

Sprawozdanie z PCA

Techniki eksploracji danych wielowymiarowych

Jakub Bożek

285665

Bioinformatyka, II rok

4 czerwca 2024

Spis treści

1	Wprowadzenie	1
1.1	Cel ćwiczenia	1
1.2	Krótkie wprowadzenie do metody PCA	1
1.3	Oprogramowanie	1
2	Metodologia	1
2.1	Standaryzacja danych	1
2.2	Macierz kowariancji	2
2.3	Wartości i wektory własne głównych składowych	2
2.4	Macierz ładunków czynnikowych	4
2.5	Macierz wartości czynnikowych	5
3	Wyniki	5
4	Obserwacje	11
5	Dyskusja i wnioski	11

1 Wprowadzenie

1.1 Cel ćwiczenia

Ćwiczenia z przedmiotu *Techniki eksploracji danych wielowymiarowych* miały za zadanie przedstawić nam oraz przećwiczyć pod nadzorem różne metody uczenia maszynowego nienadzorowanego. W tym przypadku jest to PCA.

1.2 Krótkie wprowadzenie do metody PCA

Metoda PCA polega na redukcji wymiarowości naszych danych za pomocą przekształceń algebraicznych. Kiedy to się nam uda, będziemy mogli w dużo przystępniejszy sposób zobrazować nasze dane.

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 niestandardyzowana

μ_j = średnia dla obserwacji

σ_j = odchylenie standardowe 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, stosujemy wzór:

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

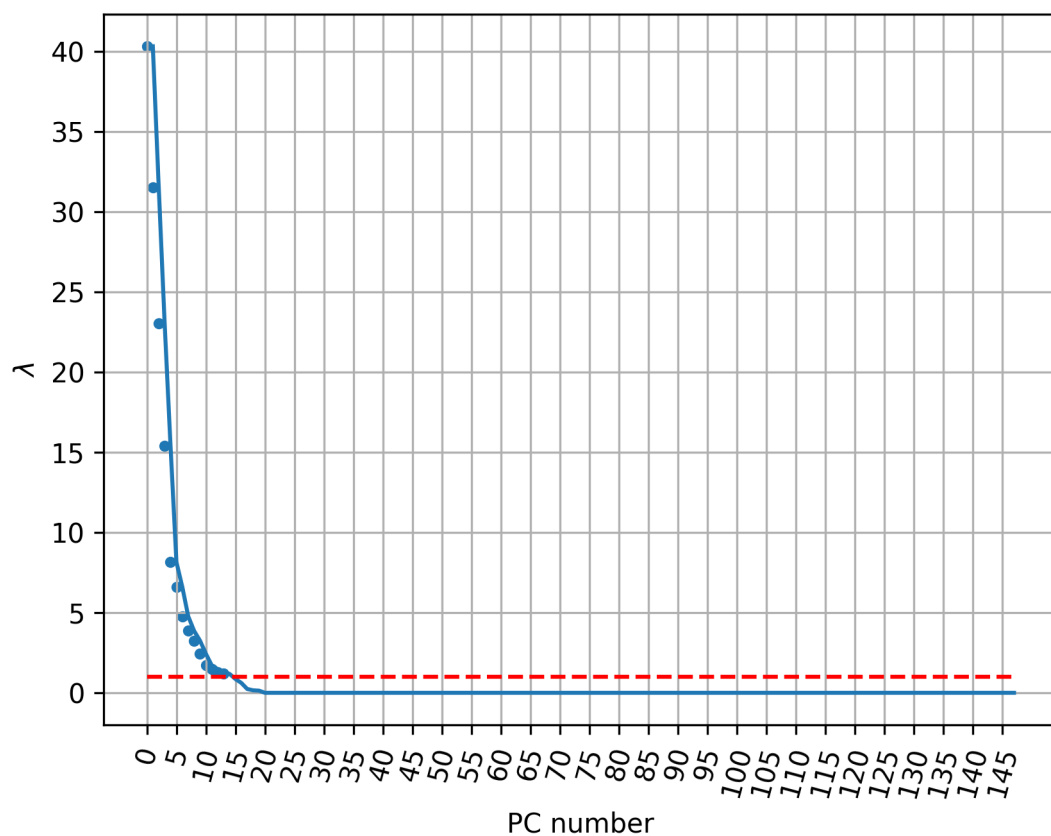
Taką macierz można przedstawić za pomocą [mapy ciepła](#). Wtedy jesteśmy w stanie zauważyć jak poszczególne zmienne korelują ze sobą.

2.3 Wartości i wektory własne głównych składowych

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

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

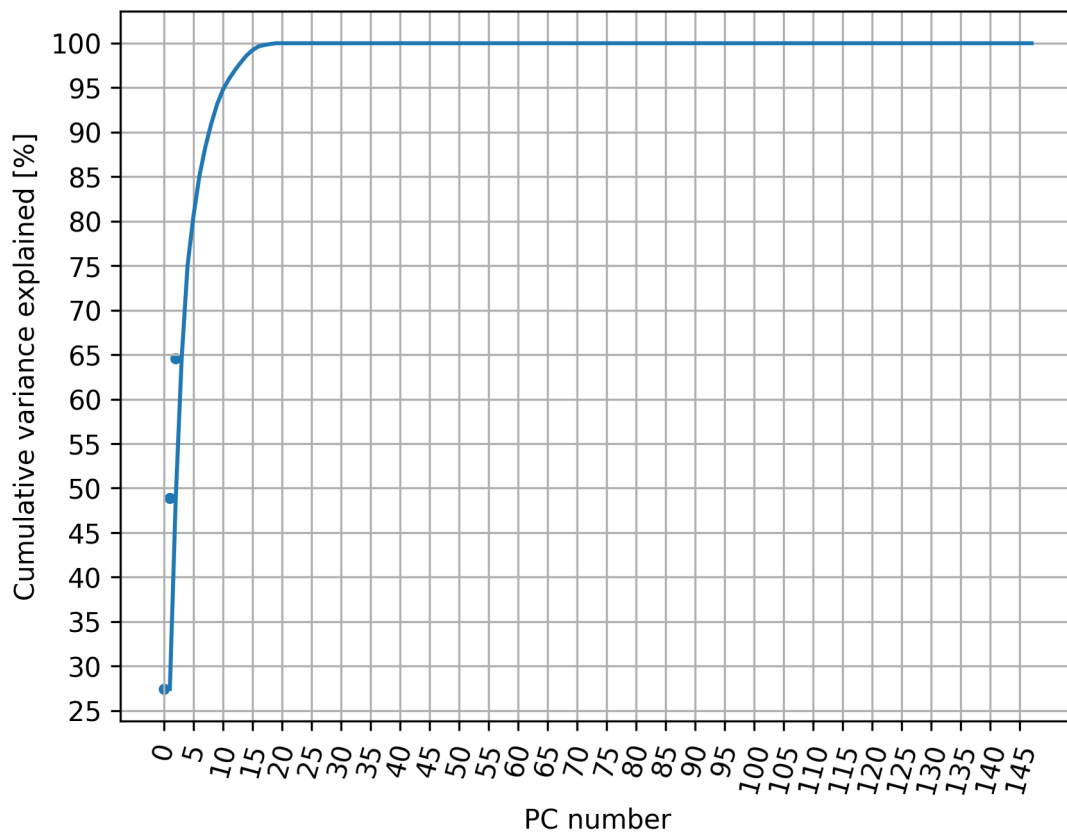
	Wartości własne	Procent wyjaśnianej wariancji
PC1	40.32	27.43
PC2	31.53	21.45
PC3	23.05	15.68
PC4	15.39	10.47
PC5	8.15	5.55
PC6	6.59	4.49
PC7	4.77	3.24
PC8	3.86	2.63
PC9	3.25	2.21
PC10	2.43	1.65



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

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

głównych składowych.



Rysunek 2: Wykres kumulatywnego procentu wyjaśnianej wariancji przez główne składowe

W tym przypadku, jak widać, dane, które zostawimy, 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:

$$M_{lad.czyn} = \vec{v} \sqrt{\lambda} \quad (3)$$

gdzie

\vec{v} = wektory własne

λ = wartości własne

2.5 Macierz wartości czynnikowych

Macierz wartości czynnikowych, jak podpowiada nam intuicja, powinna być liczona z macierzy ładunków czynnikowych. Jeśli ktoś dałby sobie za to rękę uciąć to, no cóż, nie miałby teraz ręki.

Wzór jest następujący:

$$M_{wart_czyn} = Z \vec{v} \quad (4)$$

gdzie

Z = macierz wartości autoskalowanych

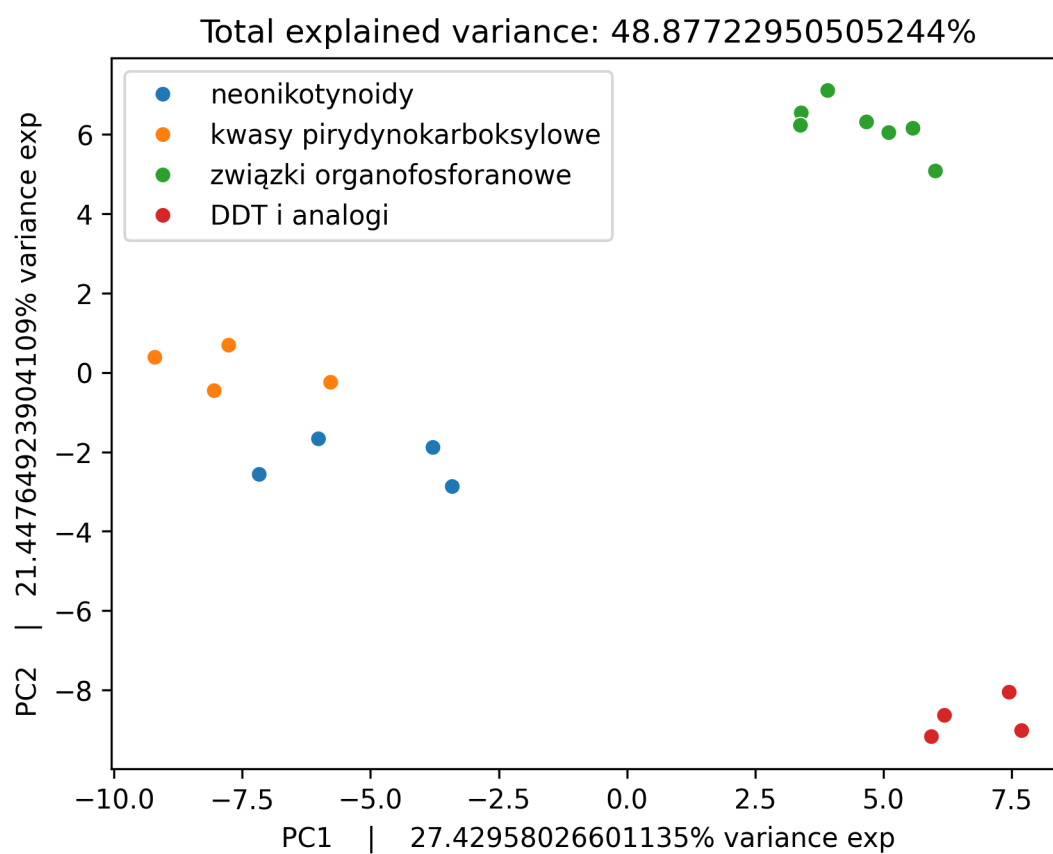
\vec{v} = wektory własne

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

3 Wyniki

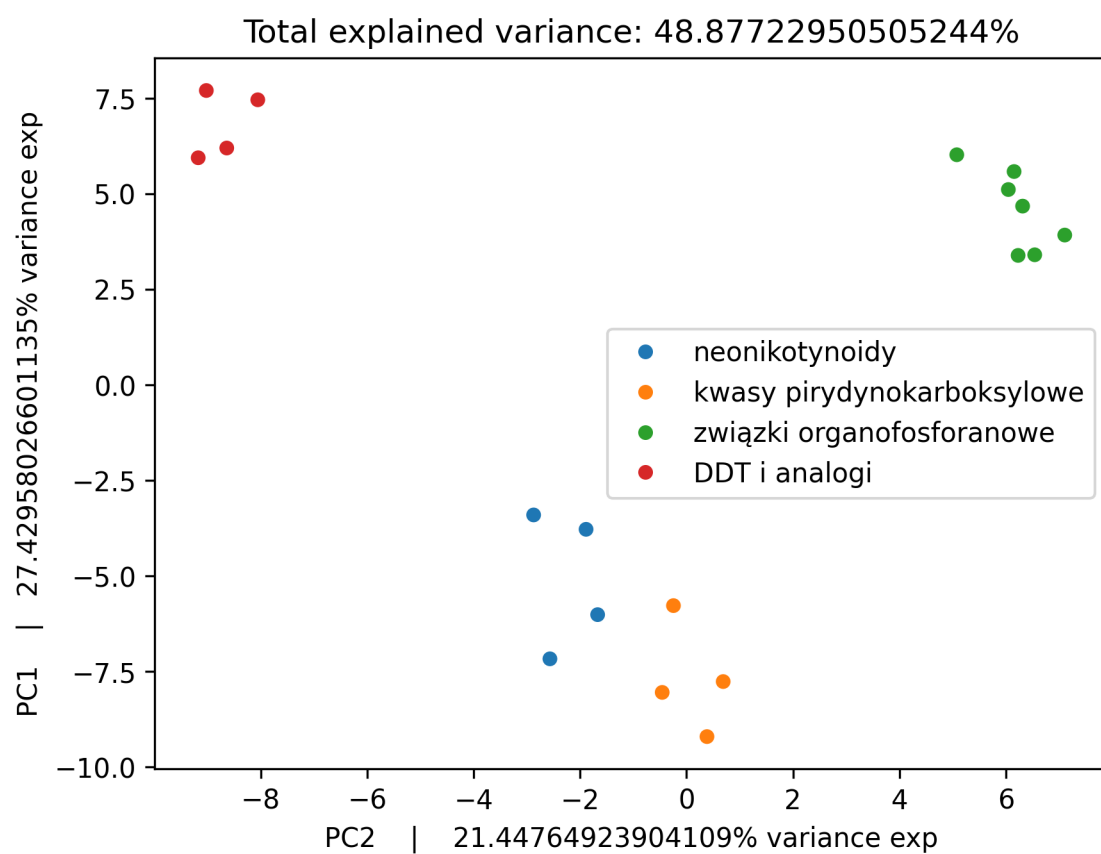
Po wykonaniu kroków opisanych w poprzednim rozdziale jesteśmy w stanie przedstawić wyniki na wykresach.

PC1:PC2



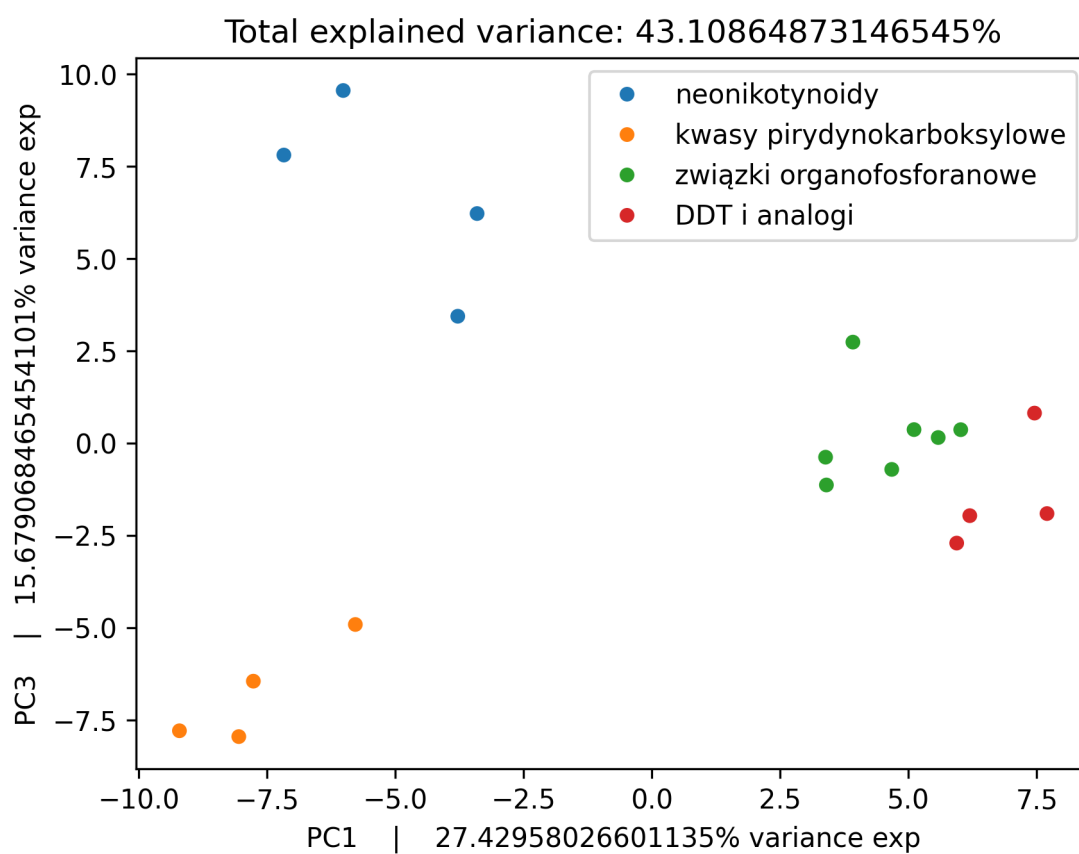
Rysunek 3: Wykres PC1 do PC2

PC2:PC1



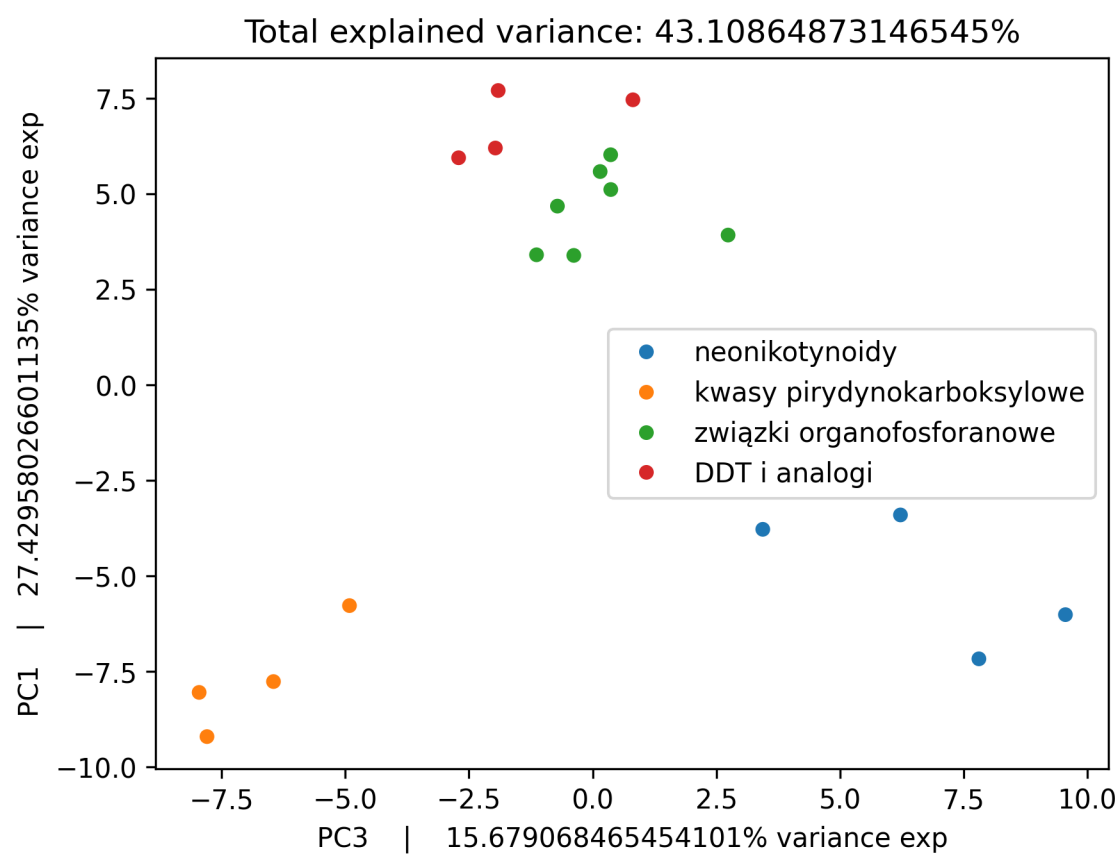
Rysunek 4: Wykres PC2 do PC1

PC1:PC3



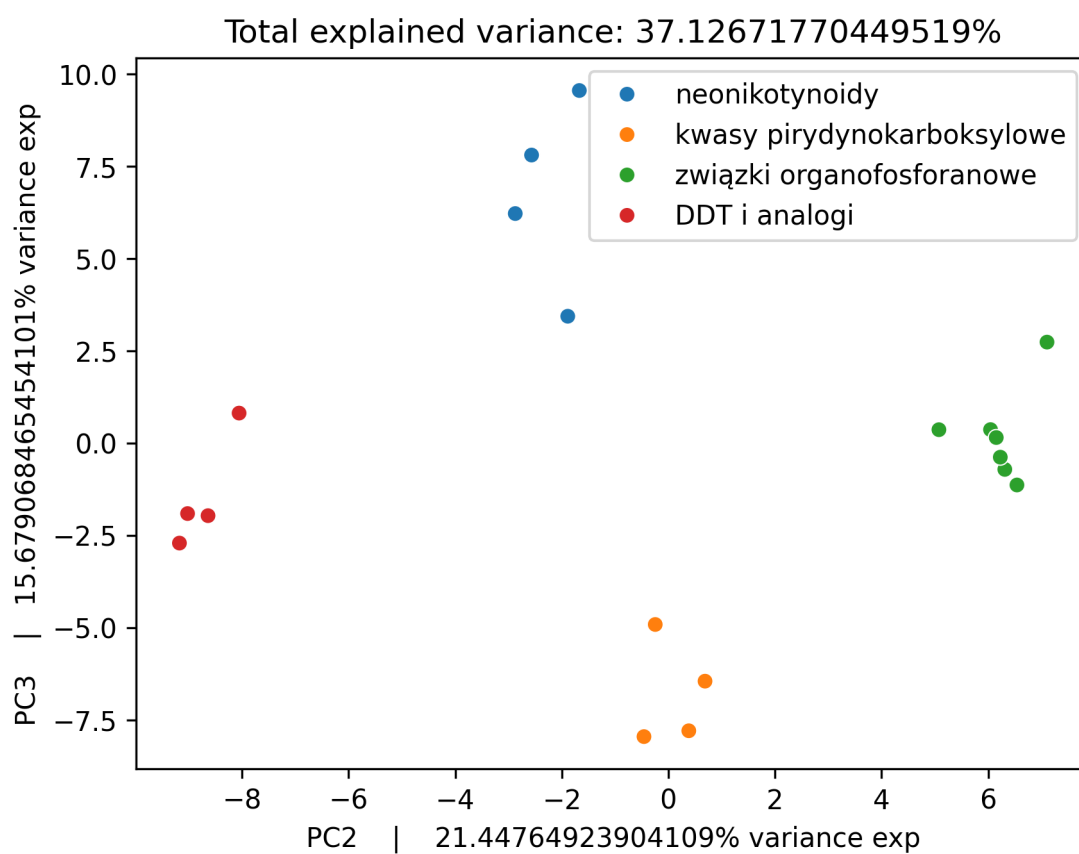
Rysunek 5: Wykres PC1 do PC3

PC3:PC1



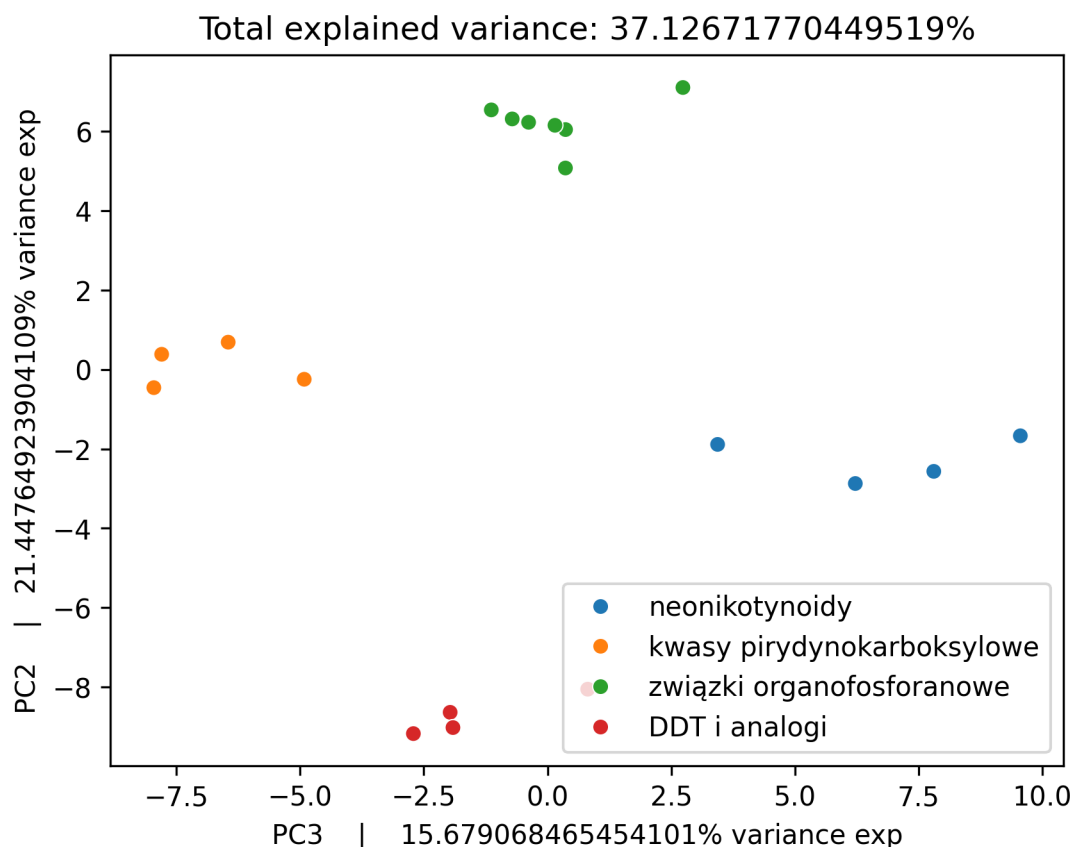
Rysunek 6: Wykres PC3 do PC1

PC2:PC3



Rysunek 7: Wykres PC2 do PC3

PC3:PC2



Rysunek 8: Wykres PC3 do PC2

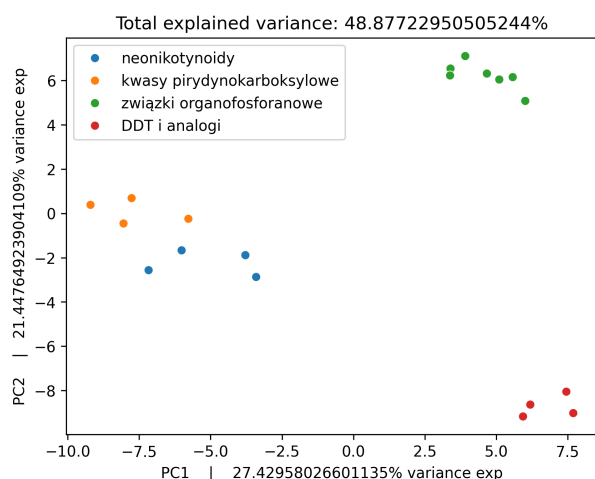
4 Obserwacje

Jak jesteśmy w stanie 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.

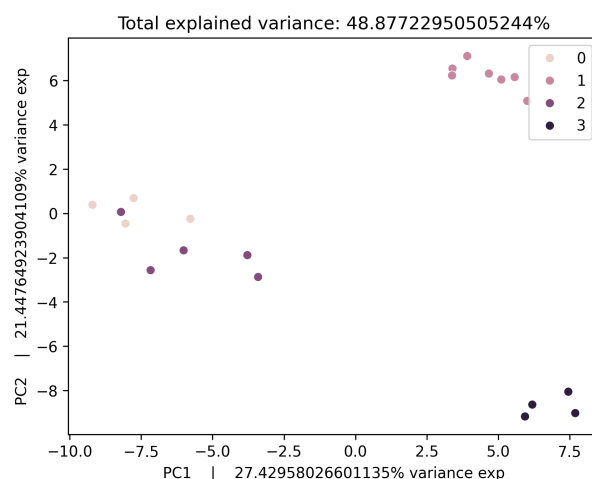
5 Dyskusja i wnioski

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 nadzorowanego. Oprócz tego metoda PCA może być wykorzystywana do odsumowania zdjęć, co jest według mnie bardzo interesujące.

W naszym przypadku wiemy, do jakich grup należą poszczególne związki, dlatego mogliśmy je zaznaczyć na wykresach. Jednak jeśli nie wiedzielibyśmy tego, można by połączyć metodę



(a) grupowanie oryginalne



(b) grupowanie KMeans

Rysunek 9: Zestawienie wykresów względem różnego grupowania

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 przymknijemy na niego oko, otrzymamy dokładnie takie samo grupowanie.