Sprawozdanie z HCA

Techniki eksploracji danych wielowymiarowych

Jakub Bożek

285665

Bioinformatyka, II rok

7 maja 2024

Spis treści

1 Wprowadzenie

1.1 Opis problemu

Analiza podobieństwa związków chemicznych dla małej ilości informacji jest relatywnie prosta. Problem zaczyna się w momencie posiadania dużej ilości danych, tak jak w naszym przypadku.

1.2 Cel analizy

Znalezienie podobieństw pomiędzy wybranymi 20 pestycydami i ich uszeregowanie w mniejsze, łatwiejsze do interpretacji grupy.

1.3 Hipoteza badawcza

Wybrane związki będą wykazywały podobieństwo między sobą tworząc tzw. klastry, inaczej zbiory zawierające elementy podobne.

1.4 Oprogramowanie

Wszystkie elementy sprawozdania zostały stworzone przy pomocy języka programowania Python 3.10.12 z pomocą bibliotek:

Numpy 1.26.2

Matplotlib 3.8.2

Pandas 2.1.3

Scipy 1.12.0

Rdkit 2023.09.5

2 Opis zbioru danych

Dla podanych nazw związków znalezione zostały ich kody Smiles, który jest uniwersalnym przedstawieniem danej molekuły za pomocą, w uproszczeniu, ułożonych szeregowo pierwiastków.

Następnie z pomocą biblioteki Rdkit[1.4] znaleźliśmy deskryptory, które są liczbowym opisem molekuły.

Tabela z zestawionymi deskryptorami posiada wymiar 20 x 210(liczba zmiennych x liczba obserwacji). Dane są w większości danymi zmiennoprzecinkowymi(float) lub całkowitymi(int).

3 Metoda analizy

W celu zredukowania wymiaru naszych danych usunęliśmy obserwacje o zerowym wkładzie. Jeżeli kolumna wartości posiada wariancję równą 0 to znaczy, że wszystkie wartości są takie same. Po tej operacji nasza tabela skurczyła się do rozmiarów 20 x 147.

Oprócz tego możnaby wykonać usunięcie wszystkich wierszy oraz kolumn w których występują puste pola, które uniemożliwiłyby nam dalszą analizę. W tym przypadku nie było to konieczne ze względu na brak ich występowania.

Następnie przeprowadziliśmy autoskalowanie danych, co określane się wzorem:

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

gdzie:

z_{ij} = zmienna ustandaryzowana

x_{ij} = zmienna niestandaryzowana

μ_j = średnia dla obserwacji

σ_j = odchylenie standardowae dla obserwacji

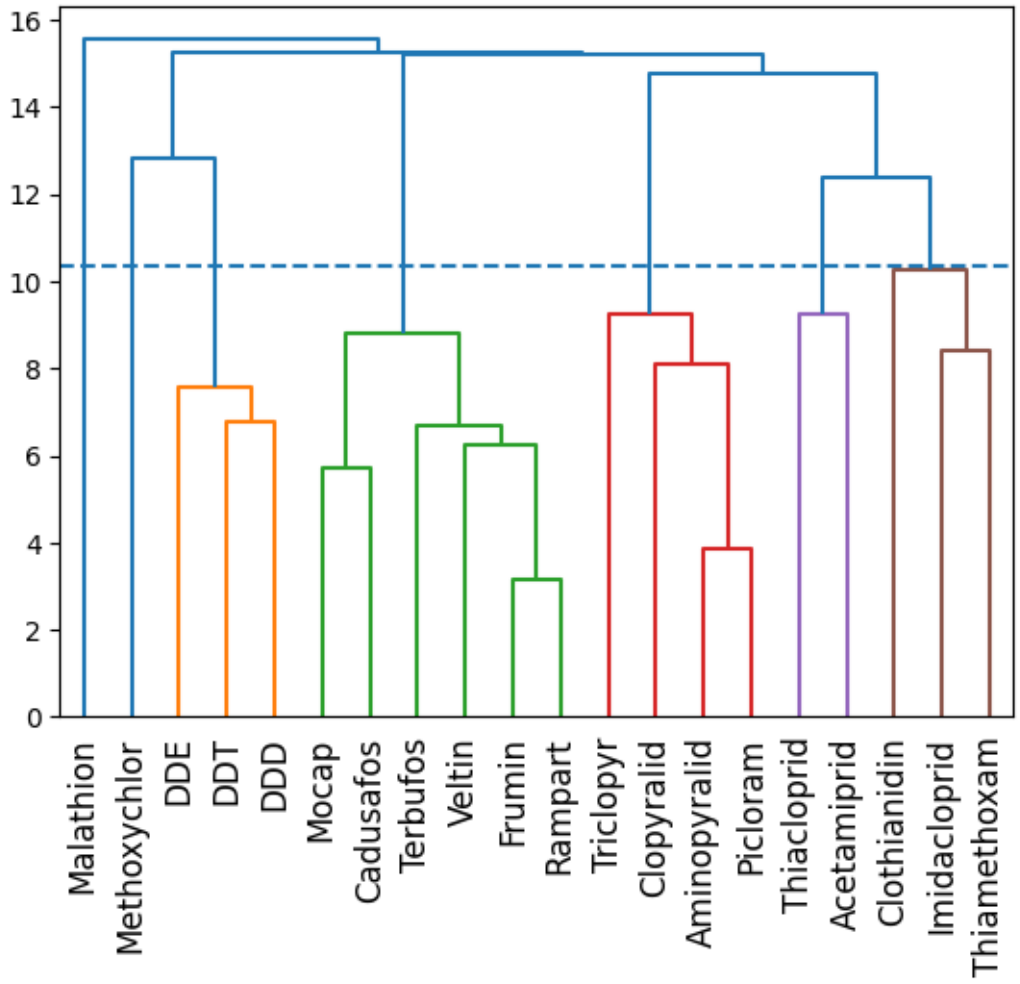
W wyniku autoskalowania, inaczej standaryzacji, zmienna uzyskuje wartość oczekiwaną 0 oraz odchylenie standardowe 1.

Dla tak przetransformowanych danych przeprowadziliśmy HCA metodą pojedynczego wiązania(najbliższego sąsiada) oraz całkowitego wiązania(najdalszego sąsiada). W celu dogłębnejszej analizy dla każdej metody wiązania wybieramy po 2 metryki odległości, Euklidesową oraz Manhattan

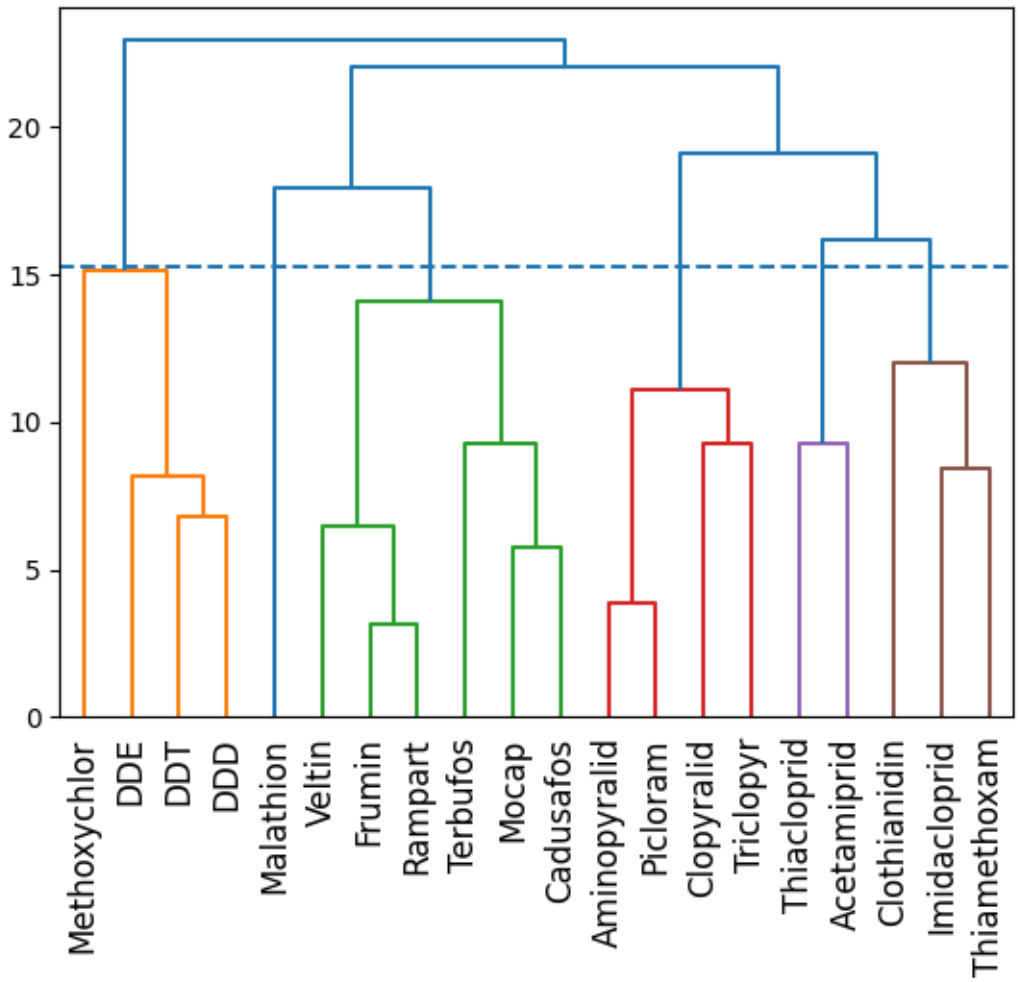
$$d_{Euk}(x,y) = \sqrt{\sum_i (x_i - y_i)^2} \qquad d_{Manh}(x,y) = \sum_i |x_i - y_i| \tag{2}$$

Oprócz tego, dodatkowo, przeprowadziliśmy HCA dla danych przed autoskalowaniem. Ja wybrałem metodę pojedynczego wiązania oraz metrykę odległości Euklidesowej.

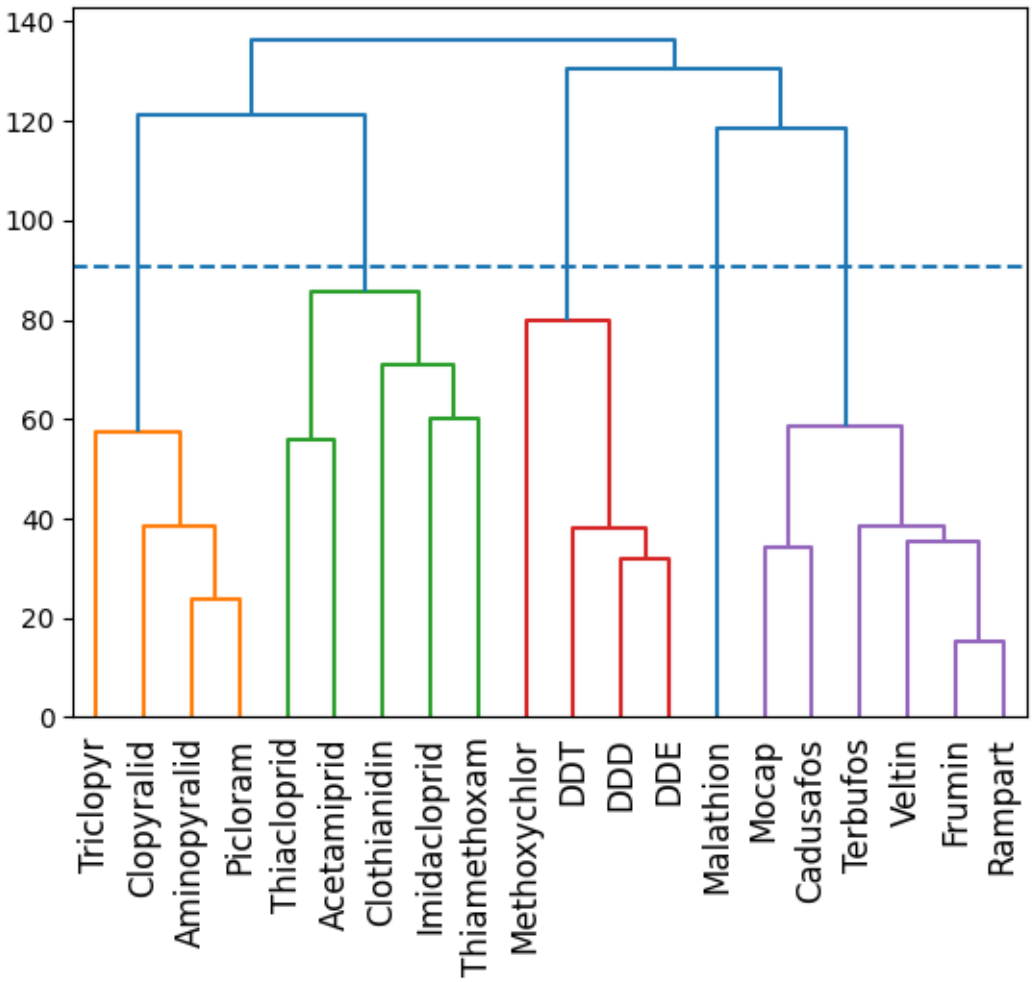
4 Wyniki analizy



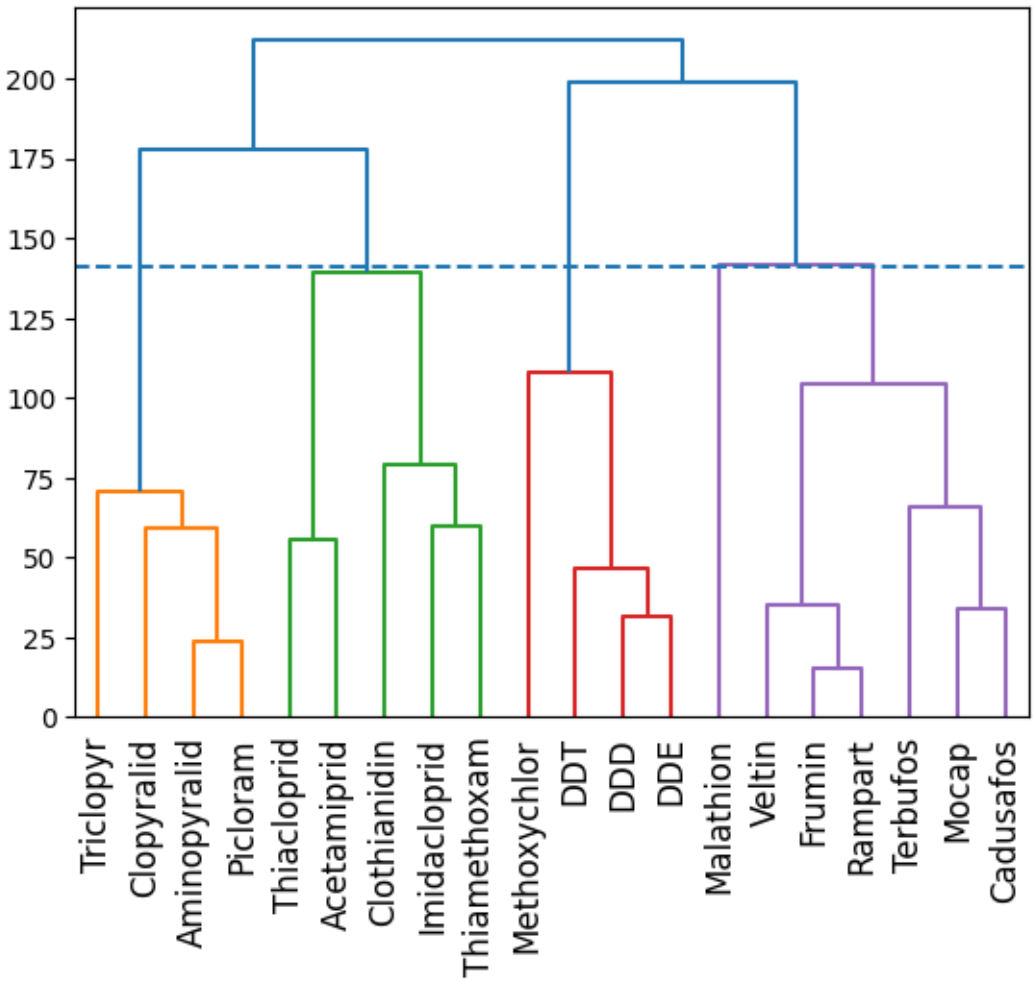
Rysunek 1: Dendrogram dla wiązania pojedynczego, metryka odległości Euklidesowej, dane autoskalowane



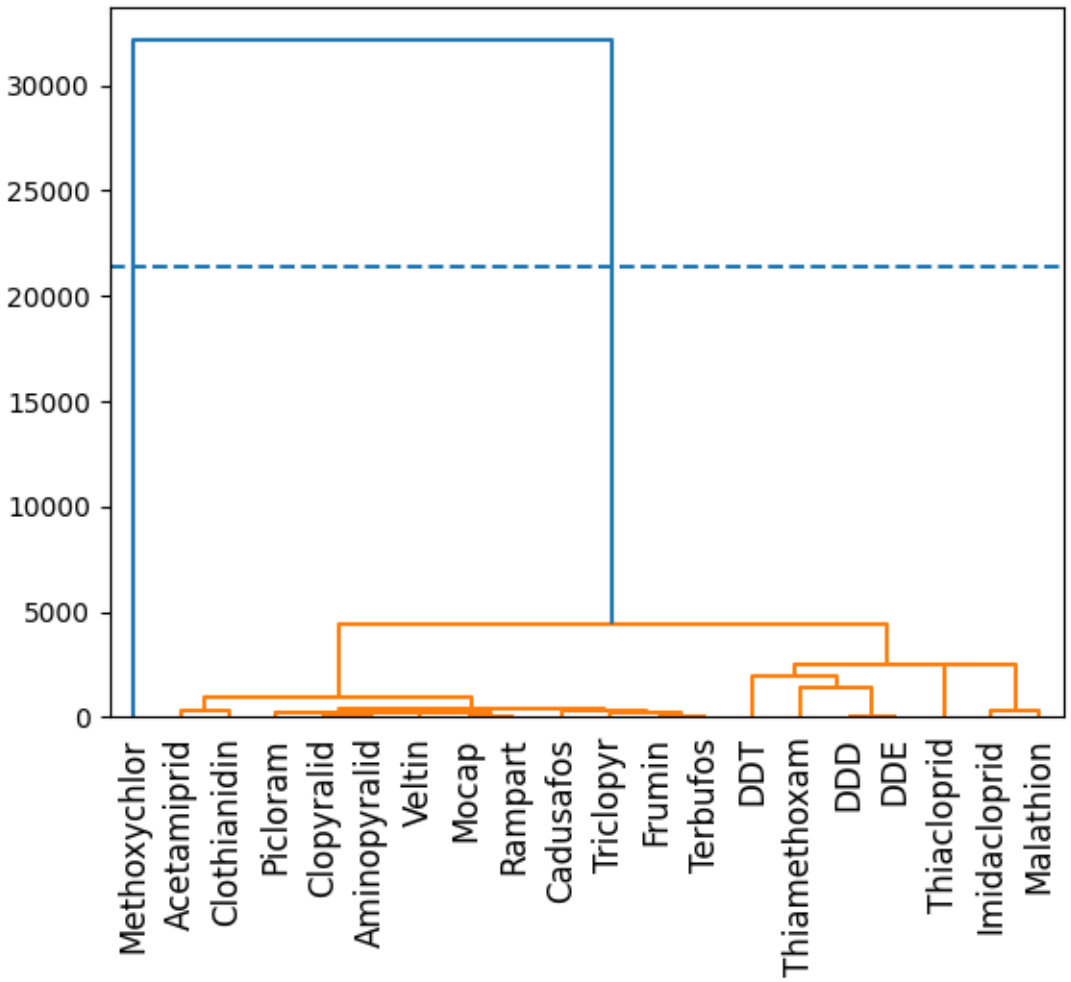
Rysunek 2: Dendrogram dla wiązania całkowitego, metryka odległości Euklidesowej, dane autoskalowane



Rysunek 3: Dendrogram dla wiązania pojedynczego, metryka odległości Manhattan, dane autoskalowane



Rysunek 4: Dendrogram dla wiązania całkowitego, metryka odległości Manhattan, dane autoskalowane



Rysunek 5: Dendrogram dla wiązania pojedynczego, metryka odległości Euklidesowej, dane nieautoskalowane

Tabela 1: Porównanie powstałej liczby klastrów oraz najwyższych wartości odległości dla dendrogramów.

	liczba klastrów	**najwyższa wartość odległości**
wiązanie pojedyncze odległość Euklidesowa	7	15.55
wiązanie całkowite odległość Euklidesowa	6	22.90
wiązanie pojedyncze odległość Manhattan	5	136.11
wiązanie całkowite odległość Manhattan	5	211.95
wiązanie całkowite odległość Euklidesowa	2	32107.06

Jak widać, powstająca liczba klastrów różni się od siebie w zależności od wybranego typu wiązania oraz metryki odległości. Dendrogramy z odległością Manhattan widocznie różnią się pod względem zarówno liczby klastrów oraz najwyższej wartości odległości.

Wartości otrzymane dla danych nieautoskalowanych różnią się znacząco od pozostałych. Powstały wykres jest bardzo skondensowany ze względu na odstającą najwyższą wartość odległości. Obserwowany dendrogram możnaby interpretować, że dane dzielą się tylko na 2 zbiory co odbiega od pozostałych wyników.

Warto zaznaczyć, że dendrogramy poniżej punktu odcięcia, który został ustawiony na $\frac{2}{3}$ wysokości, nieznacznie różnią się od siebie. Co najwyżej pojedynczymi przypadkami. Mowa oczywiście o przypadkach dla danych autoskalowanych.

5 Wnioski i dyskusja

Związki:

{ Methoxychlor, DDT, DDD, DDE } - DDT i analogi

{ Triclopyr, Clopyralid, Aminopyralid, Picloram } - kwasy pirydynokarboksylowe

{ Thiacloprid, Acetamiprid, Clothiaridin, Imidacloprid, Thiamethoxam } - neonikotynoidy

{ Malathion, Veltin, Frumin, Rampart, Terbufos, Mocap, Cadusafos } - związki organofosforanowe

w większości przypadków znajdowały się w jednym klastrze, lub ewentualnie, było blisko aby tak się stało na co wskazuje odległość maksymalnie jednego wiązania.

Taką sytuację możnaby interpretować, że dane pochodziły od najprawdopodobniej 4 typów związków. Taka informacja zgadza się z tą otrzymaną od osoby prowadzącej ćwiczenia. Po sprawdzeniu substancji potwierdziliśmy, że należą one do podanych wyżej grup pestycydów.

Jeśli jednak mielibyśmy wskazać, który dendrogram przedstawia informacje najbliższe tych prawdziwych, byłby to ten dla wiązania całkowitego i metryki odległości Manhattan dla danych autoskalowanych. Przedstawia on 5 klastrów, jednak wiązanie tworzące ów piąty klaster znajduję się tuż na granicy odcięcia.

Jak widać na powyższym przykładzie HCA, czyli hierarchiczna analiza klastrów, która jest przykładem uczenia maszynowego bez nadzoru dobrze nadaje się do podziału zbioru danych na mniejsze, łatwiejsze do interpretacji grupy, jednak potrzebne są do tego odpowiednie narzędzia oraz wiedza do interpretacji wyników.