UNIVERSITATEA „POLITEHNICA'' din BUCUREȘTI
Facultatea de Electronică, Telecomunicații si Tehnologia Informației

Proiect

Ingineria **S**istemelor cu **I**nteligență **A**rtificială

Forest Fires – Random Forest

Pescaru Anamaria Andra

Grupa: 424A

Bucuresti 2022

Scopul proiectului

Acest proiect are ca scop implementarea unui program care să prezică zonele arse de incediile forestiere, in regiunea de nord-est a Portugaliei.

Aceasta este o problema de regresie pe care am rezolvat-o folosind metoda Random Forest cu 10 arbori.

Datele inițiale ale proiectului

Pentru acest proiect s-au folosit datele de pe acest site: https://archive.ics.uci.edu/ml/datasets/Forest+Fires .

Setul de date contine 13 atribute :

1. X - coordonata spatiala x din harta parcului Montesinho : de la 1 la 9;
2. Y – coordonata spatiala y din harta parcului Montesinho : de la 2 la 9;
3. month – denumirea lunii dintr-un an : din “ianuarie” până in “decembrie”;
4. day – denumirea zilei dintr-o saptamana : de “luni” până “duminică”;
5. FFMC – FFMC index din sistemul FWI : de la 18.7 la 96.20;
6. DMC – DMC index din sistemul FWI : de la 1.1 la 291.3;
7. DC – DC index din sistemul FWI : de la 7.9 la 860.6;
8. ISI – ISI index din sistemul FWI : de la 0.0 la 56.10;
9. temp – temperatura in grade Celsius : de la 2.2 la 33.30;
10. RH - umiditatea relativa in % : de la 15.0 la 100;
11. wind – viteza vantului in km/h : de la 0.40 la 9.40;

12. rain – cantitatea de precipitatii mm/m2 : de la 0.0 la 6.4;
13. area – portiunea arsa dintr-o padure (in ha) : de la 0.0 la 1090.84;

Etapele de construire a codului

1. Importăm librariile de care avem nevoi.

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from sklearn import datasets, tree, ensemble
from sklearn.ensemble import BaggingRegressor
```

2. Extragem datele din fisierul csv .

```
datasets = pd.read_csv('forestfires.csv')
```

3. Pentru ca programul să funcționeze, este necesar ca in ‘datasets’ să existe doar date numerice.

    - Am inlocuit denumirea lunilor de pe coloana ‘month’ cu numere de la 1 – 12 in urmatorul mod :

```
month_map = {"jan": 1,"feb" : 2 ,"mar" : 3,"apr" : 4 ,"may" : 5, "jun" : 6, "jul" : 7,
             "aug" : 8,"sep" : 9, "oct" : 10, "nov" : 11, "dec" : 12}
datasets['month'] = datasets['month'].map(month_map)
```

    - Am inlocuit denumirea zilelor saptamanii de pe coloana ‘day’ cu numere de la 1 – 7 in urmatorul mod:

```
day_map = {"mon": 1,"tue" : 2 ,"wed" : 3,"thu" : 4 ,"fri" : 5, "sat" : 6, "sun" : 7}
datasets['day'] = datasets['day'].map(day_map)
```

4. Am construit o matrice in care am introdus datele din ‘datasets’.

```
datasets = datasets.to_numpy()
datasets = np.array(datasets)
```

5. Impărțim datele.

   Setul de date este format din 517 eșantioane, adica matricea noastra este compusa din 517 linii si 13 coloane.

   Imparțirea datelor se realizează astfel:

   - 75% din date sunt folosite pentru setul de antrenare

```
data_train = datasets[: 387, 0 : 13]
etichete_train = datasets[0 : 387, 12]
```

   - 25% din date sunt folosite pentru setul de testare

```
data_test = datasets[387 : , 0 : 13]
etichete_test = datasets[387 : , 12]
```

Observații:

- Datele care trebuie etichetate in această problem sunt datele care se afla pe ultima coloană.
- ‘data_train’ si ‘data_test’ sunt niste matrici.
- ‘etichete_train’ si ‘etichete_test’ sunt niste vectori.

6. Am contruit un vector care sa retina procentul "in-bag" (25%,50%, 85%) si un vector care sa retina numarul de dimensiuni alese intr-un nod (10%, 50% 80%).

```
in_bag = [0.25, 0.5, 0.85 ];
dimensiuni_nod = [0.1, 0.5, 0.8];
```

7. Am creat si antrenat arborele de regresie prin varierea concomitenta a procentului 'in-bag' si a dimeniunii nodului folosind doua for – uri.

```
for i in range(3):
    for j in range(3):
        regr = BaggingRegressor(n_estimators = 10, max_samples = in_bag[i], max_features = dimensiuni_nod[j], random_state=0)
        regr_fit = regr.fit(data_train, etichete_train)

        predictii = regr.predict(data_test);
        suma = 0;
        for k in range(0,len(etichete_test)):
            suma += (predictii[k] - etichete_test[k]) ** 2;

        print('Eroarea patratica medie pentru in_bag ', in_bag[i], ' si dimensiunea nodului de', dimensiuni_nod[j], ' este ', suma/len(etichete_test), '\n');
```

Observatii:

- Antrenarea s-a realizat cu ajutorul functiei 'fit' care a primit ca parametrii datele din cele doua seturi de antrenare (data_train si etichete_test).
- Predictiile s-au realizat cu ajutorul functiei 'predict' care a primit ca parametrii datele din "data_test".
- Datorita faptului ca este o problema de regresie, am calculat eroarea patratica medie cu ajutorul unui for.

Programul afiseaza pe rand, pentru pereachea :

- Procent in-bag de 0.25 si dimensiune nod de 0.1:

```
Eroarea patratica medie pentru in_bag  0.25  si dimensiunea nodului de 0.1  este  4644.406702252589
```

- Procent in-bag de 0.25 si dimensiune nod de 0.5:

```
Eroarea patratica medie pentru in_bag  0.25  si dimensiunea nodului de 0.5  este  2267.6379994999975
```

- Procent in-bag de 0.25 si dimensiune nod de 0.8:

```
Eroarea patratica medie pentru in_bag  0.25  si dimensiunea nodului de 0.8  este  1333.7244451769236
```

- Procent in-bag de 0.5 si dimensiune nod de 0.1:

```
Eroarea patratica medie pentru in_bag  0.5  si dimensiunea nodului de 0.1  este  4608.727481039542
```

- Procent in-bag de 0.5 si dimensiune nod de 0.5:

```
Eroarea patratica medie pentru in_bag  0.5  si dimensiunea nodului de 0.5  este  1948.3318764260675
```

- Procent in-bag de 0.5 si dimensiune nod de 0.8:

```
Eroarea patratica medie pentru in_bag  0.5  si dimensiunea nodului de 0.8  este  846.6002425769229
```

- Procent in-bag de 0.85 si dimensiune nod de 0.1:

```
Eroarea patratica medie pentru in_bag  0.85  si dimensiunea nodului de 0.1  este  4652.160440777733
```

- Procent in-bag de 0.85 si dimensiune nod de 0.5:

```
Eroarea patratica medie pentru in_bag  0.85  si dimensiunea nodului de 0.5  este  1939.8552585713676
```

- Procent in-bag de 0.85 si dimensiune nod de 0.8:

```
Eroarea patratica medie pentru in_bag  0.85  si dimensiunea nodului de 0.8  este  392.25324046538447
```