Jan Kaiser
48432
Inteligentne aplikacje

Zastosowanie algorytmu decision tree regression w celu predykcji produkcji energii elektrycznej

Zestaw danych

Date zostały pobrane z strony kaggle
https://www.kaggle.com/datasets/girumwondemagegn/dataset-for-renewable-energy-systems

“Dataset for renewable energy systems” to tabela zawierająca informacje o instalacjach energii odnawialnych. Tabela składa się 13 kolumn:

Informacje o wszystkich kolumnach (informacja ze strony zbioru danych):

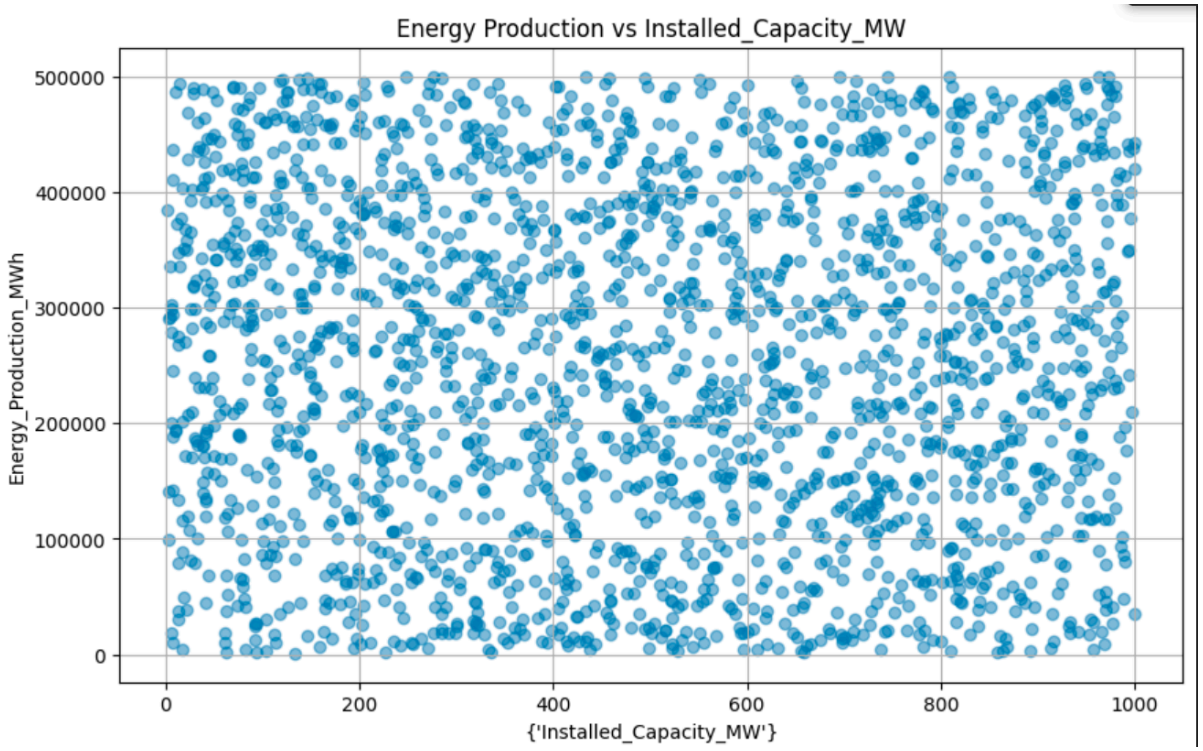
1. **Type_of_Renewable_Energy**: Numerical code representing the type of renewable energy source (1: Solar, 2: Wind, 3: Hydroelectric, 4: Geothermal, 5: Biomass, 6: Tidal, 7: Wave).
2. **Installed_Capacity_MW**: Installed capacity in megawatts (MW).
3. **Energy_Production_MWh**: Yearly energy production in megawatt-hours (MWh).
4. **Energy_Consumption_MWh**: Yearly energy consumption in megawatt-hours (MWh).
5. **Energy_Storage_Capacity_MWh**: Energy storage capacity in megawatt-hours (MWh).
6. **Storage_Efficiency_Percentage**: Efficiency of energy storage systems in percentage.
7. **Grid_Integration_Level**: Numerical code representing the level of grid integration (1: Fully Integrated, 2: Partially Integrated, 3: Minimal Integration, 4: Isolated Microgrid).
8. **Initial_Investment_USD**: Initial investment costs in USD.
9. **Funding_Sources**: Numerical code representing the funding source (1: Government, 2: Private, 3: Public-Private Partnership).
10. **Financial_Incentives_USD**: Financial incentives in USD.
11. **GHG_Emission_Reduction_tCO2e**: Reduction in greenhouse gas emissions in tons of CO2 equivalent (tCO2e).
12. **Air_Pollution_Reduction_Index**: Air pollution reduction index.
13. **Jobs_Created**: Number of jobs created.

Wykorzystany algorytm - decision tree regression - to algorytm z rodziny regresji.
Implementacja algorytmu: predykcja rocznej produkcji energii na podstawie wartości innych kolumn.

Implementacja algorytmu

Każdy poszczególny krok implementacji algorytmu jest opisany w pliku projektu.

Korelacja między poszczególnymi kolumnami a roczną energią produkcji jest praktycznie zerowa. Jest to prawdziwe w przypadku każdej kolumny (poniżej przykład relacji między wartościami rocznej produkcji, a zainstalowanej mocy).



Skuteczności algorytmu nie pomogło wykorzystanie innych parametrów, normalizacja danych, wykorzystanie różnych kolumn. Inny algorytm,(random forest regression) również nie poprawił działania algorytmu.

Błąd względny w najlepszym przypadku wyniósł około 57%.

Stwierdzono, że powodem tak dużej wartości błędu względnego jest bliska zeru korelacja między poszczególnymi kolumnami a roczną produkcją energii.

Wyniki wykonania algorytmu

Wyniki wykonania algorytmu dla parametrów:
`maxDepth=20,`
`minInstancesPerNode=100`

średni błąd kwadratowy: 57%

```
+------------------+--------------------+--------------------+
|        prediction|Energy_Production_MWh|featuresAfterScaling|
+------------------+--------------------+--------------------+
|253487.65339319356|          383838.6161|[1.33698697070147...|
|253487.65339319356|           99308.24834|[-1.3440237442314...|
|253487.65339319356|          291569.6233|[-1.3440237442314...|
| 282184.8746911296|           335762.192|[-0.4503535059204...|
|253487.65339319356|          299365.4826|[-0.4503535059204...|
|253487.65339319356|           18026.9665|[1.33698697070147...|
|253487.65339319356|           200492.119|[1.33698697070147...|
|253487.65339319356|          142363.1731|[1.33698697070147...|
|253487.65339319356|          294022.2633|[1.33698697070147...|
| 282184.8746911296|          410662.0138|[-1.3440237442314...|
+------------------+--------------------+--------------------+
only showing top 10 rows


root squared mean error (RMSE) on test data = 144464.29736683954
Mean Energy Production: 252813.675988132
Standard Deviation of Energy Production: 144990.30899856513
normalized rmse = 57.142595946281496%
```

Wyniki wykonania algorytmu dla parametrów
`maxDepth=20,`
`minInstancesPerNode=1`

średni błąd kwadratowy: 79%

Znacznie większy błąd wynika z nieprawidłowego doboru parametrów.

Dobór zbyt dużej głębokości (maxDepth) może prowadzić do nadmiernego dopasowania, gdzie model uczy się nie tylko wzorców, ale także szumów w danych treningowych (overfitting). Skutkuje to słabą predykcją nowych, nieznanych danych.

W tym przypadku dobór zbyt niskiej wartości parametru minInstancePerNode spowodował, że drzewo dobierało zbyt mało punktów danych do każdego węzła analizy. Powodowało to, że model uczył się bardzo konkretnych odpowiedzi, ale nie był gotowy na nieznane dane.

```
+------------------+--------------------+--------------------+
|        prediction|Energy_Production_MWh|featuresAfterScaling|
+------------------+--------------------+--------------------+
|       68381.08865|          383838.6161|[1.34668394263377...|
|       181815.1383|           200492.119|[1.34668394263377...|
|        395690.667|          188046.4142|[1.34668394263377...|
|169121.29850000003|          245465.8712|[-1.3828851238873...|
|       140515.0998|          30498.64368|[-0.4730287683803...|
|       444116.4606|          115504.0859|[-1.3828851238873...|
|        321660.257|           88969.3757|[1.34668394263377...|
|269978.22160000005|          367428.1968|[0.43682758712672...|
|       438504.3611|          171615.3748|[1.34668394263377...|
|       38494.45181|           428189.566|[-1.3828851238873...|
+------------------+--------------------+--------------------+
only showing top 10 rows

root squared mean error (RMSE) on test data = 199798.2499040837
Mean Energy Production: 252813.675988132
Standard Deviation of Energy Production: 144990.30899856513
normalized rmse = 79.02984248109385%
```