Sprawozdanie z PCA

Techniki eksploracji danych wielowymiarowych

Jakub Bożek

285665

Bioinformatyka, II rok

4 czerwca 2024

Spis treści

1 Wprowadzenie

1.1 Cel ćwiczenia

Celem ćwieczeń było zapoznanie się z metodą PCA(Principal Component Analysis lub po polsku analiza głównych składowych). Która jest jedną z metod uczenia maszynowego nienadzorowanego.

1.2 Krótkie wprowadzenie do metody PCA

Metoda PCA polega na redukcji wymiarowości danych za pomocą przekształceń algebraicznych. Wynikiem czego będzie możliwość przystępniejszego zobrazowania i/lub ich analizy.

1.3 Oprogramowanie

Wszystkie elementy sprawozdania zostały stworzone przy pomocy języka programowania Python 3.12.3 z pomocą bibliotek:

Numpy 1.26.4

Matplotlib 3.8.4

Pandas 2.2.2

Scikit-learn 1.4.2

Seaborn 0.13.2

2 Metodologia

2.1 Standaryzacja danych

Pierwszym elementem pracy z danymi jest ich standaryzacja. Odbywa się to poprzez zastosowanie wzoru:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{1}$$

gdzie:

z_{ij} = zmienna ustandaryzowana

x_{ij} = zmienna niestandaryzowana

μ_j = średnia dla obserwacji

σ_j = odchylenie standardowae dla obserwacji

Po autoskalowaniu każda zmienna posiada średnią równą 0 oraz odchylenie standardowe równe 1. Takie działanie sprawia, że nie bierzemy pod uwagę jednostek. Dodatkowo każda zmienna może mieć inny zakres wartości. W niektórych przypadkach może to wprowadzać nam błędną intuicję względem wyglądu naszych danych.

2.2 Macierz kowariancji

Macierz kowariancji, jak wskazuje nazwa, jest zbiorem danych zawierającym wartości kowariancji pomiędzy poszczególnymi zmiennymi.

Aby ją policzyć, stosuję się wzór:

$$Cov(X) = \begin{bmatrix} Var(x_1) & \dots & Cov(x_1, x_n) \\ \vdots & \ddots & \vdots \\ Cov(x_n, x_1) & \dots & Var(x_n) \end{bmatrix} \tag{2}$$

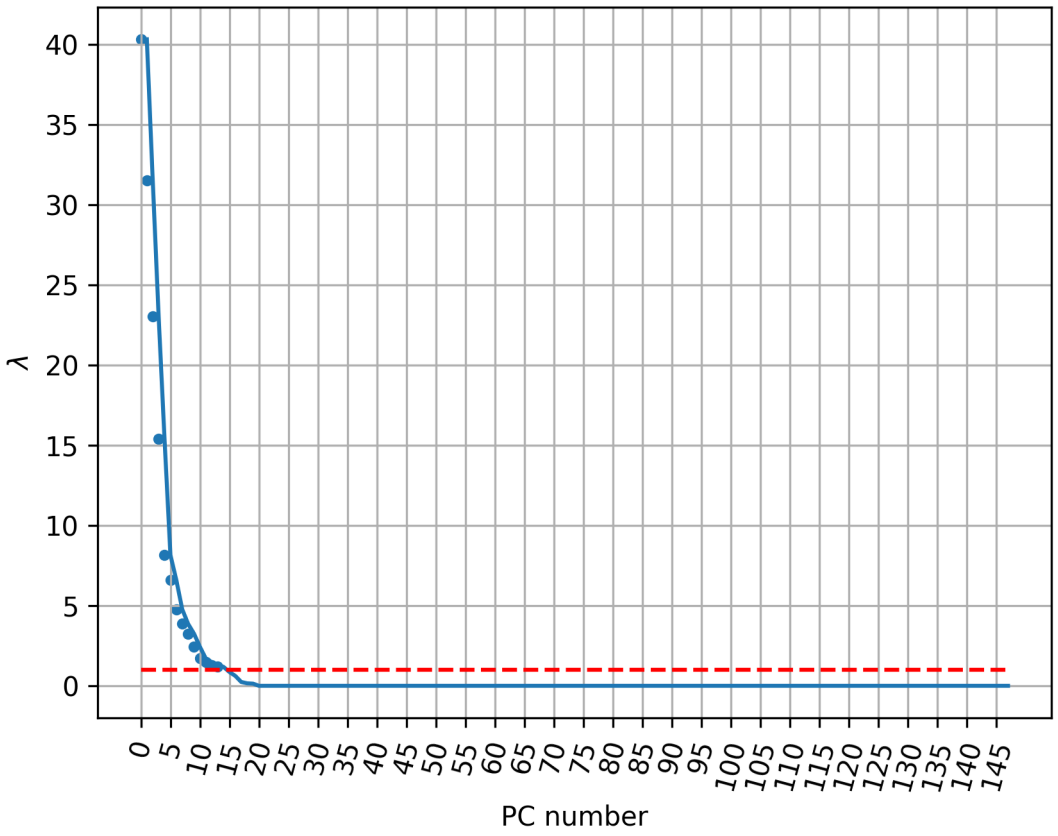
Taką macierz można przedstawić za pomocą mapy ciepła. Wtedy można zauważyć, jak poszczególne zmienne korelują ze sobą.

2.3 Wartości i wektory własne głównych składowych

Z macierzy kowariancji należy następnie policzyć wartości i wektory własne. W tym celu skorzystałem z metody eig. Zwraca ona wartości posortowane w kolejności malejącej względem wartości własnych, nazywanymi również eigenvalues. Jeśli by tak nie było, należałoby samemu wykonać sortowanie w powyższy sposób. Jest to ważne, ponieważ wartości własne wyznaczają istotność zmiennej w zbiorze danych.

Tabela 1: Zestawienie wartości własnych oraz procent wyjaśnianej wariancji przez 10 pierwszych głownych składowych

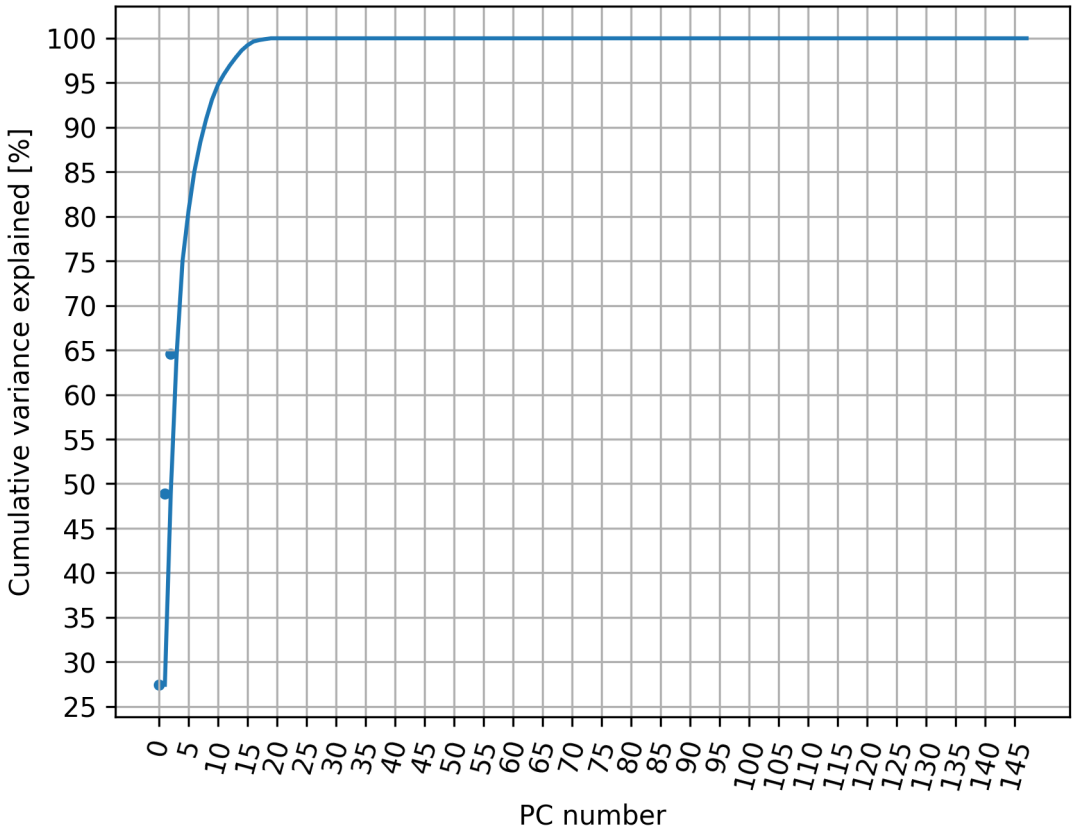
	Wartości własne	Procent wyjaśnianej wariancji
PC1	40.32	27.43
PC2	31.53	21.45
PC3	23.05	15.68
PC4	15.39	10.47
PC5	8.15	5.55
PC6	6.59	4.49
PC7	4.77	3.24
PC8	3.86	2.63
PC9	3.25	2.21
PC10	2.43	1.65



Rysunek 1: Wykres osypiska(wartości własnych) dla głównych składowych

Czerwona linia odcina główne składowe na poziomie 1. Jest to oczekiwana praktyka, jednak na potrzeby zwięzłości sprawozdania będę operować jedynie na 3 pierwszych głównych

składowych.



Rysunek 2: Wykres kumulatywnego procentu wyjaśnianej wariancji przez główne składowe

W tym przypadku, dane będą opisywały $\sim 65\%$ wariancji całości. Normalnie byłoby to prawdopodobnie $> 90\%$.

2.4 Macierz ładunków czynnikowych

Macierz ładunków czynnikowych służy do przedstawienia, które zmienne mają największe znaczenie, tzn. wnoszą największą część informacji poszczególnym z głównych składowych. Zakłada się, że znaczenie mają te, dla których wartość bezwzględna jest większa od 0.7.

Aby ją policzyć, wykorzystujemy wzór:

$$M_{lad\_czyn} = \overrightarrow{v}\sqrt{\lambda} \tag{3}$$

gdzie

$\overrightarrow{v}$ = wektory własne

λ = wartości własne

Tabela 2: Typ korelacji zmiennych dla głównych składowych

	PC1		PC2
MolWt	+	BCUT2D_MWHI	−
ExactMolWt	+	BCUT2D_MWLOW	+
FpDensityMorgan1	−	BCUT2D_MRHI	+
FpDensityMorgan2	−	BalabanJ	+
FpDensityMorgan3	−	BertzCT	−
BCUT2D_LOGPHI	+	Chi3n	−
Chi0n	+	Chi4n	−
Chi0v	+	SMR_VSA7	−
Chi1n	+	SlogP_VSA3	+
Chi1v	+	SlogP_VSA6	−
Chi2v	+	VSA_EState1	+
Kappa1	+	VSA_EState10	−
Kappa2	+	VSA_EState6	−
PEOE_VSA11	−	FractionCSP3	+
PEOE_VSA3	−	NumAromaticCarbocycles	−
PEOE_VSA7	+	NumAromaticRings	−
SMR_VSA3	−	NumHAcceptors	+
TPSA	−	NumRotatableBonds	+
EState_VSA7	+	RingCount	−
VSA_EState2	−	fr_benzene	−
VSA_EState3	−	fr_halogen	−
NHOHCount	−		
NOCount	−		
NumAromaticHeterocycles	−		
NumHDonors	−		
MolLogP	+		
MolMR	+		
fr_Ar_N	−		
fr_pyridine	−		

	PC3
SPS	+
AvgIpc	+
PEOE_VSA8	+
SMR_VSA4	+
SlogP_VSA4	+
NumAliphaticHeterocycles	+
NumAliphaticRings	+
NumSaturatedHeterocycles	+
NumSaturatedRings	+
fr_COO	−
fr_COO2	−

Jeżeli korelacja jest dodatnia(+) to wtedy wraz ze wzrostem wartości, wartość skorelowana z nią będzie również rosła. W przypadku korelacji ujemnej(−) wzrost wartości oznacza spadek wartości skorelowanej.

2.5 Macierz wartości czynnikowych

Macierz wartości czynnikowych, jak możnaby uważać, powinna być liczona z poprzednio liczonej macierzy ładunków czynnikowych. Jednak jest to jedynie błędnie działająca intuicja, powieważ liczy się ją w sposób następujący:

$$M_{wart\_czyn} = Z\overrightarrow{v} \tag{4}$$

gdzie

Z = macierz wartości autoskalowanych

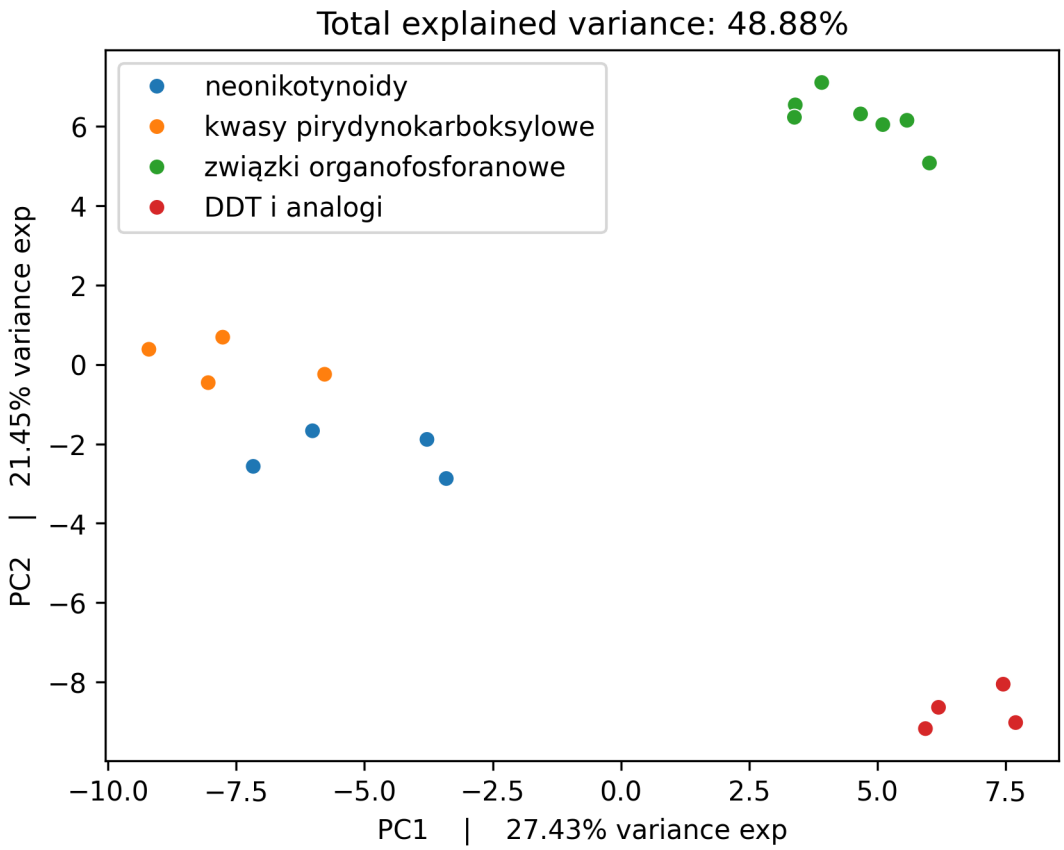
$\overrightarrow{v}$ = wektory własne

Przedstawia ona wartości naszego zbioru obserwacji dla każdej z wybranych głównych składowych. Te dane posłużą później do przedstawienia na wykresie zawierającym pary głównych składowych.

3 Wyniki

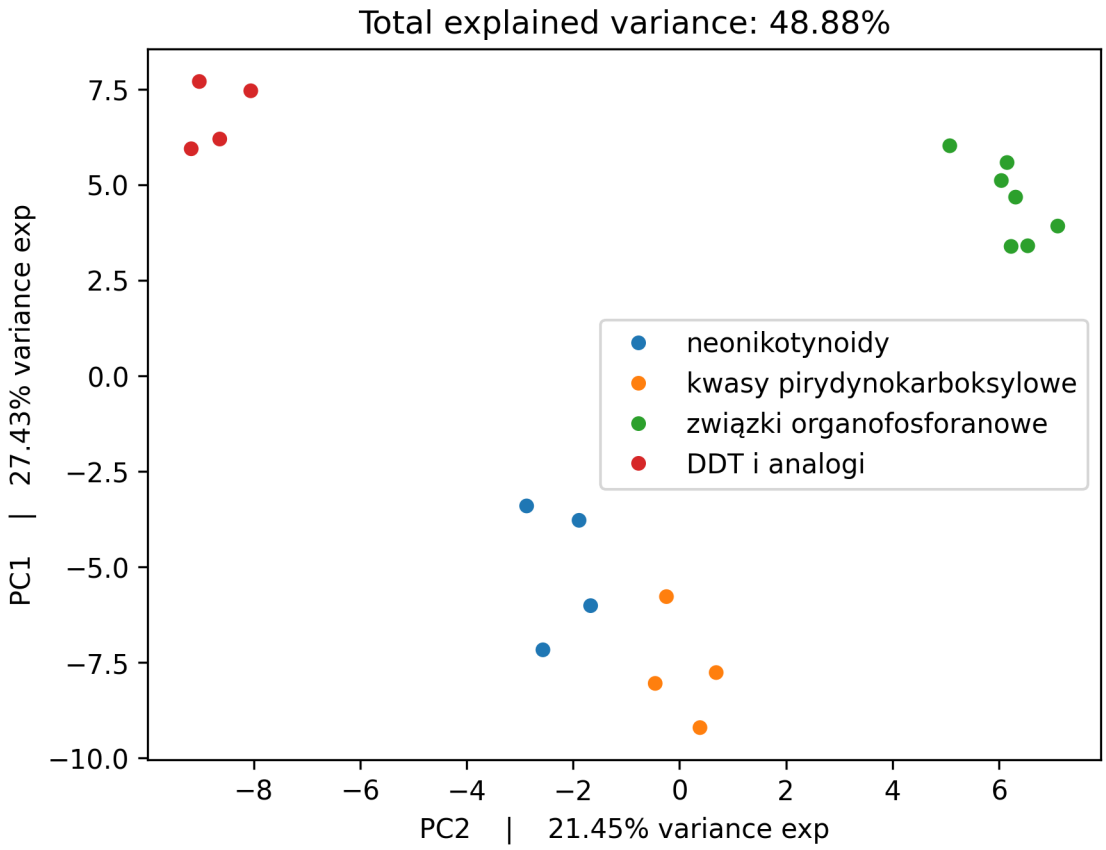
Po wykonaniu kroków opisanych w poprzednim rozdziale można przedstawić wyniki na wykresach.

PC1:PC2



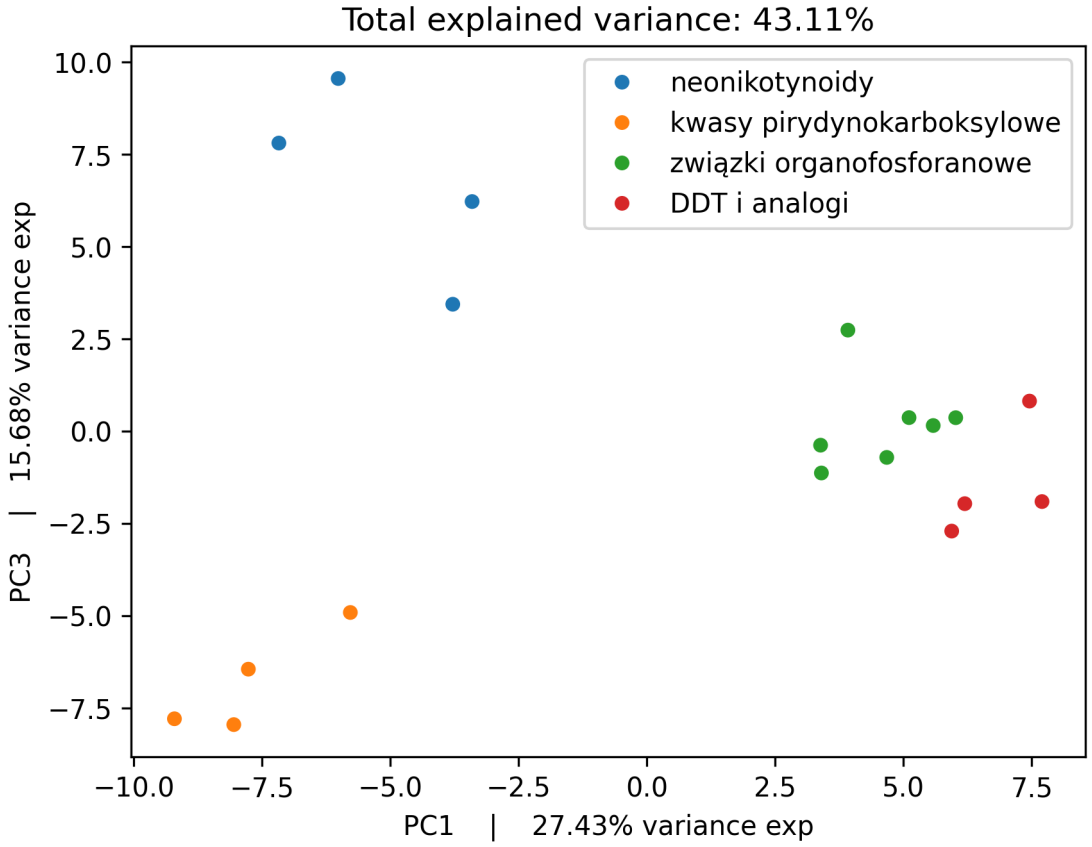
Rysunek 3: Wykres PC1 do PC2

PC2:PC1



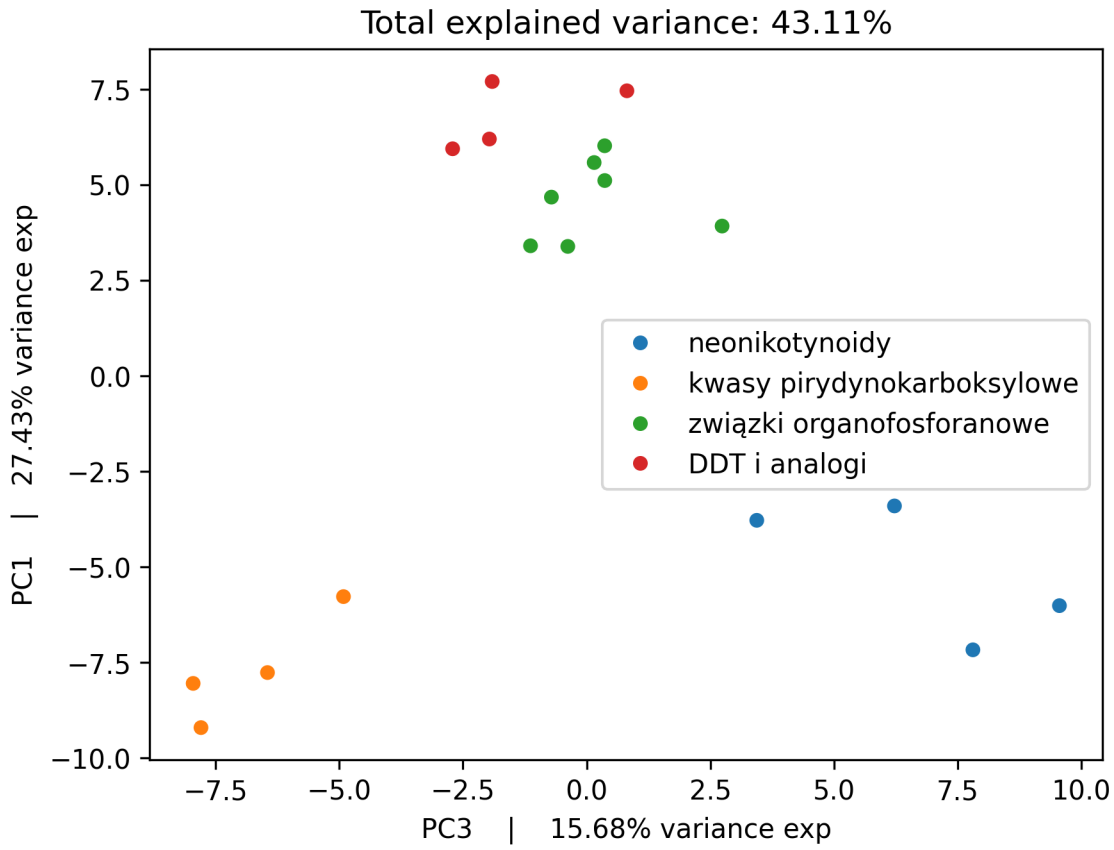
Rysunek 4: Wykres PC2 do PC1

PC1:PC3



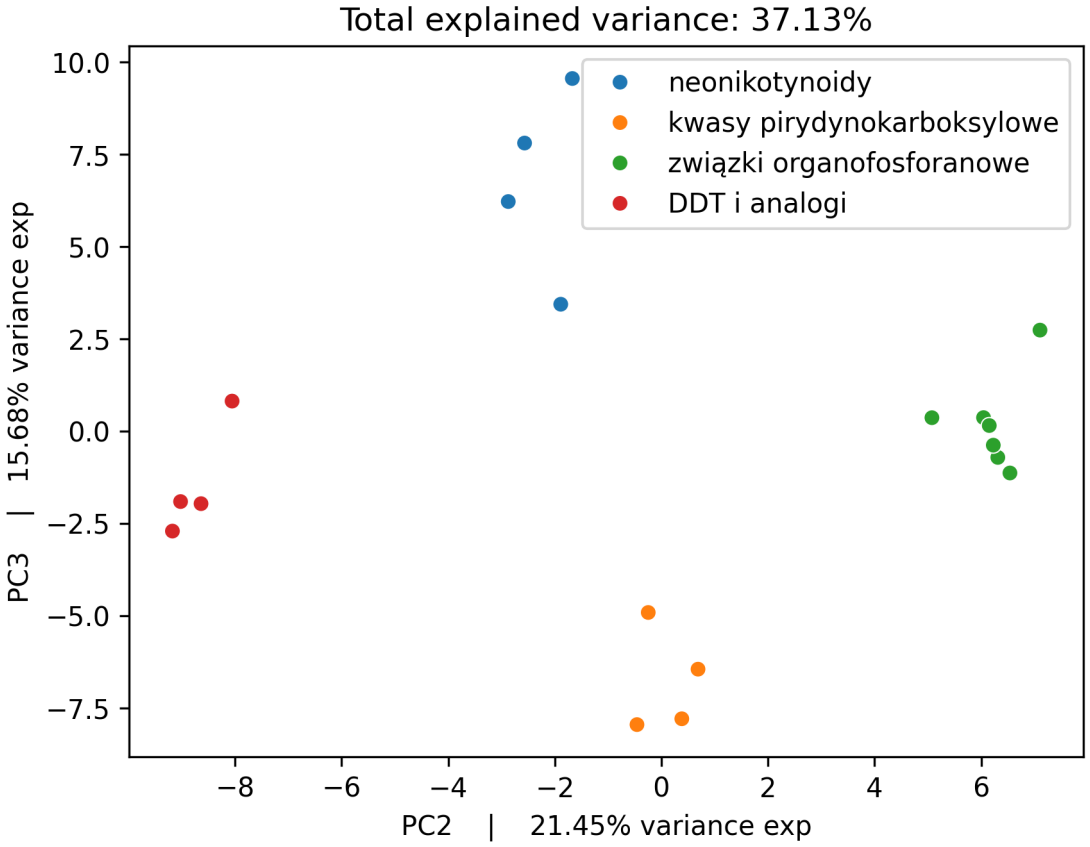
Rysunek 5: Wykres PC1 do PC3

PC3:PC1



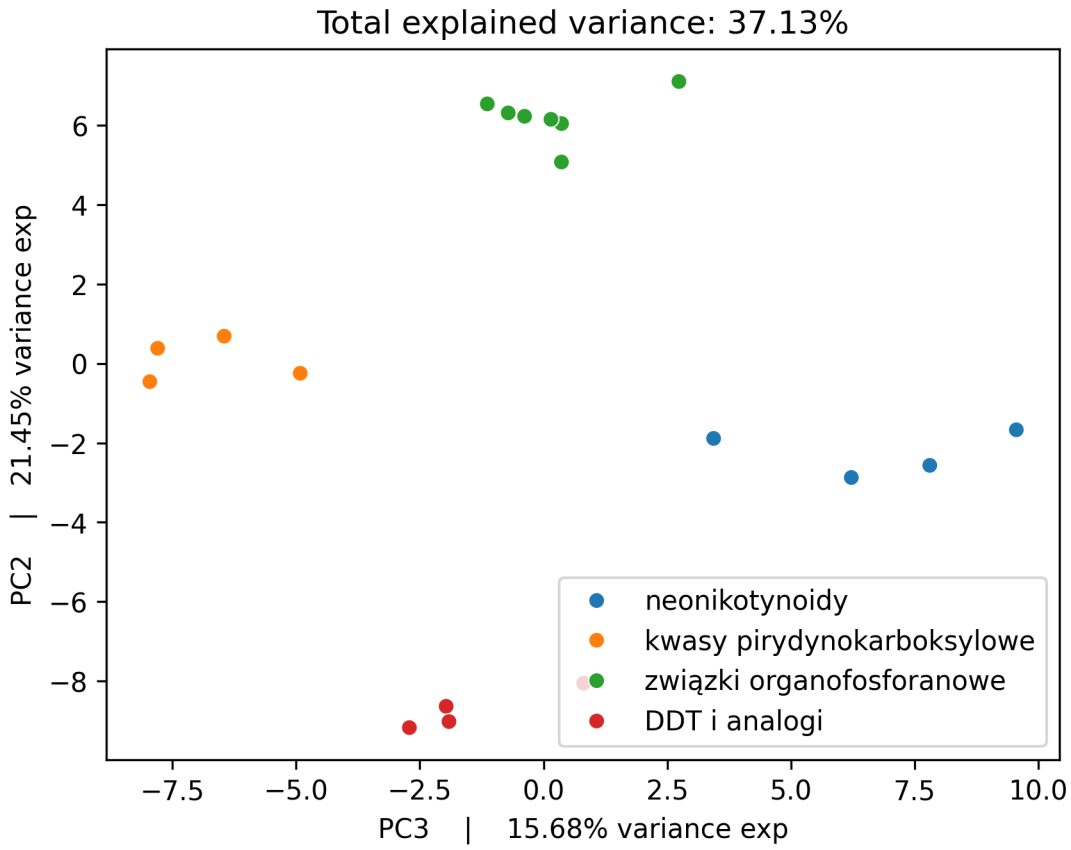
Rysunek 6: Wykres PC3 do PC1

PC2:PC3



Rysunek 7: Wykres PC2 do PC3

PC3:PC2



Rysunek 8: Wykres PC3 do PC2

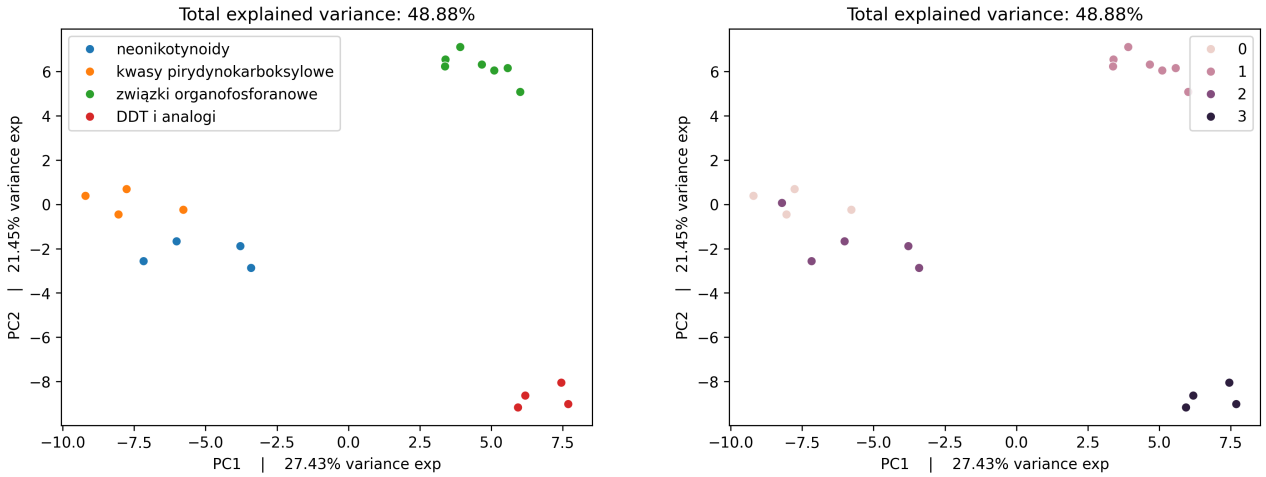
4 Obserwacje

Jak można zauważyć, powstałe wykresy oddzielają nam 4 grupy pestycydów od siebie w sposób bardzo czytelny. W większości przypadków są one skumulowane oraz oddzielone od siebie. Jest to widoczne najlepiej na zestawieniu PC2 oraz PC3. Warto zaznaczyć, że dzieje się tak pomimo najniższej sumy opisywanej wariancji.

5 Dyskusja i wnioski

Spośród 4 grup, które zostały włączone do PCA tzn. (neonikotynoidy, kwasy pirydynokarboksylowe, związki organofosforowe, DDT i analogi) wszystkie są w podobnej odległości od siebie, dlatego ciężko byłoby stwierdzić, które się od siebie najbardziej różnią.

Zgrupowania są raczej zwarte, oprócz neonikotynoidów, które wizualnie wykazują się najluźniejszą strukturą.



(a) grupowanie oryginalne

(b) grupowanie KMeans

Rysunek 9: Zestawienie wykresów względem różnego grupowania

PCA jest bardzo użyteczną techniką używaną w procesie analizy danych. Może ona służyć zarówno w celu przystępniejszego wyświetlenia danych, jak i jako dane wejściowe do algorytmów uczenia maszynowego. Oprócz tego metoda PCA może być wykorzystywana do odszumiania zdjęć, co jest według mnie bardzo interesujące.

W tym przypadku znana jest przynależność poszczególnych związków do grup, dlatego można było je zaznaczyć na wykresach. Jednak jeśli to nie byłoby wiadome, możnaby połączyć metodę PCA z metodą grupowania np. KMeans.

Muszę przyznać, że na wykresie z grupowaniem KMeans jest jedna wartość więcej, niż powinna być. Po analizie kodu muszę przyznać, że nie wiem, skąd ten błąd się wziął. Jeśli przymknąć na niego oko, otrzymamy dokładnie takie samo grupowanie.