Robert Kamiński
Michał Jagielski

Raport

Analiza i modelowanie charakterystyki energetycznej budynków.

1. Motywacja i cele

Celem tego projektu jest poznanie metod data science do analizy praktycznych problemów inżynierskich oraz rozwinięcie umiejętności myślenia analitycznego. Jest to okazja do nabycia doświadczenia z zakresu wykorzystania technik komputerowych w celu usprawniania tworzenia rzeczywistych obiektów i rozwiązywania trudnych problemów analitycznych.

2. Struktura zbiorów

Nasz zbiór danych składa się z 8 parametrów opisujących różne właściwości budynku takie jak obszar powierzchni czy współczynnik dystrybucji przeszklenia. Mają one pomóc w przewidywaniu dwóch współczynników dotyczących energii budynku: obciążenie chłodzące i cieplne.

- Obciążenie chłodzenia to to ilość energii cieplnej, która musiałaby zostać usunięta z przestrzeni poprzez chłodzenie, aby utrzymać temperaturę na wymaganym poziomie.

- Obciążenie cieplne to ilość energii cieplnej, która musiałaby zostać dodana do przestrzeni poprzez ogrzewanie, by utrzymać temperaturę na wymaganym poziomie.

Parametrami opisującymi nasz budynek są:

- Względna zwartość (X1)
- Obszar powierzchni (X2)
- Powierzchnia ściany (X3)
- Powierzchnia dachu (X4)
- Całkowita wysokość (X5)
- Orientacja (X6)
- Obszar zaszklony (X7)
- Dystrybucja zaszklenia (X8)
- Obciążenie cieplne (Y1)
- Obciążenie chłodzenia (Y2)

Wszystkie dane składają się z danych liczbowych. Dane posiadają 768 rekordów.

	X1	X2	X3	X4	X5	X6	X7	X8	Y1	Y2
0	0.98	514.5	294.0	110.25	7.0	2	0.0	0	15.55	21.33
1	0.98	514.5	294.0	110.25	7.0	3	0.0	0	15.55	21.33
2	0.98	514.5	294.0	110.25	7.0	4	0.0	0	15.55	21.33
3	0.98	514.5	294.0	110.25	7.0	5	0.0	0	15.55	21.33
4	0.90	563.5	318.5	122.50	7.0	2	0.0	0	20.84	28.28

Ilustracja 1: Pięć pierwszych wierszy zbioru danych.

Dane zostały pobrane ze strony pod adresem:

https://archive.ics.uci.edu/ml/datasets/energy+efficiency?fbclid=IwAR2e0tDnmKY71ziUfI0fEAQkbnnjHFK2jLzBz8Opo6QbtUnBqhnE227axUY

3. Przygotowanie danych

Pierwszym krokiem było pobranie i zaimportowanie danych do analizy. Plik danych jest zapisany w formacie xlsx. Pobranie go do pamięci programu python jest możliwe poprzez użycie pakietu „pandas".

Dane zostały sprawdzone pod kątem brakujących i pustych rekordów:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 10 columns):
X1    768 non-null float64
X2    768 non-null float64
X3    768 non-null float64
X4    768 non-null float64
X5    768 non-null float64
X6    768 non-null int64
X7    768 non-null float64
X8    768 non-null int64
Y1    768 non-null float64
Y2    768 non-null float64
dtypes: float64(8), int64(2)
memory usage: 60.1 KB
```

	X1	X2	X3	X4	X5	X6	X7	X8	Y1	Y2
count	768.000000	768.000000	768.000000	768.000000	768.00000	768.000000	768.000000	768.00000	768.000000	768.000000
mean	0.764167	671.708333	318.500000	176.604167	5.25000	3.500000	0.234375	2.81250	22.307195	24.587760
std	0.105777	88.086116	43.626481	45.165950	1.75114	1.118763	0.133221	1.55096	10.090204	9.513306
min	0.620000	514.500000	245.000000	110.250000	3.50000	2.000000	0.000000	0.00000	6.010000	10.900000
25%	0.682500	606.375000	294.000000	140.875000	3.50000	2.750000	0.100000	1.75000	12.992500	15.620000
50%	0.750000	673.750000	318.500000	183.750000	5.25000	3.500000	0.250000	3.00000	18.950000	22.080000
75%	0.830000	741.125000	343.000000	220.500000	7.00000	4.250000	0.400000	4.00000	31.667500	33.132500
max	0.980000	808.500000	416.500000	220.500000	7.00000	5.000000	0.400000	5.00000	43.100000	48.030000

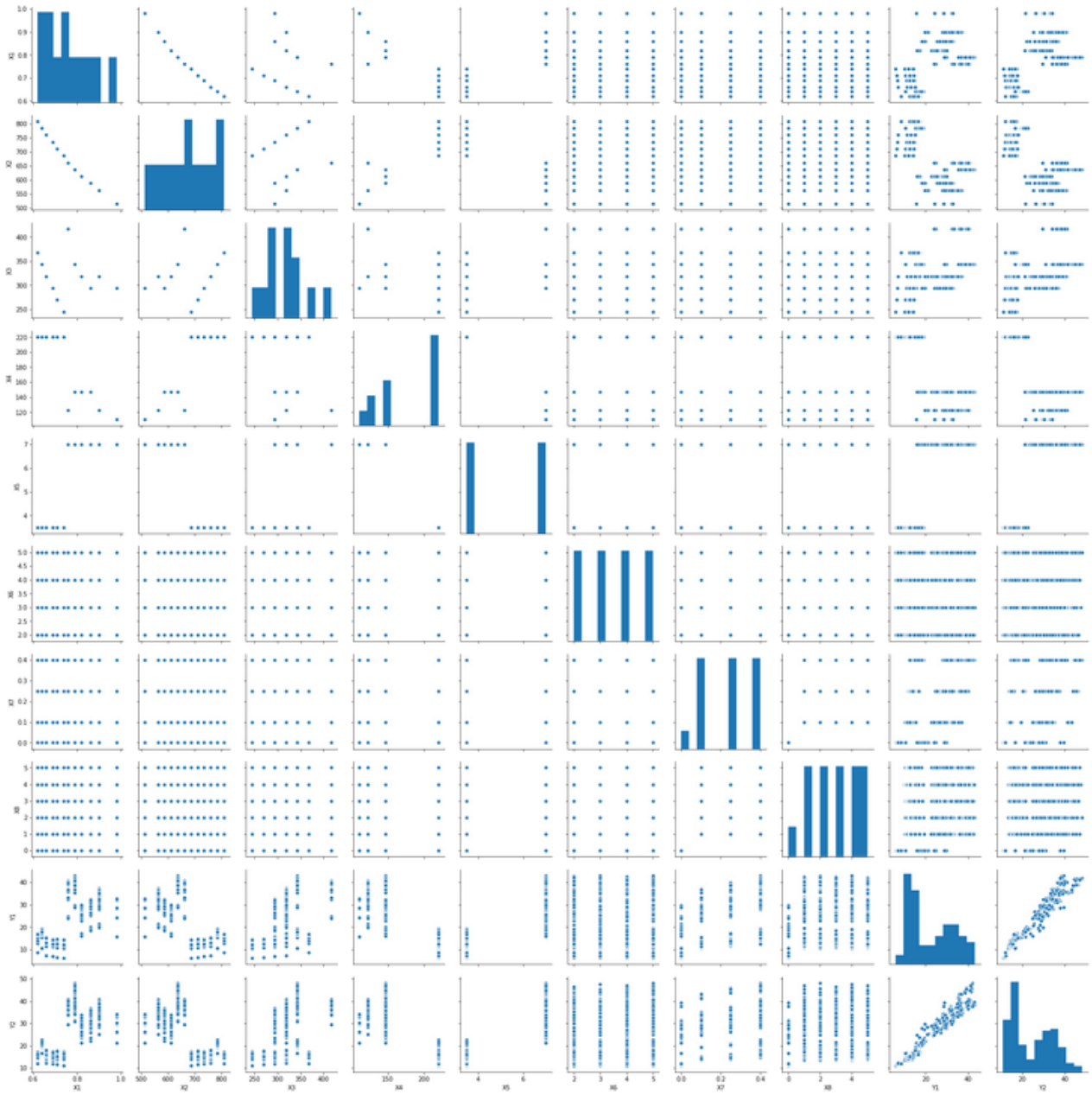
Zgodnie z oczekiwaniami posiadamy 768 rekordów. Wszystkie rekordy są uzupełnione. Warto zwrócić uwagę, iż parametry takie jak orientacja i dystrybucja zaszklenia określane są przez liczbę całkowitą.

Wartości zmiennych są ciężkie do zweryfikowania pod względem ilościowym, jednakże nie widzimy tu szczególnych odstępstw od norm (brak różnic rzędów wielkości).

Dane pobrane ze strony wyglądają na wcześniej obrobione, dlatego z tego miejsca można przejść do analizy danych.

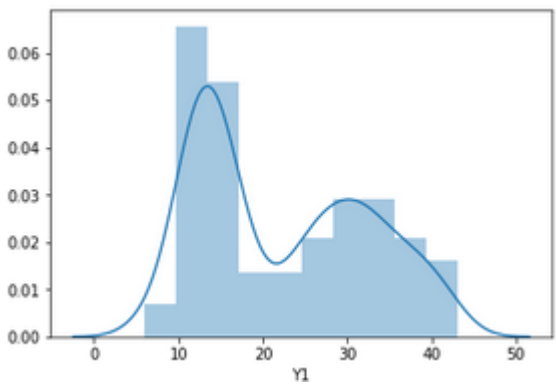
4. Analiza danych

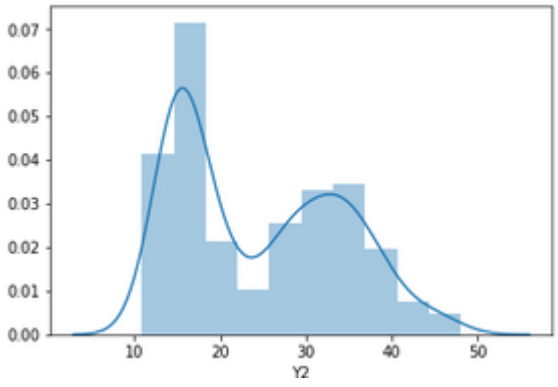
Przy pomocy metody pairplot z pakietu „seaborn" możemy zapoznać się ze wzajemnymi relacjami pomiędzy zmiennymi.



Pozwala nam to określić interesujące nas zmienne. Zgodnie z oczekiwaniami można dostrzec ,iż Y1 i Y2 są zależne od siebie.

Spójrzmy na ich rozkłady:

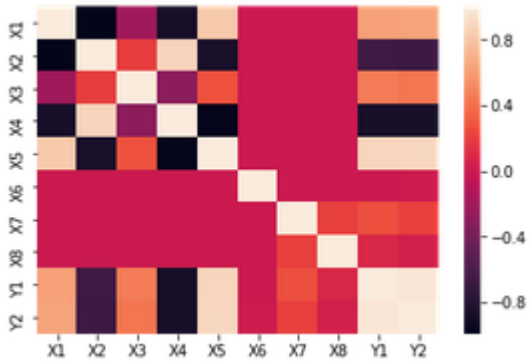




W celu dalszej analizy sprawdzamy wzajemne oddziaływania zmiennych. Do tego celu

wykorzystujemy współczynniki korelacji:

	X1	X2	X3	X4	X5	X6	X7	X8	Y1	Y2
X1	1.000000e+00	-9.919015e-01	-2.037817e-01	-8.688234e-01	8.277473e-01	0.000000	1.283986e-17	1.764620e-17	0.622272	0.634339
X2	-9.919015e-01	1.000000e+00	1.955016e-01	8.807195e-01	-8.581477e-01	0.000000	1.318356e-16	-3.558613e-16	-0.658120	-0.672999
X3	-2.037817e-01	1.955016e-01	1.000000e+00	-2.923165e-01	2.809757e-01	0.000000	-7.969726e-19	0.000000e+00	0.455671	0.427117
X4	-8.688234e-01	8.807195e-01	-2.923165e-01	1.000000e+00	-9.725122e-01	0.000000	-1.381805e-16	-1.079129e-16	-0.861828	-0.862547
X5	8.277473e-01	-8.581477e-01	2.809757e-01	-9.725122e-01	1.000000e+00	0.000000	1.861418e-18	0.000000e+00	0.889430	0.895785
X6	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000	0.000000e+00	0.000000e+00	-0.002587	0.014290
X7	1.283986e-17	1.318356e-16	-7.969726e-19	-1.381805e-16	1.861418e-18	0.000000	1.000000e+00	2.129642e-01	0.269842	0.207505
X8	1.764620e-17	-3.558613e-16	0.000000e+00	-1.079129e-16	0.000000e+00	0.000000	2.129642e-01	1.000000e+00	0.087368	0.050525
Y1	6.222719e-01	-6.581199e-01	4.556714e-01	-8.618281e-01	8.894305e-01	-0.002587	2.698417e-01	8.736846e-02	1.000000	0.975862
Y2	6.343391e-01	-6.729989e-01	4.271170e-01	-8.625466e-01	8.957852e-01	0.014290	2.075050e-01	5.052512e-02	0.975862	1.000000



Z przyczyn oczywistych nie uwzględniamy wartości na przekątnej. Szukamy natomiast zmiennych, które mają jak najwyższy współczynnik korelacji. Przykładowo można zaobserwować, że współczynnik korelacji pomiędzy Y1 a X5 jest bardzo wysoki ~0.9.

5. Modelowanie danych

Do celów modelowania nasz zbiór danych został podzielony na zbiór testowy oraz zbiór uczący w stosunku 2:8. Jako model wybraliśmy jeden z modeli liniowych czyli regresję liniową. Wpływ na to miała wcześniejsza analiza danych korelacji zmiennych.

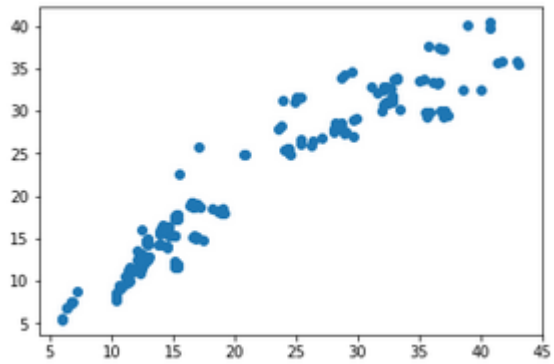
Otrzymana jakość modelu to 86.87. Korelacje z badaną zmienną to:

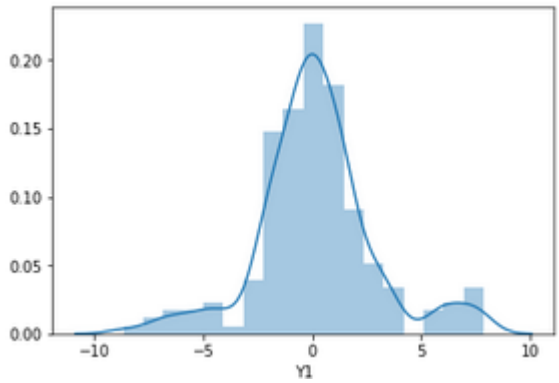
	Coeff
X1	-6.536622e+01
X2	2.232206e+11
X3	-2.232206e+11
X4	-4.464412e+11
X5	3.987993e+00
X6	-4.271881e-02
X7	2.034014e+01
X8	2.053594e-01

Dość dużym wpływem na nasz wynik jest udział zmiennej X1(względnej zwartości).

Teraz, gdy posiadamy już nauczony model nadszedł czas na wypróbowanie go na utworzonej wcześniej grupie testowej (20%).

Uzyskany model dla regresji liniowej oraz różnica między wartościami testowymi a uzyskaną predykcją:



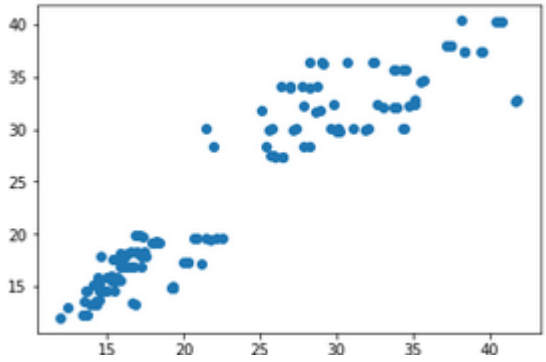


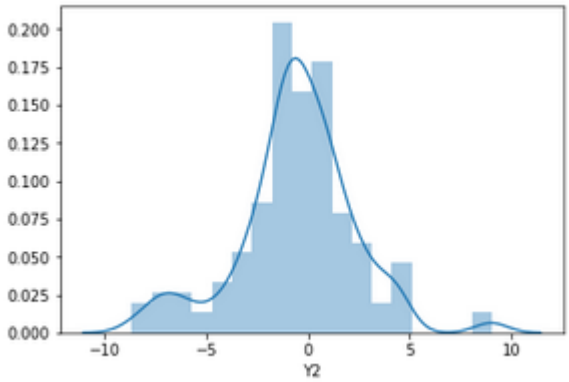
Błędy:

- mean_absolute_error = 1.8958237818679766
- mean_squared_error = 7.410019587128868
- sqrt(metrics.mean_squared_error) = 2.722135115516654

Poprawność modelu wynosi ~92.44%. Wcześniejszy wynik dla zbioru danych to ~91.59%. Jest to dobra poprawa biorąc pod uwagę to, aby upewnić się żeby nasz model nie był przeuczony.

Podobny proces tworzenia modelu dla Y2:





Błędy:

- mean_absolute_error= 2.1376398102532903
- mean_squared_error = 8.727775080688563
- sqrt(metrics.mean_squared_error) = 2.9542808059980628

6. Podsumowanie

Po dokładniejszej analizie naszych danych można wywnioskować parę rzeczy:

- Dostarczone dane zostały już przygotowane do analizy (zbiór danych nie potrzebował obróbki).
- W zbiorze argumentów nauczonego modelu nie występuje jeden wyraźnie dominujący czynnik wpływający na naszą funkcję celu, jednakże można zauważyć ich kolektywny wpływ. Widać tendencję 3 krotnie większego wpływu jednakże nie jest to wpływ zauważalnie większy względem innych (wszystkie wartości mają relatywnie duży udział). Dzięki temu można zaobserwować zmiany korelacji zmiennych w modelu oraz w

pierwotnych danych.

Niestety wybrany przez nas zbiór danych nie pozwala na zastosowanie bardziej zaawansowanych technik eksploracji danych. Jednym z istotnych wniosków naszej pracy jest to jak istotną rzeczą jest sam dobór danych. Istotne jest również zrozumienie sensu danych w celu wyciągnięcia wniosków. Operowanie na samych liczbach/danych bez znajomości kontekstu może utrudnić uzyskanie prawidłowego modelu (przykładowo zły dobór funkcji celu).