuie să fie bazată pe algoritmii de învățare automată **ID3 (arbore de decizie)** și **clasificare bayesiană**, adaptați pentru o problemă de regresie. În plus, proiectul impune limitări, cum ar fi excluderea datelor din decembrie pentru antrenarea modelelelor.

1.2 Scopul proiectului

Scopul principal al programului este de a prezice soldul energetic final pentru fiecare zi din luna decembrie 2024 pe baza datelor istorice disponibile. Principalele obiective:

1.2.1 Analiza Datelor

- Înțelegerea variabilelor furnizate în setul de date (producție, consum, sold) și relațiile dintre ele.
- Preprocesarea datelor pentru eliminarea zgomotului și pregătirea unui set adecvat de antrenament.

1.2.2 Adaptarea Algoritmilor

- **ID3 (Arbore de Decizie):** Transformarea algoritmului pentru a suporta probleme de regresie prin discretizarea intervalelor de valori ale soldului (*bucketing*).
- **Clasificare Bayesiană:** Discretizarea variabilelor continue și calcularea probabilităților condiționate.

1.2.3 Evaluarea Performanței

- Pentru evaluarea performanței modelului de regresie, a fost utilizată metrica **Mean Squared Error (MSE)**. Aceasta calculează eroarea pătratică medie între valorile prezise și valorile reale. Modelul a fost folosit pentru a prezice valorile soldului pentru fiecare zi din luna decembrie 2024, iar rezultatele au fost evaluate folosind funcția `mean_squared_error`.

Logica Programului

1. **Citirea și curățarea datelor**: Programul încarcă fișierul Excel care conține date energetice și elimină rândurile incomplete sau cu valori negative, asigurând calitatea datelor utilizate.
2. **Preprocesarea datelor**: Datele sunt filtrate pentru luna decembrie, iar anul este ajustat pentru a reflecta anul 2024. Valorile zilnice medii sunt calculate pentru caracteristicile relevante.
3. **Antrenarea modelului**: Un arbore de decizie (ID3) este antrenat folosind datele istorice pentru a captura relațiile dintre caracteristicile energetice și soldul final.
4. **Aplicarea metodei Bayesiene**: În paralel cu modelul ID3, este utilizată o metodă bayesiană pentru a estima probabilitățile de distribuție a soldului pe baza datelor istorice. Aceasta ajută la îmbunătățirea preciziei predicțiilor, oferind o abordare complementară.
5. **Generarea predicțiilor**: Soldul final este prezis pentru fiecare zi a lunii decembrie 2024 utilizând modelele antrenate (ID3 și Bayesian).
6. **Exportarea rezultatelor**: Rezultatele sunt salvate într-un fișier Excel prietenos utilizatorului pentru analiză ulterioară.

Codul Implementat

1.3 Citirea fișierului

Fișierul Excel este citit utilizând `pandas`, iar coloana `Data` este convertită în format `datetime` pentru a facilita manipulările. Astfel, putem manipula ușor datele și le putem filtra după dată.

```
import pandas as pd

# Citirea fiierului Excel
file_path = 'grafic/Grafic_SEN.xlsx'
data = pd.ExcelFile(file_path)
df = data.parse("Grafic SEN")

# Convertirea coloanei Data n format datetime
df['Data'] = pd.to_datetime(df['Data'], errors='coerce', dayfirst=True)
```

1.4 Curățarea datelor

Sunt eliminate rândurile care conțin valori lipsă sau negative în coloanele cheie. Această etapă asigură că datele sunt valide și pregătite pentru modelare.

```
# Eliminarea rndurilor cu valori lips sau negative
cleaned_df = df.dropna()
cleaned_df = cleaned_df[cleaned_df[['Foto[MW]', 'Consum[MW]',
                          'Productie[MW]']].min(axis=1) >= 0]
```

1.5 Împărțirea datelor

Datele sunt împărțite în seturi de antrenare și testare pentru a evalua performanța modelului în mod obiectiv.

```
from sklearn.model_selection import train_test_split

#   mprirea   datelor  n  seturi de antrenare i testare
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]'

X = cleaned_df[features]
y = cleaned_df[target]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
    random_state=25)
```

1.6 Antrenarea modelelor

Se utilizează `DecisionTreeRegressor` și `GaussianNB` pentru a antrena modelele de arbore de decizie și respectiv Bayes Naiv.

```
from sklearn.tree import DecisionTreeRegressor
from sklearn.naive_bayes import GaussianNB

# Antrenarea modelului Decision Tree
id3_model = DecisionTreeRegressor(random_state=25)
id3_model.fit(X_train, y_train)

# Antrenarea modelului Naive Bayes
nb_model = GaussianNB()
nb_model.fit(X_train, y_train_binned)
```

1.7 Predicția pentru decembrie

Datele din decembrie 2024 sunt procesate și prezise folosind modelele antrenate.

```
# Predicia pentru decembrie folosind modelele antrenate
december_data = cleaned_df[cleaned_df['Data'].dt.month == 12]
december_data['Predicted Sold[MW] - Decision Tree'] = id3_model.predict(
    X_daily)
december_data['Predicted Sold[MW] - Naive Bayes'] = nb_model.predict(
    X_daily)
```

1.8 Exportarea rezultatelor

Rezultatele sunt salvate într-un fișier Excel care conține doar coloanele `Data` și `Predicted Sold[MW]`.

Codul pentru exportarea rezultatelor este:

```
# Exportarea rezultatelor ntr-un fiier Excel
data_decembrie[['Data', 'Predicted Sold[MW]']].to_excel('
    Predicted_December_2024_Sold.xlsx', index=False)
```

Rezultatele Prezise

Rezultatele reprezintă soldul energetic prezis pentru fiecare zi a lunii decembrie 2024. Predicțiile sunt generate pe baza caracteristicilor medii calculate din datele istorice și oferă o estimare zilnică a soldului final.

Data	Predicted Sold[MW] - Decision Tree	Predicted Sold[MW] - Naive Bayes
2024-12-01	-940	-1556.722222
2024-12-02	-895	-1556.722222
2024-12-03	-1088	-1556.722222
2024-12-04	1040	1193.833333
2024-12-05	431	643.7222222
2024-12-06	-24	643.7222222
2024-12-07	205	643.7222222
2024-12-08	34	643.7222222
2024-12-09	-377	93.61111111
2024-12-10	-149	93.61111111
2024-12-11	818	1193.833333
2024-12-12	1196	1193.833333
2024-12-13	815	643.7222222
2024-12-14	402	643.7222222
2024-12-15	-439	93.61111111
2024-12-16	-220	93.61111111
2024-12-17	-288	93.61111111
2024-12-18	938	643.7222222
2024-12-19	837	643.7222222
2024-12-20	808	1193.833333
2024-12-21	948	1193.833333
2024-12-22	-1137	-1006.611111
2024-12-23	-1287	-2106.833333
2024-12-24	-465	-1556.722222
2024-12-25	-1864	-1556.722222
2024-12-26	-1115	-1556.722222
2024-12-27	-1161	-1556.722222
2024-12-28	-1053	-1006.611111
2024-12-29	-1088	-1006.611111
2024-12-30	-924	-1006.611111
2024-12-31	-1277	-1556.722222

Fișierul Excel exportat conține următoarele informații:

- **Data**: Fiecare zi din decembrie 2024.
- **Predicted Sold[MW] - Decision Tree**: Soldul energetic prezis pentru ziua respectivă cu ajutorul modelului ID3.
- **Predicted Sold[MW] - Naive Bayes**: Soldul energetic prezis pentru ziua respectivă cu ajutorul metodei Bayesiene.

Observații

- Performanța modelului poate depinde de calitatea și cantitatea datelor de antrenament.
- `random_state=25` asigură reproducibilitatea antrenării, fără a afecta semnificativ performanța modelului.
- Exportul rezultatelor într-un format simplu facilitează analiza ulterioară.

Concluzie

Programul utilizează un arbore de decizie și un model Bayes Naiv pentru predicția soldului energetic zilnic, cu rezultate utile pentru planificarea energetică. Datele exportate permit integrarea ușoară în rapoarte și analize ulterioare.