1. Cel

Celem ćwiczenia jest dobór odpowiedniego sposobu skalowania i jego przeprowadzenie dla różnych zestawów danych.

2. Materiały i metody

Zbiory danych:

a) iris.csv:
Zbiór danych dotyczy 3 gatunków irysów. Każdy irys jest opisany za pomocą 4 cech (długość i szerokość kielicha, długość i szerokość płatka) + informacja o gatunku.

W szczególności, dane te obejmują:

1. *Długość działki kielicha (sepal_length),*
2. *Szerokość działki kielicha (sepal_width),*
3. *Długość płatka (petal_length),*
4. *Szerokość płatka (petal_width),*
5. *Gatunek (species) – etykieta wskazująca na gatunek kwiatu: Setosa, Versicolor, Virginica.*

Charakterystyka danych:

- *Liczba próbek: 150 (w tym 50 próbek na każdy z trzech gatunków irysów).*

- *Podstawowe statystyki:*

  Z danych wynika, że średnie wartości poszczególnych cech morfologicznych różnią się, np. średnia długość działki kielicha wynosi około 5.84 cm, natomiast średnia szerokość płatka to 1.19 cm. Różnice te mogą świadczyć o tym, że cechy te mają zróżnicowaną wartość w klasyfikacji gatunków.

  Odchylenie standardowe (std) pokazuje zmienność pomiarów w obrębie poszczególnych cech – np. długość płatka (1.76) cechuje się większym rozrzutem niż szerokość działki kielicha (0.43), co może mieć znaczenie w dalszej analizie.

  Wartości minimalne (min) i maksymalne (max) pozwalają ocenić zakres rozkładu cech, np. szerokość płatka waha się od 0.1 cm do 2.5 cm.

Procedury przygotowania danych:

- *Opis statystyczny danych:*
  Wykorzystano metodę .describe() do uzyskania podstawowych statystyk opisowych takich jak średnia, odchylenie standardowe, wartości minimalne, kwartyle i wartości maksymalne dla każdej zmiennej.

- *Sprawdzenie braków danych:*
  Metoda .isnull().sum() została użyta w celu sprawdzenia brakujących wartości w danych. Wynik wskazuje, że nie ma brakujących wartości w żadnej z kolumn.

Techniki przekształcania danych:

- *Normalizacja Min-Max (0,1):*
  Przy użyciu klasy MinMaxScaler z zakresu (0, 1) przekształcono dane wejściowe, aby każda cecha mieściła się w tym przedziale.

- *Normalizacja Min-Max (-1, 1):*
  Przeprowadzono również normalizację w zakresie (-1, 1) za pomocą MinMaxScaler. Każda cecha po normalizacji mieści się w tym przedziale.

- *Standaryzacja:*
  Zastosowano StandardScaler, który przekształca cechy tak, aby miały średnią 0 i odchylenie standardowe 1.

b)Zad2_L1.csv :
Zbiór danych dotyczy serii zarejestrowanych pomiarów w postaci widm Ramana w czasie. Widmo Ramana to wykres intensywności rozproszonego promieniowania Ramana w funkcji różnicy częstotliwości w stosunku do promieniowania padającego. W pierwszym wierszu znajdują się jednostki (Wavenumber [cm^-1], Intensity [a.u.]), w drugim oznaczenia kolejnych pomiarów, ti. Pierwsza kolumna odnosi się do zakresu pomiarowego, w którym rejestrowane były widma (oś horyzontalna).Pozostałe kolumny odnoszą się do poszczególnych pomiarów.

W szczególności dane te obejmują:

Charakterystyka danych:

- *Liczba falowa: Jest to zmienna niezależna, która opisuje pozycję na osi X spektrum Ramana.*

- *Intensywność: Reprezentuje pomiary dla każdego punktu widma przy danej liczbie falowej, które odpowiadają wartości sygnału Ramana w jednostkach arbitralnych (a.u.).*
  Kolumny intensywności zawierają różne zestawy danych różne próbki).

Techniki analizy danych:

Normalizacja: Jednym z kluczowych kroków było znormalizowanie intensywności widm do amplitudy przy liczbie falowej 985 cm^{-1}. Normalizacja ta pozwala na porównanie różnych widm niezależnie od ich oryginalnej skali amplitudy.

- *Normalizacja była wykonana przez podzielenie wartości intensywności w każdym widmie przez wartość intensywności przy liczbie falowej najbliższej 985 cm^{-1}.*

Obsługa braków danych:
Na początku wykryto brakujące dane w kolumnie "Wavenumber", co mogło wymagać uzupełnienia (np. interpolacji) lub usunięcia tych wierszy z analizy, co jest typowym krokiem przygotowania danych (w pracy wyszczególniono odpowiednie wiersze i kolumny dokonując podziału zawierającego odpowiednie dane numeryczne:
wavenumber = data_Raman_spectra.iloc[:, 0]
intensity_spectra = data_Raman_spectra.iloc[:, 1:])

c)Zad3_L1.csv:
Zbiór danych dotyczy serii zarejestrowanych pomiarów w postaci widm FTIR w czasie.
Widmo w podczerwieni obrazuje intensywność widma w podczerwieni.
Pierwsza kolumna odnosi się do zakresu pomiarowego (oś pozioma), w którym rejestrowane były widma (Wavenumber [cm^-1]). Pozostałe kolumny (Absorbance [a.u.])) odnoszą się do widm rejestrowanych po upływie określonego czasu (oś wertykalna).

W szczególności dane te obejmują:

Charakterystyka danych:

Rodzaj danych: Numeryczne, z wartościami absorbancji w jednostkach bezwymiarowych, które mogą być dodatnie, ujemne lub równe zeru.

Zakres liczby fal: Od 525.0251 cm^{-1} do 3999.8810 cm^{-1}, co obejmuje szereg częstotliwości w podczerwieni.

Techniki wykorzystane do analizy danych:

Podstawowe statystyki: Analiza opisowa zbioru danych, obejmująca takie miary, jak średnia, odchylenie standardowe, wartości minimalne i maksymalne oraz kwartyle.

Normalizacja: W celu porównania danych z różnych czasów, zastosowano normalizację, która polegała na podzieleniu wartości absorbancji przez pole powierzchni pod wykresem. Normalizacja umożliwia porównywanie intensywności widm bez względu na ich całkowitą wartość, co jest istotne w analizach ilościowych.

Procedury przygotowania danych:

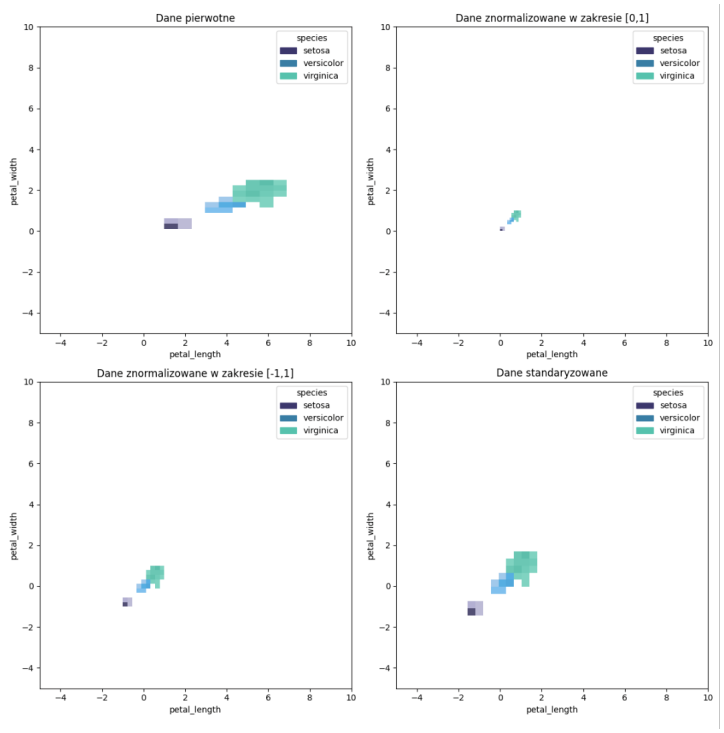
Zamiana przecinków na kropki: Konwersja formatów liczbowych (z przecinków na kropki) w celu zapewnienia prawidłowej interpretacji danych liczbowych.

Obsługa brakujących wartości: Przeprowadzono kontrolę na obecność brakujących danych, co wykazało, że brakujące wartości nie występują w zbiorze, co potwierdza integralność danych do analizy.

3. Wyniki i dyskusja

Zadanie 1.

Wykres zbiorczy, składający się z 4 wykresów, zależności długości płatka [cm] od szerokości płatka [cm] różnicowany na podstawie gatunku dla danych pierwotnych, znormalizowanych w zakresie [0,1], znormalizowanych w zakresie [1,1] oraz standaryzowanych.



Dane pierwotne

- *Wykres pokazuje, że istnieje wyraźna różnica między gatunkami, z wyraźnie zaznaczonymi grupami: setosa, versicolor i virginica.*
- *Irys setosa ma mniejsze długości i szerokości płatków w porównaniu do pozostałych gatunków.*

Dane znormalizowane do zakresu [0, 1]

- *Po normalizacji, różnice między gatunkami są nadal widoczne, ale wartości zostały przeskalowane do mniejszego zakresu, co może ułatwić dalszą analizę.*
- *Oś X i Y są w skali 0 do 1, co wpływa na interpretację danych, ale nie zmienia ich struktury.*

Dane znormalizowane do zakresu [-1, 1]

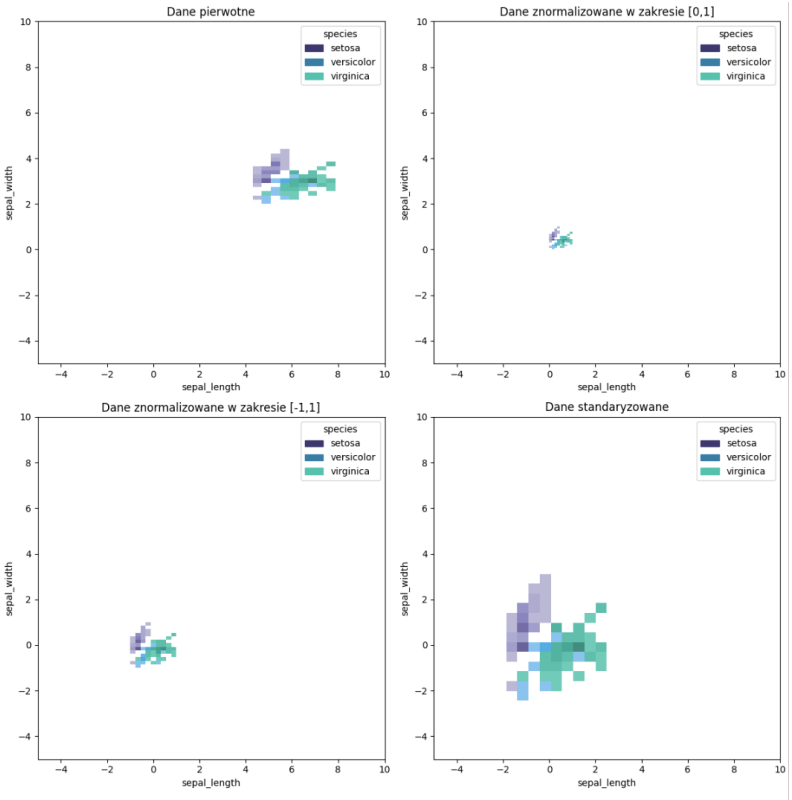
- *Rozkład danych jest symetryczny, a wartości są rozłożone równomiernie w zakresie od -1 do 1.*
- *Umożliwia to lepszą wizualizację w kontekście algorytmów opartych na symetrii.*

Dane standaryzowane

- *W tym przypadku wartości są rozłożone wokół zera, co pozwala na lepsze porównanie cech między gatunkami.*
- *Różnice między gatunkami są nadal wyraźne, a struktura danych wciąż pozostaje niezmieniona.*

Wykres zbiorczy, składający się z 4 wykresów, zależności długości kielicha [cm]

od szerokości kielicha [cm] różnicowany na podstawie gatunku dla danych pierwotnych, znormalizowanych w zakresie [0,1], znormalizowanych w zakresie [-1,1] oraz standaryzowanych.



Dane pierwotne

- *Podobnie jak w przypadku płatków, wyraźnie widoczne są różnice między gatunkami. Setosa ponownie ma mniejsze wartości, co sugeruje, że jest to gatunek charakteryzujący się mniejszymi rozmiarami.*

Dane znormalizowane do zakresu [0, 1]

- *Normalizacja ułatwia interpretację, zachowując jednocześnie różnice między gatunkami.*
- *Umożliwia to porównanie z innymi zestawami danych, które mogą mieć różne zakresy.*

Dane znormalizowane do zakresu [-1, 1]

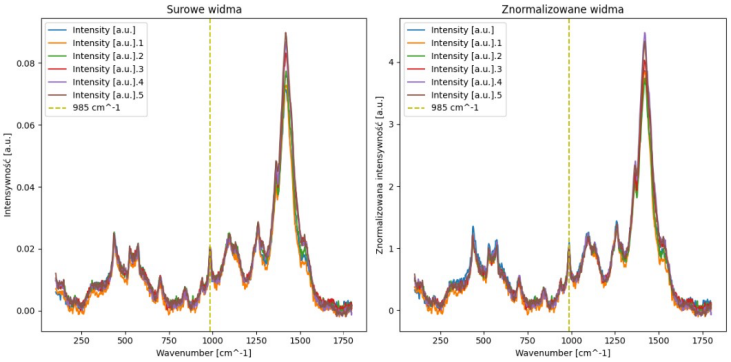
- *Wizualizacja po normalizacji do [-1, 1] ukazuje podobne cechy jak w przypadku danych znormalizowanych do [0, 1], ale z bardziej symetrycznym rozkładem.*

Dane standaryzowane

- *W przypadku standaryzacji, różnice między gatunkami stają się bardziej widoczne, a wartości mają bardziej homogeniczny charakter.*
- *Daje to możliwość lepszego porównania z danymi, które mogą mieć różne rozkłady.*

Zadanie 2.

wykres zbiorczy składający się z 2 wykresów: widma dla danych surowych oraz widma znormalizowane.



Normalizacja

Znormalizowano każde widmo (pomiar) przez podzielenie wszystkich wartości intensywności przez intensywność w punkcie odpowiadającym liczbie falowej 985 cm^{-1}. W wyniku tej operacji amplituda w 985 cm^{-1} przyjęła wartość 1 dla każdego z widm, co umożliwia łatwe porównanie kształtu widm w różnych zakresach liczby falowej.

Przykład normalizacji

Wiersze przed normalizacją zawierały różne wartości amplitudy dla widm w zależności od pomiaru, np.:

- *Dla pierwszego pomiaru intensywność przy liczbie falowej 985 cm^{-1} wynosiła ok. 0.0185.*
- *Po normalizacji intensywność w tym samym miejscu wynosi 1, a wszystkie pozostałe wartości zostały przeskalowane względem tej wartości.*

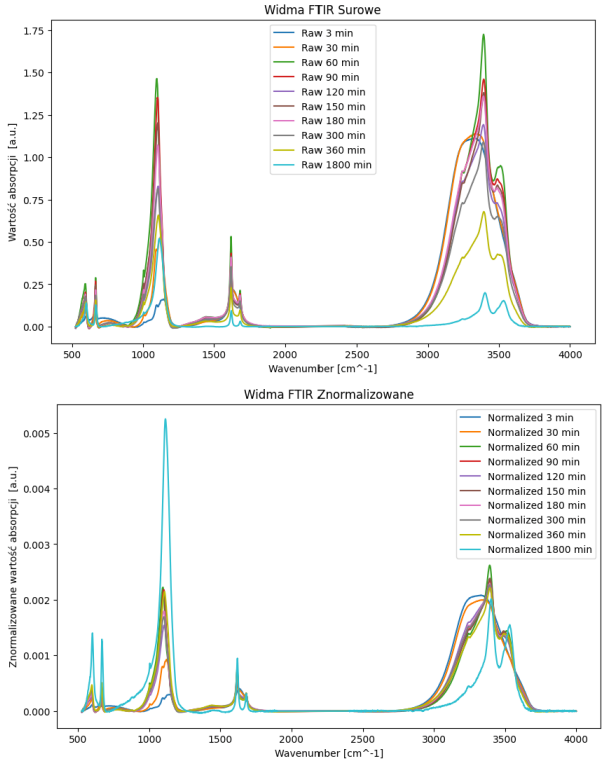
Efekty normalizacji

Normalizacja widm pozwoliła na zniwelowanie różnic w amplitudach pomiarów, zachowując jednocześnie istotne informacje o względnych zmianach intensywności w innych zakresach liczby falowej. Dzięki temu:

- *Możliwe jest porównanie kształtu widm między sobą bez względu na różnice w ogólnej intensywności sygnału.*
- *Wartości intensywności, które były wyższe niż wartość przy 985 cm^{-1}, są teraz większe niż 1, a te niższe — mniejsze od 1.*

Zadanie 3.

wykres zbiorczy składający się z 2 wykresów: widma dla danych surowych oraz widma znormalizowane.



Normalizacja:

Widma te obrazują zmiany intensywności absorpcji (Absorbance [a.u.]) w funkcji liczby falowej (Wavenumber [cm^-1]) w różnych odstępach czasu. Normalizacja miała na celu przeskalowanie każdego z widm do porównywalnych wartości poprzez normalizację do pola powierzchni pod wykresem (integralna wartość absorpcji).

Po normalizacji uzyskano dane, które zostały przeskalowane, co pozwala na lepsze porównanie różnych pomiarów, szczególnie w kontekście zmian intensywności absorpcji w czasie. Wartości te zostały znormalizowane poprzez podzielenie każdej wartości intensywności przez pole powierzchni pod wykresem dla danego widma. Wyniki normalizacji wykazują zmiany w amplitudzie, jednak zachowują ogólny kształt pierwotnych widm.

Uwagi:

Stabilność danych po normalizacji:

- *Wartości intensywności po normalizacji są znacznie mniejsze niż przed procesem, co wskazuje, że pole powierzchni pod wykresem było stosunkowo duże dla wszystkich widm.*
- *Wartości intensywności dla każdego widma są teraz znormalizowane w taki sposób, że różnice między nimi są bardziej widoczne. Dzięki temu można łatwiej identyfikować subtelne zmiany w czasie.*

Zmiany intensywności w czasie:

- *W danych można zaobserwować stopniowe zmiany intensywności absorpcji w czasie. Wczesne widma (np. „3 min", „30 min") różnią się od późniejszych widm (np. „1800 min"), co może sugerować, że zachodzą pewne procesy chemiczne lub fizyczne, które zmieniają charakter materiału w miarę upływu czasu.*

Znaczenie normalizacji:

- *Normalizacja do pola powierzchni pod wykresem jest kluczowa w takich analizach, ponieważ różne pomiary mogą różnić się całkowitą intensywnością z powodu różnorodnych warunków eksperymentalnych. Bez normalizacji porównanie widm byłoby trudne, ponieważ wartości mogłyby zależeć od takich czynników jak grubość próbki, intensywność źródła promieniowania itp.*

4. Podsumowanie

Zadanie 1:

Różnorodność gatunków: Analiza wykresów pokazuje, że gatunki irysów różnią się między sobą pod względem długości i szerokości kielichów oraz płatków. Setosa jest wyraźnie oddzielona od pozostałych gatunków, co sugeruje, że jest to dobrze zdefiniowana kategoria.

Wpływ normalizacji i standaryzacji: Wprowadzenie normalizacji oraz standaryzacji nie zmienia relacji między danymi, ale może poprawić interpretację wyników w kontekście analizy statystycznej i algorytmów uczenia maszynowego.

Zadanie 2:

Normalizacja do amplitudy przy liczbie falowej 985 cm^{-1} pozwala na eliminację różnic w skali pomiarów, umożliwiając porównanie widm w sposób względny.

- *Fluktuacje i różnice pomiędzy widmami po normalizacji mogą być wynikiem zmienności w badanym materiale, błędów pomiarowych lub innych czynników zewnętrznych.*
- *Wartość normalizacji jest kluczowa w analizie danych widmowych, ponieważ umożliwia łatwiejsze porównanie i identyfikację wzorców pomiędzy różnymi pomiarami. Zadanie 3:*

Równoważność widm: Normalizacja sprawiła, że możliwe jest bezpośrednie porównywanie widm zarejestrowanych w różnych odstępach czasowych. Dzięki temu łatwiej można zauważyć zmiany w charakterystyce materiału w trakcie trwania eksperymentu.

Zastosowanie w praktyce: Normalizacja do pola powierzchni pod wykresem jest efektywnym narzędziem w analizie danych spektroskopowych, szczególnie gdy badane są zmiany zachodzące w czasie lub w różnych próbkach, co pozwala na eliminację wpływu zmiennych zewnętrznych.

Równoważność widm: Normalizacja sprawiła, że możliwe jest bezpośrednie porównywanie widm zarejestrowanych w różnych odstępach czasowych. Dzięki temu łatwiej można zauważyć zmiany w charakterystyce materiału w trakcie trwania eksperymentu.

- *Zastosowanie w praktyce: Normalizacja do pola powierzchni pod wykresem jest efektywnym narzędziem w analizie danych spektroskopowych, szczególnie gdy badane są zmiany zachodzące w czasie lub w różnych próbkach, co pozwala na eliminację wpływu zmiennych zewnętrznych.*

5. Bibliografia

https://scikit-learn.org/0.21/documentation.html

https://seaborn.pydata.org/

https://pandas.pydata.org/docs/

https://www.youtube.com/watch?v=6eJHk8JYK2M

https://www.youtube.com/watch?v=wipZ-olfoqU

https://www.youtube.com/watch?v=bqhQ2LWBheQ