Klastrowanie publicznych baz związków

Iwona Raczkowska
Jakub Kościukiewicz

Tematyka projektu

Pozwala na tworzenie ukierunkowanych kampani marketingowych, algorytmów rekomendujących filmy, kategoryzowanie klientów

KLASTROWANIE

Polega na wyznaczaniu grup danych o podobnych cechach.

Technika uczenia nienadzorowanego

Algorytmy klastrowania pomagają w systemach rekomendacyjnych, jak te stosowane przez Netflix czy Spotify, do grupowania użytkowników o podobnych gustach.

Iris
setosa

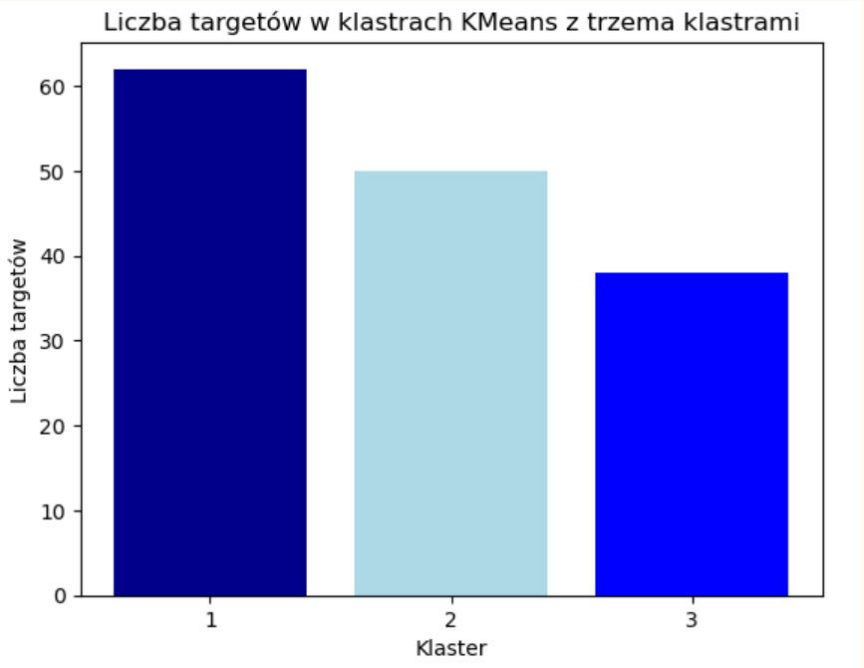
Iris versicolor

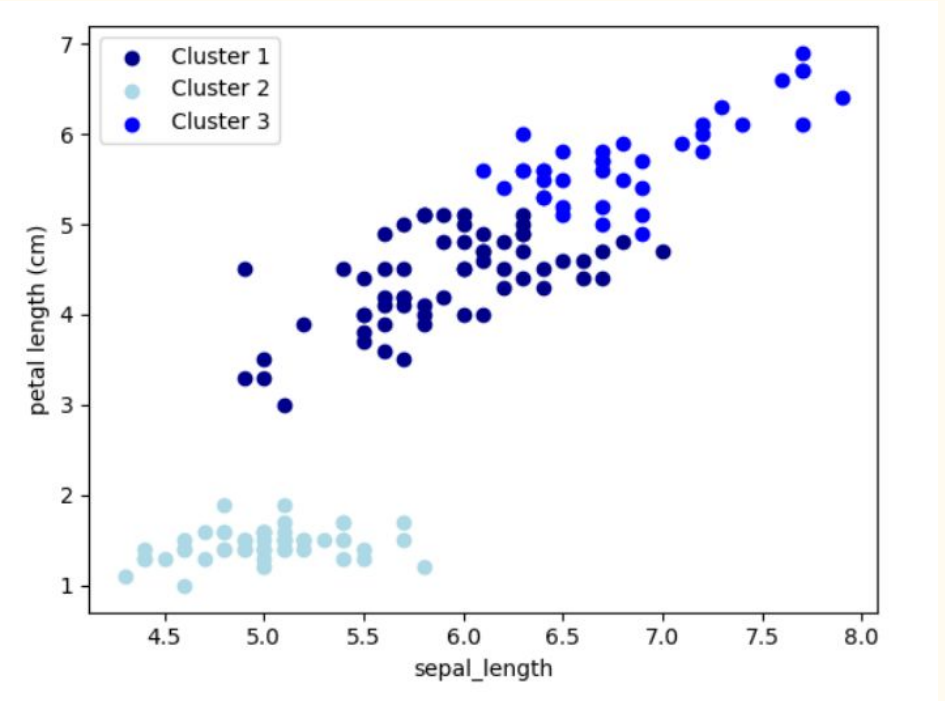
Iris virginica

```
irysy["target"]
```

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2])
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2


Liczba targetów w klastrach KMeans z trzema klastrami
Liczba targetów
Klaster
0
10
20
30
40
50
60
1
2
3


Cluster 1
Cluster 2
Cluster 3
petal length (cm)
sepal_length
1
2
3
4
5
6
7
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

Co jest celem projektu?

Znalezienie ciekawych zależności w dużych publicznych zbiorach baz związków co pozwoli na weryfikację hipotez:

1. W obrębie poszczególnych grup związków chemicznych można zaobserwować wspólne cechy korelujące ze źródłem ich danych.

Wybór reprezentacji związków chemicznych: alkaloidy + związki, które nie są alkaloidami

Alkaloidy - grupa naturalnie występujących zasadowych związków organicznych (na ogół heterocyklicznych), głównie pochodzenia roślinnego, zawierających azot.

W obrębie alkaloidów do wspólnych cech można zaliczyć obecność w strukturze pierścienia z azotem, możliwość wykorzystania jako narkotyczny lek przeciwbólowy, naturalne pochodzenie i na podstawie tych informacji przeprowadzona zostanie klasteryzacja.

2. Związki chemiczne nie są równomiernie rozłożone w przestrzeni chemicznej, co można zbadać poprzez analizę ich rozmieszczenia w wybranej przestrzeni.

Przestrzeń chemiczna to zbiór wszystkich możliwych związków chemicznych, w tym wszystkich znanych cząsteczek leków oraz tych, które dopiero mają zostać odkryte. Szacuje się, że całkowita liczba związków tworzących przestrzeń chemiczną wynosi 10^{60}. Do tej pory zbadano tylko niewielki ułamek tej przestrzeni. W bazie ChEMBL znajduje się 2,4 mln związków.

Klastry jako obszary wysokiej gęstości

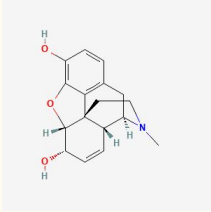
Związki o podobnych właściwościach chemicznych i biologicznych grupują się w przestrzeni chemicznej, tworząc klastry. Leki działające na podobne cele biologiczne mogą wykazywać podobieństwa w ich profilach chemicznych, co skutkuje ich zbliżonym położeniem w przestrzeni. Zatem obecność klastrów będzie w stanie potwierdzić hipotezę.

Wybór reprezentacji związków chemicznych: alkaloidy + związki, które nie są alkaloidami

3. Klastrowanie związków chemicznych na podstawie wartości ich aktywności odkryje niespójności pomiędzy różnymi testami biologicznymi.

- Różne grupy badawcze mogą przeprowadzać badania na tym samym związku, ale w różnych warunkach laboratoryjnych, co może prowadzić do powstania wielu zestawów danych dotyczących tego samego związku. Różnice między sprzętem stosowanym w różnych ośrodkach naukowych mogą wpływać na niespójność danych.
- Różne parametry aktywności: Związek może być testowany pod kątem różnych typów aktywności, takich jak inhibicja, powinowactwo, toksyczność, itp. Miary aktywności mogą być określane jako Ki, IC50 itp.
- Różne cele biologiczne: Aktywność związku mogła być testowana przeciwko różnym wariantom konkretnego receptora lub przeciwko innym receptorom w różnych organizmach lub systemach komórkowych.

Wybór reprezentacji związków chemicznych: morfina



4. Leki spełniające cztery kryteria Lipińskiego mają szansę stać się skutecznym lekiem doustnym.

- Masa molowa < 500 Da
- Współczynnik podziału oktanol: woda (log P) < 5
- Liczba donorów wiązania wodorowego < 5
- Liczba akceptorów wiązania wodorowego < 10

Chemiczne/biologiczne znaczenie projektu

Szukanie zależności między związkami w celu klasyfikacji lub znalezienia anomalii w publicznych bazach danych.

Potencjalna trudność w wykonaniu projektu

- Problem z ukazaniem co tak naprawdę dzieje się w danym zbiorze. W publicznych bazach danych związków chemicznych dane są zazwyczaj wielowymiarowe. Wielowymiarowość danych wynika z różnorodnych właściwości takich jak: struktura chemiczna, fizykochemiczne właściwości, aktywność biologiczna, bezpieczeństwo i toksykologia.
- Dane o wielu wymiarach - problem z wizualizacją danych.
- Problem z identyfikowanie wzorców w danych. Aby zapewnić powtarzalność wyników, potrzeba ustawienia parametru random state w konstruktorze K-Means. K-Means rozpoczyna proces klasteryzacji przez losowe wybranie początkowych punktów jako centroidów klastrów. Jeśli punkty startowe są wybierane losowo, każde uruchomienie algorytmu może prowadzić do różnych wyników.

Spodziewany efekt projektu

Dla hipotezy 1.: Utworzenie klastra, który będzie się składać z alkaloidów, poza klastrem związki, które nie są alkaloidami.

Dla hipotezy 2.: Powstanie obszarów o wysokiej gęstości - klastrów.

Dla hipotezy 3.: Uzyskanie co najmniej jednego klastra z punktami (wartościami aktywnościowymi), które będą odstawać.

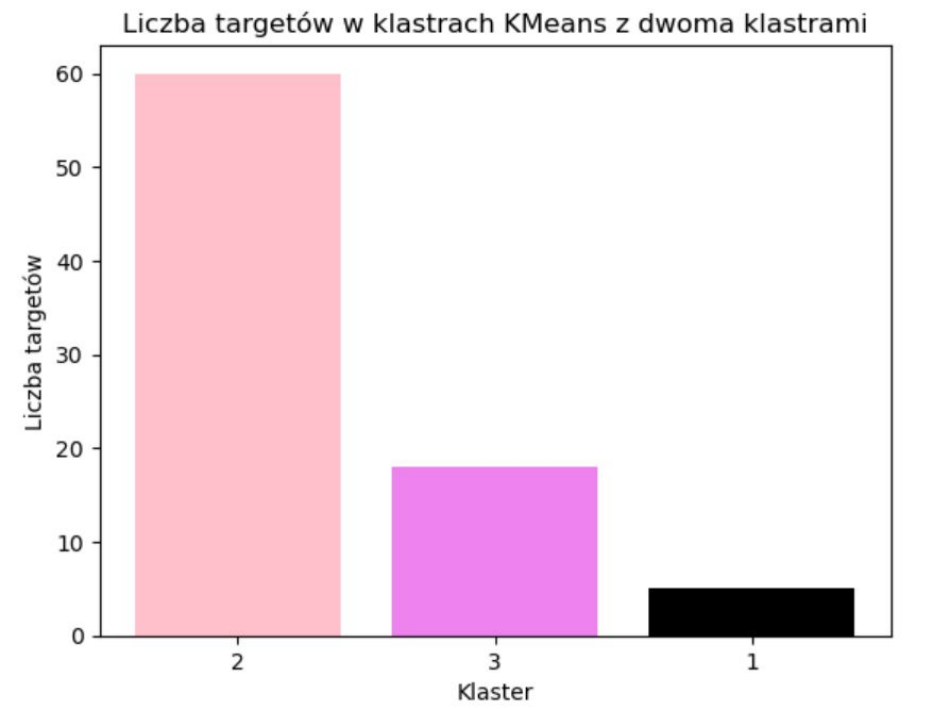
Dla hipotezy 4.: Uzyskanie jednego klastra składającego się na związki, które mają szansę stać się skutecznym lekiem doustnym.

Wybór reprezentacji związków chemicznych: Związki, które są popularnymi i skutecznymi lekami doustnymi oraz te, które nie są.

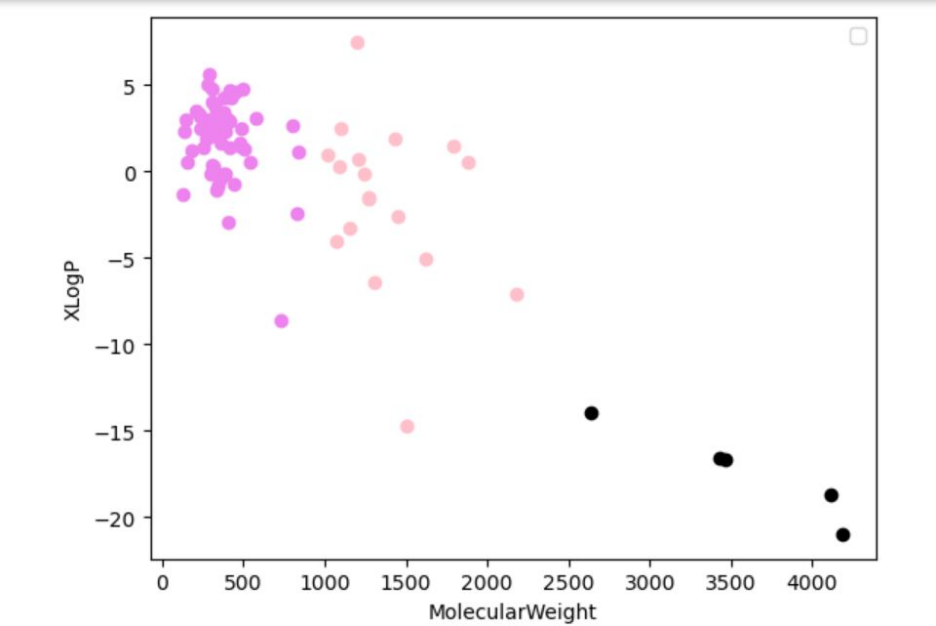
- Statyny - obniżanie poziomu cholesterolu we krwi (simwastatyna);
- Leki przeciwbólowe, przeciwgorączkowe (paracetamol, aspiryna, ibuprofen);
- Leki stosowane w leczeniu depresji, zaburzeń lękowych (fluoksetyna, diazepam).

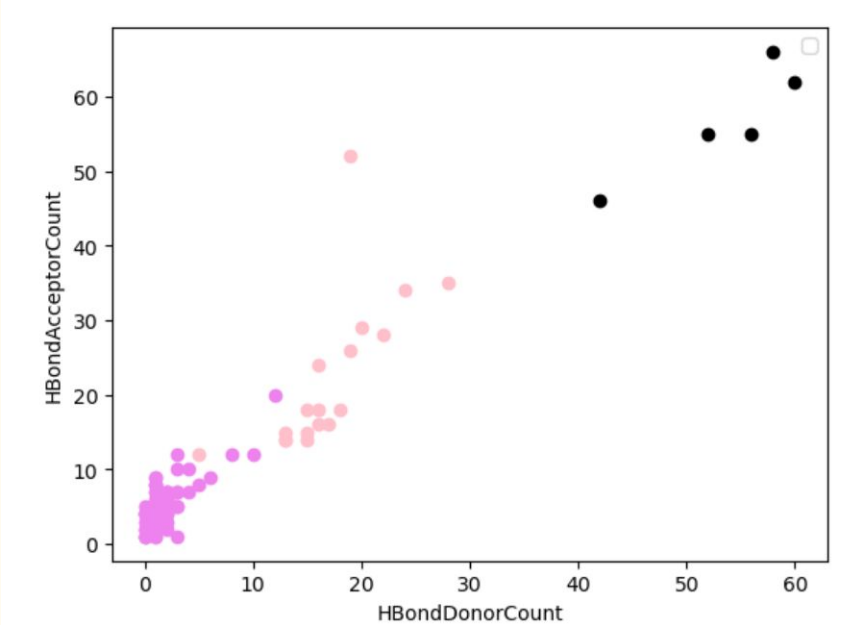
- Antybiotyki stosowany do leczenia ciężkich zakażeń opornych na inne antybiotyki (kolistyna - nie jest wchłaniana po podaniu doustnym i musi być podawana dożylnie lub przez inhalację);
- Peptydy stosowany w leczeniu silnego przewlekłego bólu (Zikonotyd podawany dooponowo, bezpośrednio do płynu rdzeniowego, z powodu jego niezdolności do przekroczenia bariery krew-mózg po podaniu doustnym.).

Wyniki eksploracyjnej analizy danych: wizualizacja danych i wnioski

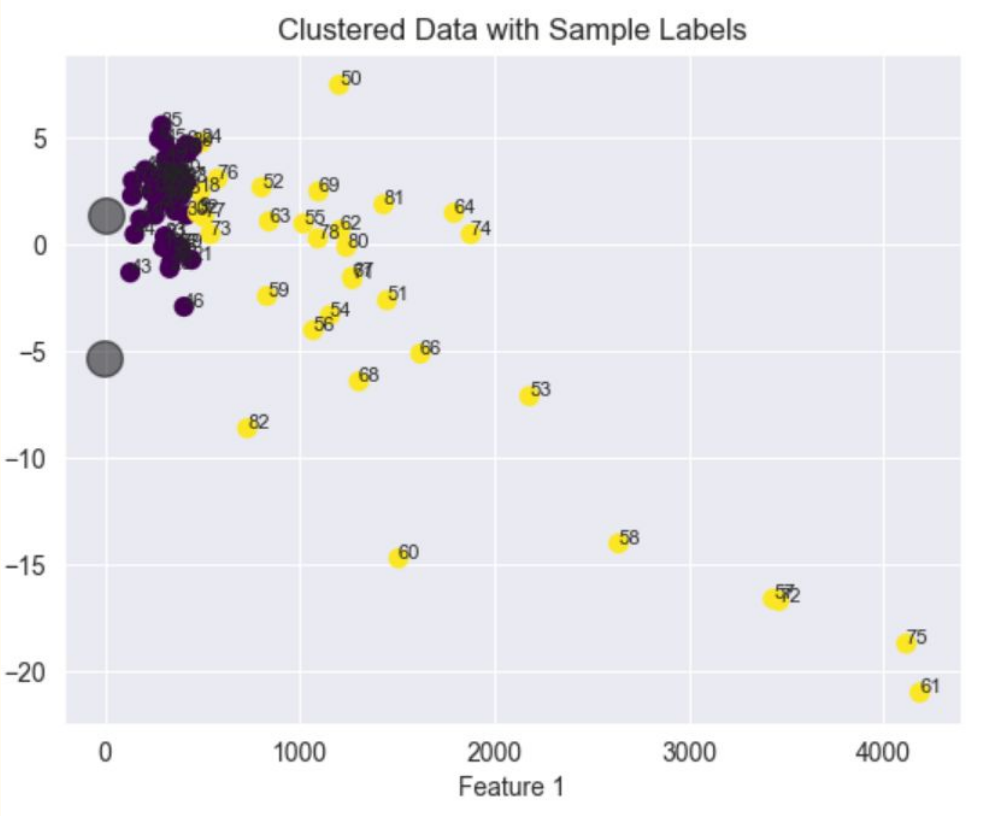


Wyniki eksploracyjnej analizy danych: wizualizacja danych i wnioski

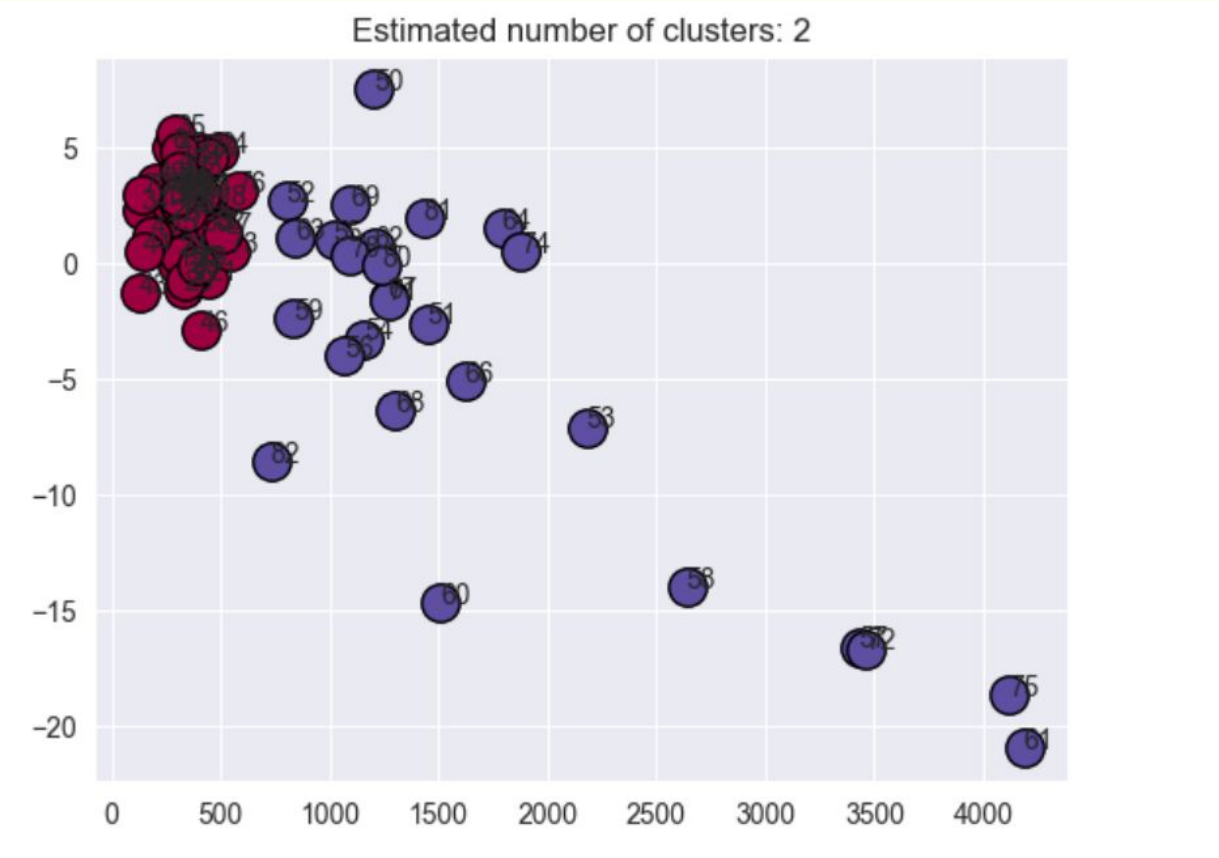




KMeans


Clustered Data with Sample Labels
5
0
−5
−10
−15
−20
0
1000
2000
3000
4000
Feature 1
50
52
69
81
64
74
63
55
62
78
80
73
59
51
54
56
66
68
53
82
58
60
75
61

DBSCAN


Estimated number of clusters: 2
5
0
−5
−10
−15
−20
0
500
1000
1500
2000
2500
3000
3500
4000

Dane wejściowe użyte w projekcie

- baza PubChem i ChEMBL

Jak zdefiniowane będzie wejście modelu?

Wejście:

Typ danych: dane liczbowe, reprezentacje binarne (fingerprinty Morgana).

API (Application Programming Interface) interfejs który, dostarcza zestaw specyfikacji i narzędzi, które umożliwiają zdalne interakcje z bazą danych. Pozwala to na automatyczne wyszukiwanie, pobieranie oraz analizowanie informacji chemicznych bez konieczności ręcznego przeszukiwania serwisu internetowego.

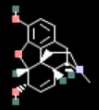
Format danych: Dane zwracane przez API są w formacie JSON lub XML.

3.1 Computed Properties

Property Name	Property Value	Reference
Molecular Weight	151.16 g/mol	Computed by PubChem 2.2 (PubChem release 2021.10.14)
XLogP3	0.5	Computed by XLogP3 3.0 (PubChem release 2021.10.14)
Hydrogen Bond Donor Count	2	Computed by Cactvs 3.4.8.18 (PubChem release 2021.10.14)
Hydrogen Bond Acceptor Count	2	Computed by Cactvs 3.4.8.18 (PubChem release 2021.10.14)

2.1.4 Canonical SMILES

CN1CCC23C4C1CC5=C2C(=C(C=C5)O)OC3C(C=C4)O



50% Inhibition of stereospecific [3H]-naltrexone (10e-9 M) binding towards opiate receptor in rat brain homogenate

Activity Outcome: Active Activity Type: IC50 Activity Value: 0.000027 µM

BioAssay AID: 145933

Target Name: Oprm1 - opioid receptor, mu 1 (Norway rat)

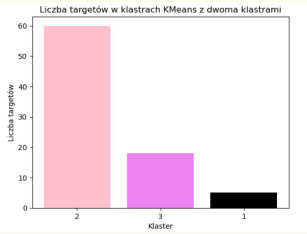
Substance SID: 103169185 Compound CID: 5288826

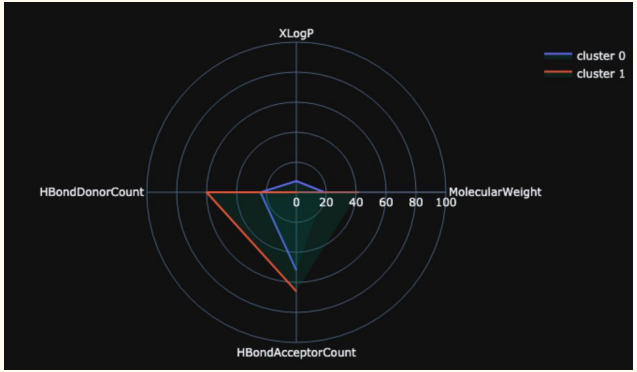
Jak zdefiniowane będzie wyjście modelu?

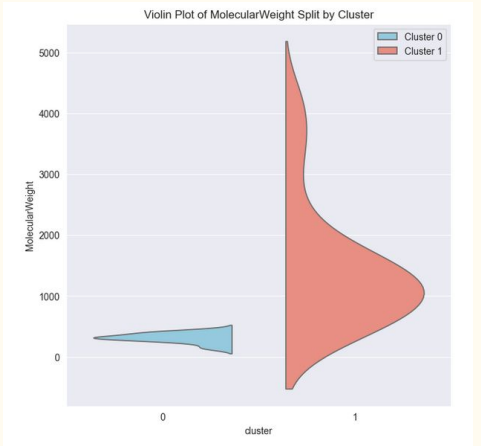
Wyjście:

- Wizualizacja za pomocą histogramów, wykresów rozproszenia, wykresów radarowych, wykresów skrzypcowych;

- Etykiety związków należących do klastrów: Każdy element zestawu danych, reprezentowany przez etykietę, jest przyporządkowany do konkretnego klastra. Co będzie przydatne przy weryfikacji klastrowania.

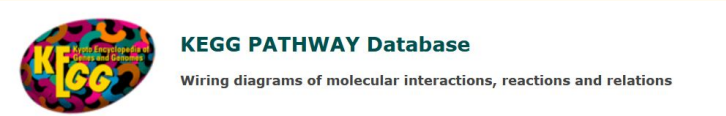






Skąd można pozyskać więcej danych do projektu

Inne bazy danych:



Jakie problemy w dostępnych danych są widoczne?

Reguła Lipińskiego jest przede wszystkim użyteczna w projektowaniu małych cząsteczek, które są przeznaczone do wchłaniania przez przewód pokarmowy i które działają wewnątrzkomórkowo.

- ChEMBL i PubChem to bazy danych, które skupiają się głównie na małych cząsteczkach organicznych, które mają zastosowanie w chemii medycznej i farmakologii. Informacje na temat cząstek niespełniających reguły Lipińskiego są mniej dostępne w tych bazach danych;

- Różnorodne zapisy notacji, np. potrzeba zamiany przecinków na średniki przy imporcie na csv;

- Duplikowanie nagłówka, który trzeba było wyeliminować;

- Brak niektórych danych w poszczególnych bazach - scalenie danych z dwóch baz.

Jakie dodatkowe informacje (metadane) są dostępne?

- Dostępność narzędzi do pobierania danych takich jak API;
- Informacja o pochodzeniu danych;
- Informacja o receptorze, przeciwko któremu były prowadzone badania, może być traktowana jako metadane w kontekście wartości aktywności chemicznej. Takie metadane są szczególnie ważne w badaniach farmakologicznych, w kontekście zrozumienia eksperymentu i interpretacji danych.

Jakie metody uczenia maszynowego będą wykorzystane w projekcie:

K-means

Jest to metoda klastrowania, która grupuje dane na podstawie ich cech do k różnych skupień. K-means jest algorytmem uczenia nienadzorowanego, co oznacza, że próbuje znaleźć strukturę w nieoznakowanych danych poprzez minimalizację wariancji wewnątrz klastra.

t-SNE

Jest to technika redukcji wymiarowości, szczególnie przydatna do wizualizacji wysokowymiarowych zbiorów danych. Działa poprzez próbę zachowania podobieństwa między punktami w przestrzeni wielowymiarowej, kiedy są one rzutowane na przestrzeń o niższej liczbie wymiarów, zazwyczaj 2D lub 3D.

DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to algorytm klastrowania w ramach uczenia nienadzorowanego wykorzystywany w analizie danych, który jest szczególnie efektywny w identyfikacji klastrów o nieregularnych kształtach i rozmiarach, jak również w obsługiwaniu punktów odstających (outliers). DBSCAN skupia się na grupowaniu punktów, które są blisko siebie, bazując na gęstości, co odróżnia go od innych metod klastrowania, takich jak K-means, które wymagają wstępnej specyfikacji liczby klastrów.

- Nie wymaga określenia liczby klastrów: W przeciwieństwie do K-means, DBSCAN automatycznie określa liczbę klastrów na podstawie danych.
- Dobrze radzi sobie z punktami odstającymi: Punkty, które nie pasują do żadnego klastra, są łatwo identyfikowane jako odstające.
- Zdolność do identyfikacji klastrów o nieregularnych kształtach: Może znajdować klastry o dowolnym kształcie, co jest trudne dla wielu innych algorytmów klastrowania.
- Trudności z różnymi gęstościami: Jeśli klastry mają znacznie różne gęstości, DBSCAN może mieć trudności z ich poprawnym zidentyfikowaniem.

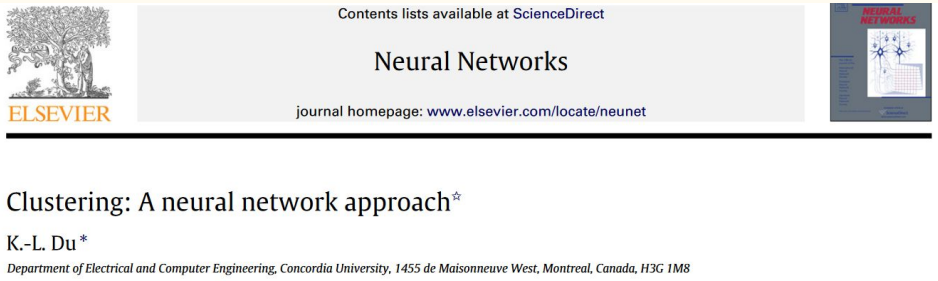
Czy ten temat był już poruszany w literaturze? Jeśli tak, to jakie narzędzia są dostępne?

Comparison of hierarchical clustering and neural network clustering: an analysis on precision dominance

Nazish Shahid

Scientific Reports **13**, Article number: 5661 (2023) | Cite this article

4012 Accesses | **5** Citations | Metrics



Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

ELSEVIER

Clustering: A neural network approach☆

K.-L. Du*

Department of Electrical and Computer Engineering, Concordia University, 1455 de Maisonneuve West, Montreal, Canada, H3G 1M8

Clustering of chemical structures on the basis of two-dimensional similarity measures

J. M. Barnard and G. M. Downs

Clustering of chemical data sets for drug discovery

Dostępne narzędzia do klastrowania danych: Scikit-learn - Biblioteka Pythona zapewniająca proste i wydajne narzędzia do analizy danych i modelowania predykcyjnego, w tym różne algorytmy klastrowania, takie jak K-means, DBSCAN.

Jakie miary będą zastosowane do zmierzenia skuteczności modelu

Klasteryzacja jest przykładem uczenia nienadzorowanego - trudniej niż w uczeniu nadzorowanym mierzyć i sprawdzać czy algorytmy działają poprawnie, ponieważ nie mamy punktu odniesienia.

Do zmierzenia skuteczności wyników klastrowania: Porównanie informacji między różnymi bazami związków, np. w przypadku hipotezy 4. skorzystano z informacji z bazy PubChem, a weryfikację wyniku przeprowadzono na podstawie bazy ChEMBL, gdzie można znaleźć informację czy dany związek spełnia regułę Lipińskiego.

Do zmierzenia skuteczności określenia ilości klastrów:

- metoda łokcia
- model inertia

Metoda łokcia

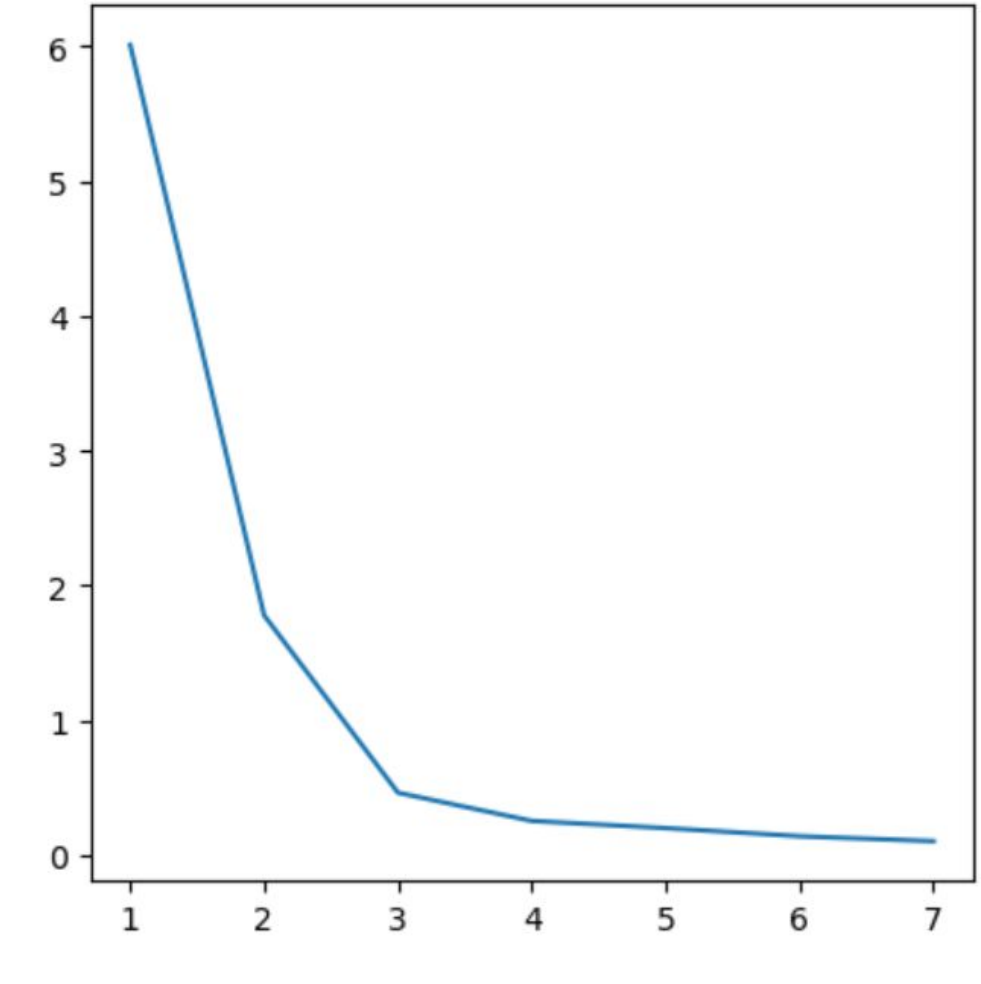
Metoda łokcia jest popularnym podejściem używanym do określania odpowiedniej liczby klastrów w algorytmie klastrowania, takim jak K-means. Jest to wizualna metoda, która pomaga wybrać optymalną liczbę klastrów poprzez analizę zmiany wartości funkcji kosztu (na przykład inercji w K-means) w zależności od liczby klastrów.

- Inercja odnosi się do sumy kwadratów odległości każdego punktu od najbliższego centrum klastra. Jest to miara, która pomaga ocenić, jak dobrze dane są sklastrowane. Inercja maleje w miarę zwiększania liczby klastrów, ponieważ punkty są bliżej swoich centroidów.

- Używając różnych wartości dla liczby klastrów, oblicza się inercję dla każdego przypadku i tworzy się wykres zależności inercji od liczby klastrów.

- Wykres zazwyczaj pokaże szybki spadek inercji przy niskiej liczbie klastrów, po czym zmiana staje się mniej wyraźna i wykres zaczyna się wypłaszczać. Punkt, w którym spadek inercji łamie się i zaczyna być mniej stromy, jest nazywany "łokciem". Jest to moment, w którym dodanie kolejnych klastrów przynosi coraz mniejszą poprawę w agregacji danych

- "Łokieć" na wykresie sugeruje optymalną liczbę klastrów. Jest to punkt, po którym dodatkowe klastry nie wprowadzają znaczącej poprawy do wartości inercji, a tylko zwiększają złożoność modelu.



Jaki stos technologiczny będzie użyty do wykonania projektu

- Język programowania: Python;
- Biblioteki: Pandas, NumPy, Scikit-learn, RDKit i Matplotlib;
- Narzędzia do monitorowania i zarządzania projektami: Git;
- Zarządzanie środowiskami: Conda.

Źródła:

- Rishal Hurbans, *Algorytmy sztucznej inteligencji*
- Bogocz Jacek. (2016). *Metody eksploracji baz danych w poszukiwaniu nowych reguł projektowania leków*. Praca doktorska. Katowice: Uniwersytet Śląski
- https://en.m.wikipedia.org/wiki/File:Spotify_logo_with_text.svg
- https://pl.m.wikipedia.org/wiki/Plik:Netflix_2015_logo.svg
- https://pl.wikipedia.org/wiki/Alkaloidy
- https://pl.wikipedia.org/wiki/Mak_lekarski
- https://extrapolations.com/what-is-chemical-space/
- https://www.herbapol-polana.com/vademecum-botaniczne/mak-lekarski
- https://www.theguardian.com/film/2019/jan/21/from-red-pills-to-red-white-and-blue-brexit-how-the-matrix-shaped-our-reality
- https://pl.pinterest.com/pin/563794447074793442/
- https://pl.pinterest.com/pin/442197257153939610/
- https://pubchem.ncbi.nlm.nih.gov/compound/Morphine
- https://www.artstation.com/artwork/6bNG65
- https://www.ebi.ac.uk/chembl/
- https://pubchem.ncbi.nlm.nih.gov/
- https://www.slideshare.net/100006619533516/zinc-database
- https://go.drugbank.com/
- https://www.genome.jp/kegg/pathway.html
- https://www.rcsb.org/
- https://ieeexplore.ieee.org/abstract/document/7036702
- https://fulmanski.pl/zajecia/seminarium_katedry/tenn/TymoszczukPrace/old/1.pdf
- https://www.nature.com/articles/s41598-023-32790-3
- https://pubs.acs.org/doi/pdf/10.1021/ci00010a010