10. VAJA: STROJNO UČENJE

Dominik Primožič

3. JUNIJ 2025

1. Naloga

Združite organske spojine v skupine na podlagi njihove kemijske zgradbe ter fizikalno-kemijskih lastnosti. Uporabite enak nabor kot v prvi nalogi (vsaj 1000 organskih spojin), za katere so na voljo naslednji podatki:

- Strukturni deskriptorji, izračunani s knjižnico RDKit: molekulska masa, LogP (lipofilnost), polarna površina (TPSA), število donorjev in akceptorjev vodikovih vezi, število rotacijskih vezi ipd.
- Eksperimentalne lastnosti: vrelišče, tališče in topnost v vodi (npr. v obliki logS).

Naredite analizo glavnih komponent (PCA) za zmanjšanje dimenzionalnosti podatkov ter nato gručenje z algoritmom k-means ali DBSCAN. Rezultate gručenja interpretirajte: ali se spojine v posameznih skupinah ujemajo glede na velikost, polariteto, topnost, temperaturo tališča ali vrelišča? Rezultate vizualizirajte in komentirajte smiselnost dobljenih skupin.

Metode

Zaradi omejene razpoložljivosti eksperimentalnih podatkov (vrelišče, tališče, topnost) sem gručenje izvedel ločeno za vsak nabor podatkov. Zaradi različnih numeričnih meril (npr. molekulska masa, LogP, TPSA) sem predhodno izvedel standardizacijo vseh spremenljivk. Nato sem s principalno komponentno analizo (PCA) zmanjšal dimenzionalnost:

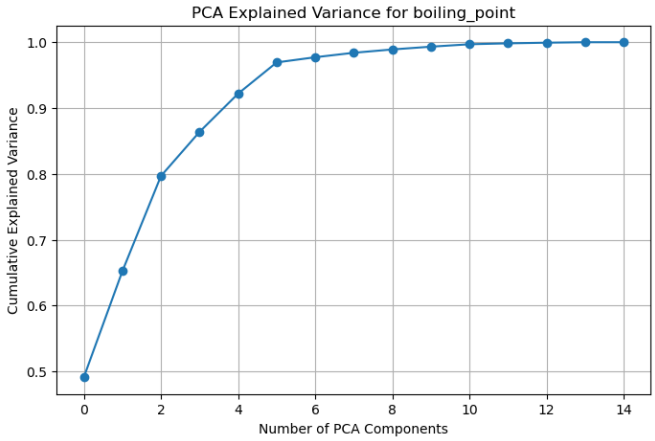
$$Z = XW$$

kjer je X matrika standardiziranih podatkov, W matrika lastnih vektorjev, Z pa projekcija v prostor glavnih komponent.

Za nadaljnjo analizo sem izbral prvih nekaj komponent, ki pojasnjujejo največ variance, in na tej osnovi izvedel gručenje z algoritmoma k-means in DBSCAN.

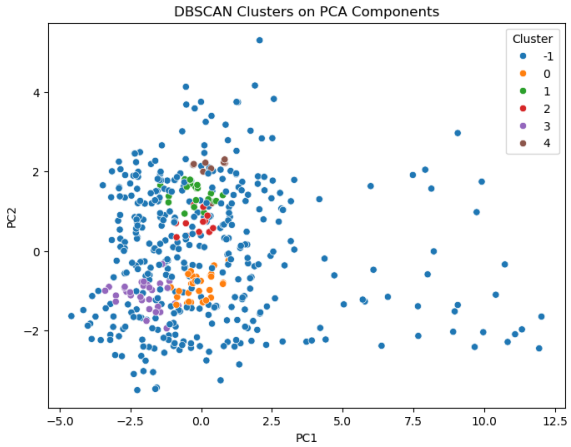
Rezultati

Najprej sem se lotil grupiranja za set podatkov z vreliščem. Izvedel sem PCA in preveril koliko PCA osi zares potrebujem za nadaljnje delo.



Graf 1: PCA varianca

Že samo 4 osi so dovolj dobre za zajem podatkov, a sem vseeno uporabil 6 PCA osi. Poskusil sem različne variante DBSCAN parametrov, a nisem dobil večjih grupiranj kot eps=1 in po 10 vzorcev v klastre. Dobljene skupine so precej majhne.



Graf 2: DBSCAN (-1 so negrupirani)

Skupina	Število v skupini
0	29
1	24
2	14
3	31
4	11

Tabela 1: Velikosti skupin DBSCAN-a

Skupine so tako majhne, da je težko govoriti o kakšni podobnosti. Vseeno pa sem naredil grafično analizo porazdelitve določenih lastnosti v skupinah.

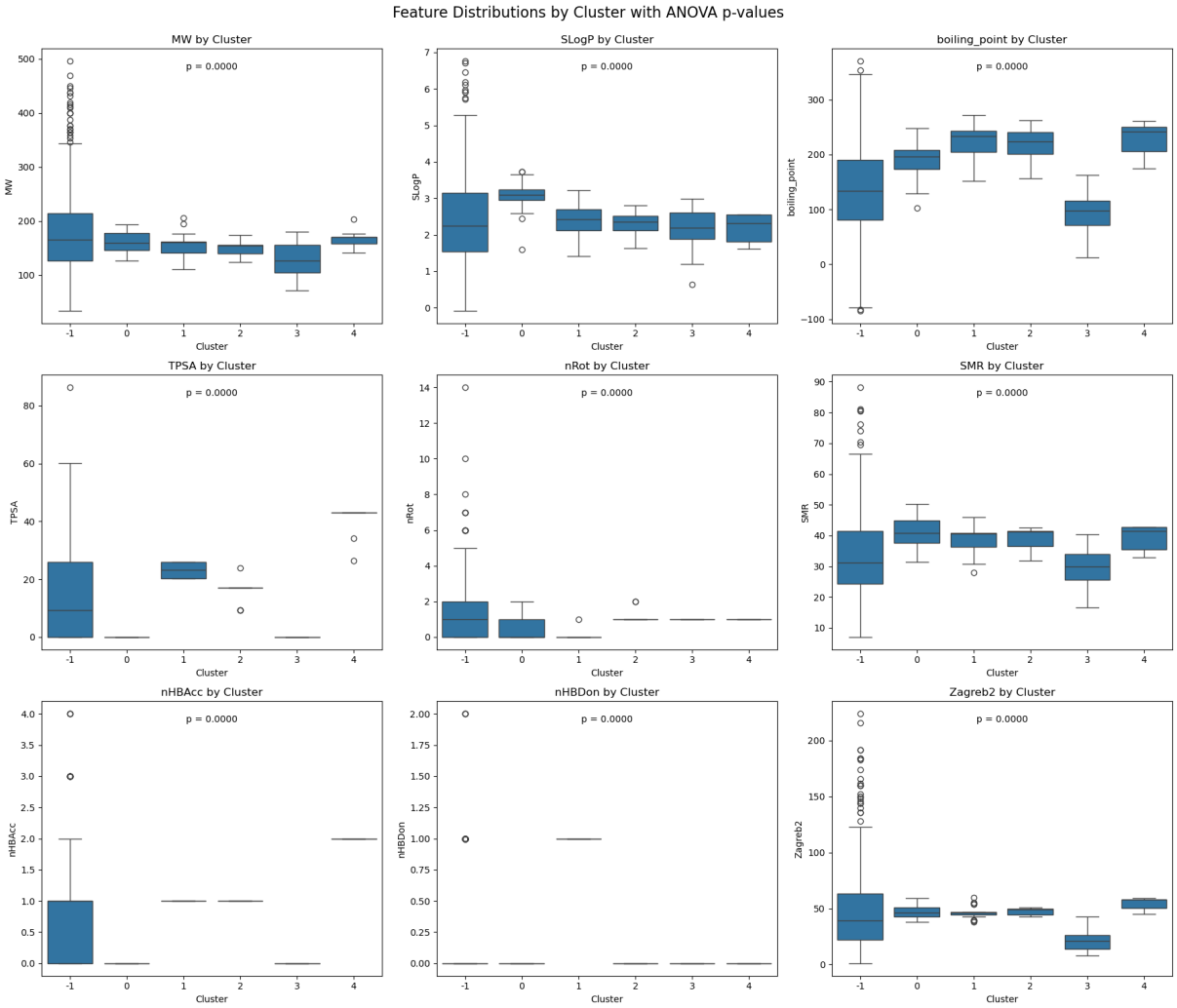
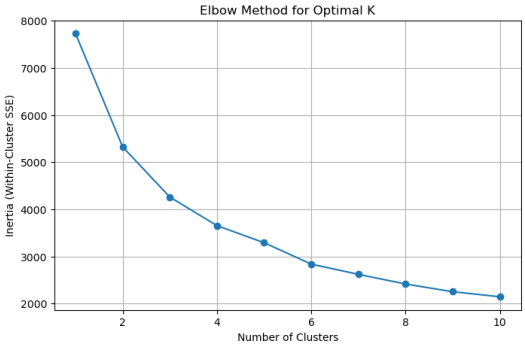


Tabela 2: Lastnosti v posameznih skupinah

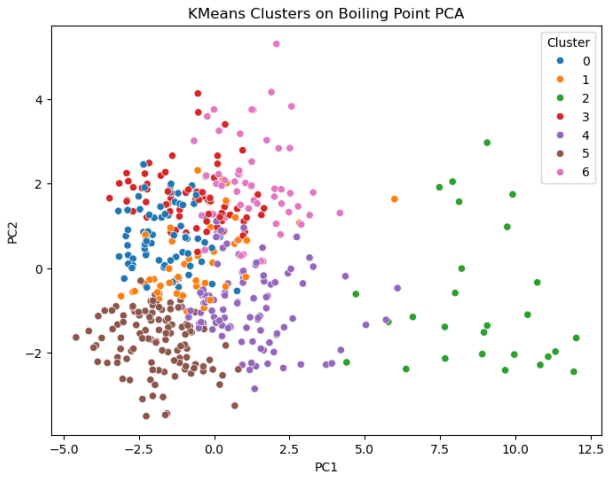
Za vse so p vrednosti iz ANOVA testa enake 0, kar pomeni, da ni nobene statistične povezanosti med vrednostmi v grupi. Glede na same grafe bi edina TPSA lahko bil parameter, ki najbolj vpliva na skupine. Druge lastnosti so med skupinami preveč podobne, da bi lahko govorili o grupiranju glede na njih.

Poskusil sem tudi k-means grupiranje v upanju, da bodo nastale bolj smiselne skupine. Naredil sem elbow plot, da sem določil optimalno število skupin, to je bilo 7.

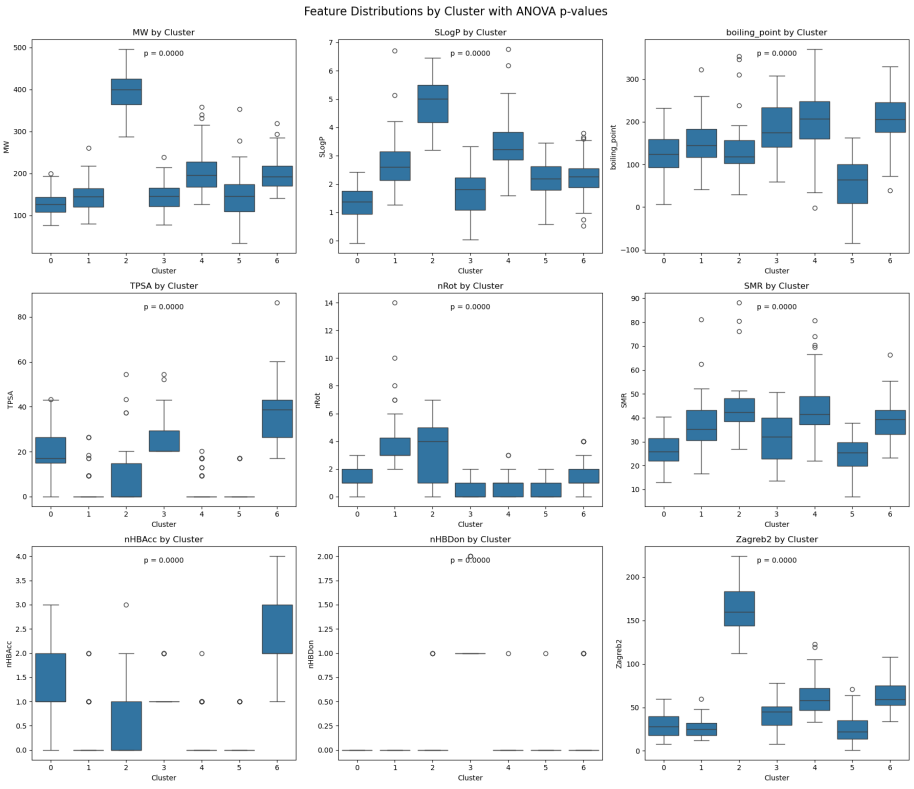


Graf 3: Elbow plot za določitev optimalnega števila skupin

Dobljene skupine so potem bile:



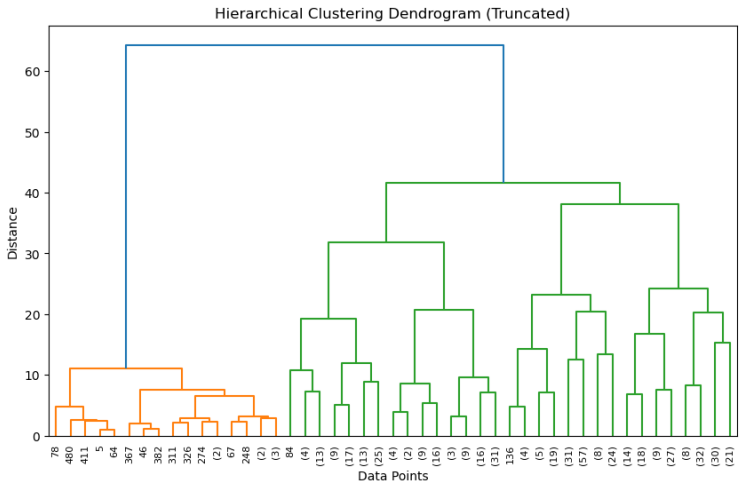
Graf 4: KMeans skupine



Graf 5: Lastnosti po KMenas skupinah

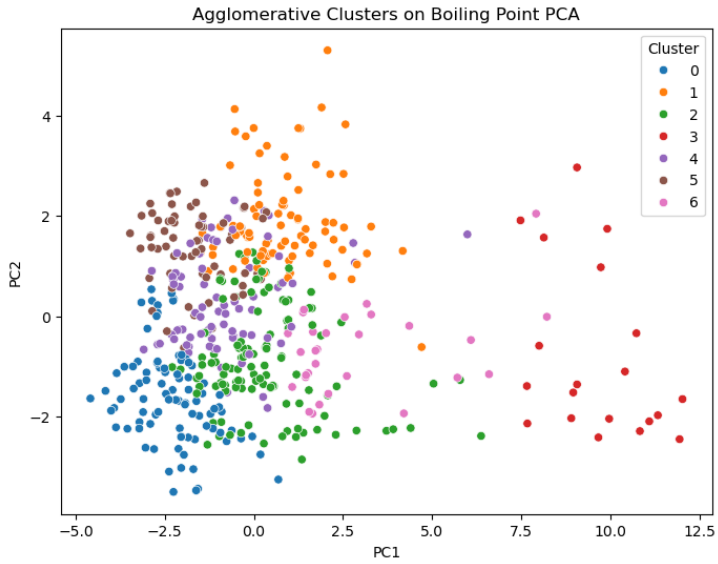
Tudi v tem primeru ni nobenih izrazitih povezovalnih lastnosti. Glede na vrelišče se posebno ujame skupina 3, ki odstopa malo tudi v drugih lastnostih. Tu bi lahko bil kemijski razlog za grupiranje.

Odločil sem se poskusiti še aglomerativno grupiranje. Najprej sem naredil dendrogram, da sem določil optimalno število skupin.



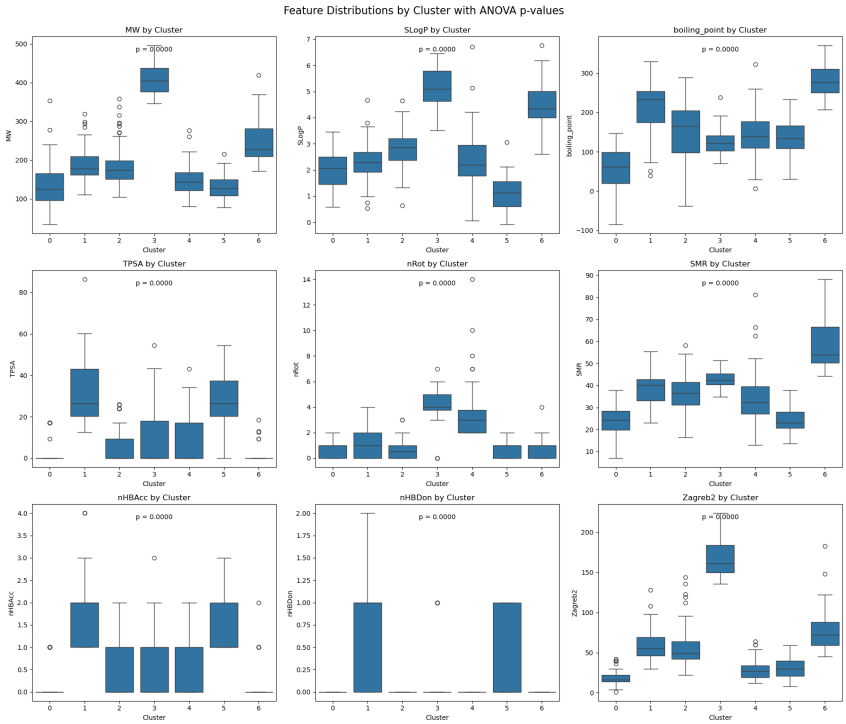
Graf 6: Dendrogram

Odločil sem se za 7 skupin.



Graf 7: Aglomerativne skupine

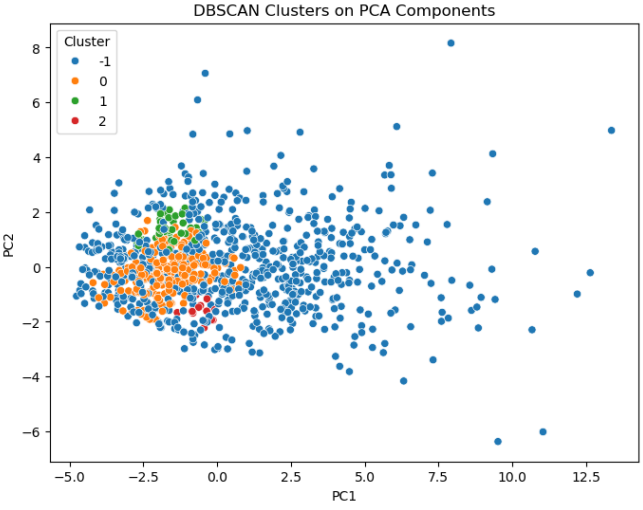
Na prvi pogled tudi tu ni nobenih posebno urejenih skupin, razen skupine 3, ki so vsi outlier-ji.



Graf 8: Lastnosti aglomerativnih skupin

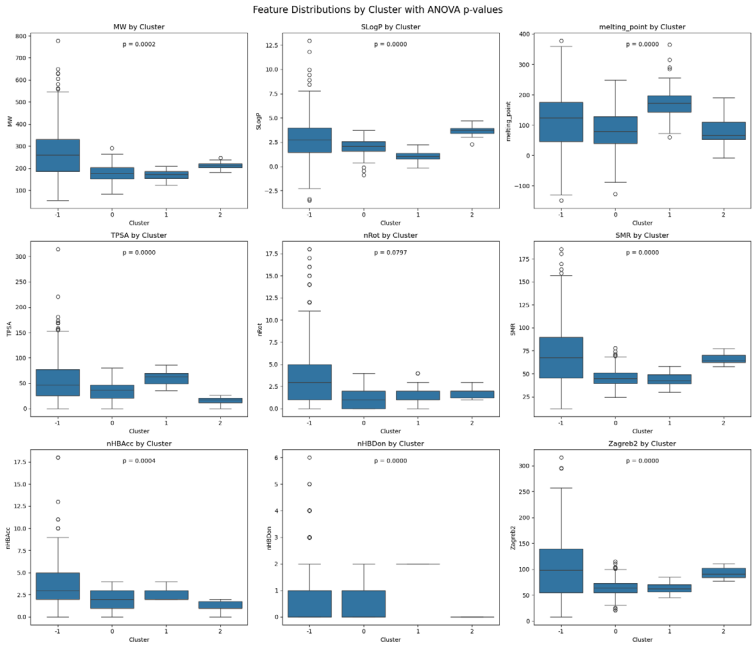
V teh skupinah so določene lastnosti, a so p vrednosti zopet 0, zato ne moremo govoriti o statistično signifikantnih grupah.

Podobno sem naredil še za tališča in topnost. Spodaj predstavljam le rezultate DBSCANa, saj podobno nisem dobil nobenih dobrih grupiranj.



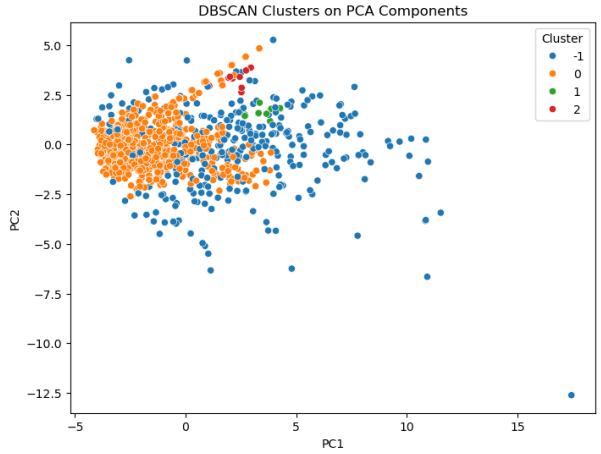
Graf 9: DBSCAN za tališče

Tu sem dobil še manj skupin, ki pa so vse enake.



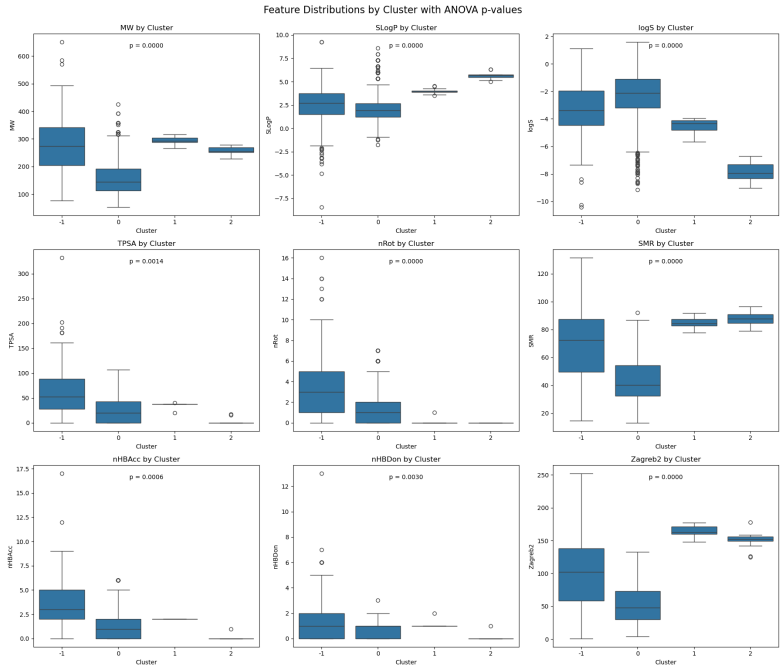
Graf 10: Lastnosti DBSCAN tališča

KMeans in aglomerativno, nista nič izboljšala rezultata.



Graf 11: DBSCAN za topnost

Zopet dobim le malo skupin, je pa ena zelo velika.



Graf 12: Lastnosti topnostnega DBSCANA

V tem primeru bi lahko rekli, da je grupiranje izvedeno glede na logS vrednost v kombinaciji z drugimi vrednostmi. Nastale skupine imajo distintkno različne logS vrednosti, torej je možno grupiranje glede

Grupiranja bi seveda lahko izboljšal, če bi uporabil večji set podatkov. Ni pa nujno, da so podatki povsem nepovezani, lahko bi več časa namenil obdelavi in priredbi za uporabo v algoritmu, a za to nisem imel dovolj časa. Drugače bi izvedel drugačno bolj sofisticirano normiranje in odstranil kup nepomembnih podatkov, ki motijo algoritme.

2. Naloga

Uporabite isti nabor organskih spojin kot v prejšnjih nalogah (vsaj 1000 spojin z deskriptorji in eksperimentalnimi lastnostmi). Na podlagi vektorske razdalje (npr. Evklidske ali Mahalanobisove) identificirajte spojine, ki se najbolj razlikujejo od povprečja celotnega nabora. Izračunajte središče podatkovne množice ter razdaljo vsake spojine do tega središča. Označite in interpretirajte najbolj oddaljene spojine (npr. top 1–5 %). Ali so to spojine z zelo nizko topnostjo, visoko lipofilnostjo ali nenavadnim številom funkcionalnih skupin? Pojasnite kemijski pomen odstopanj.

Metode

Analizo sem izvedel na istih treh podatkovnih podmnožicah kot v prejšnjih nalogah (različni nabori eksperimentalnih lastnosti in strukturnih deskriptorjev). Za vsak nabor sem izračunal Evklidsko in Mahalanobisovo razdaljo vsake spojine do središča podatkovne množice (tj. povprečnega vektorja lastnosti):

$$d_E(\mathbf{x}, \boldsymbol{\mu}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T (\mathbf{x} - \boldsymbol{\mu})}$$

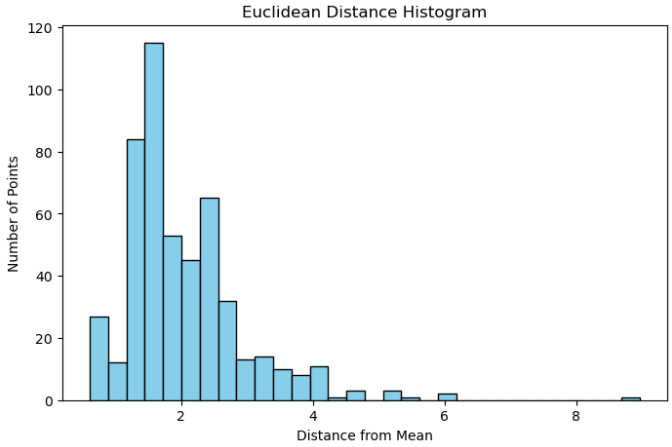
$$d_M(\mathbf{x}, \boldsymbol{\mu}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{S^{-1}} (\mathbf{x} - \boldsymbol{\mu})}$$

kjer je $\mathbf{x}$ vektor lastnosti posamezne spojine, $\boldsymbol{\mu}$ povprečni vektor (središče), $\mathbf{S}$ pa kovariančna matrika podatkov.

Na podlagi teh metrik sem identificiral najbolj oddaljene spojine (top 1–5 %), ter analiziral njihove skupne lastnosti – npr. izstopajoča topnost (logS), lipofilnost (LogP) ali ekstremna vrednost TPSA.

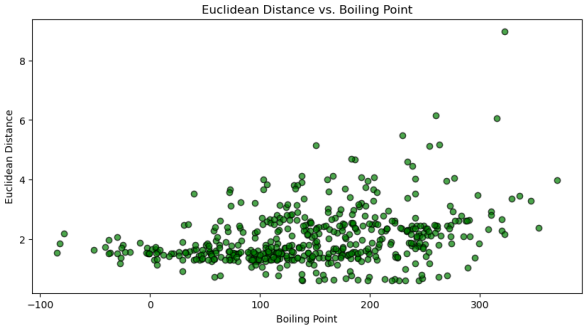
Rezultati

Izračunal sem povprečje lastnostnih vektorjev in odstopanja podatkovnih točk ter prikazal porazdelitev na histogramih.



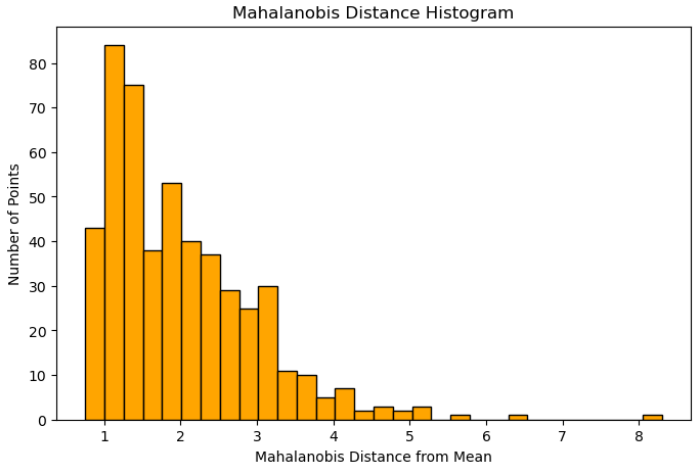
Slika 1: Evklidska norma odstopanja od povprečja

Razpršenost podatkov je precej velika, saj je večina podatkov dve enoti stran od povprečja. To lahko vzamemo v zakup, potem vidimo, da je nad tem vrhom le malo podatkov, ki zelo odstopajo od povprečja. Vseeno pa bo v 5% odstopanja od povprečja velika večina podatkov. Odnos sem preveril še grafično kot razdaljo od povprečja v odvisnosti od vrelišča.



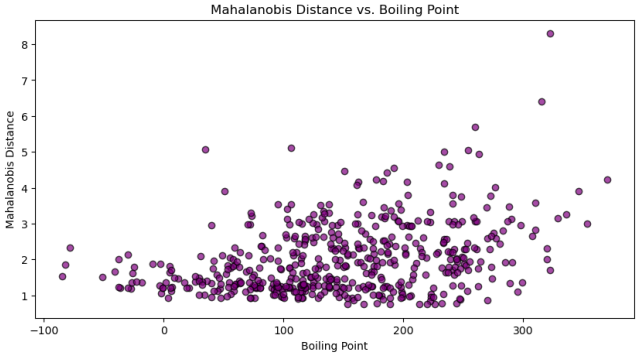
Graf 13: Oddaljenost od povprečja glede na vrelišče

Evklidska metrika je očitno slaba, če so podatki seveda statistično dobri, zato sem poskusil še Mahalanobisovo razdaljo.



Graf 14: Mahalanobisova razdalja od povprečja

S to metriko je porazdelitev boljša, večina podatkov leži bližje povprečju, a se velik del še vedno sipa izven povprečja. Pojavi pa se tudi nekaj zelo jasni osamelcev. To se vidi tudi na grafu razdalje v odvisnosti od vrelišča.



Graf 15: Mahalanobisova razdalja v odvisnosti od vrelišča

Najlepše pa se vidi, da osamelci pripadajo predvsem visokim vreliščem. V nadaljnjih analizah bi bilo najbolje izpustiti podatke z vreliščem nad 300°C. Naredil sem še test za osamelce in izstopajoče lastnosti.

Mahalanobis osamelec	Povprečje	Mediana
Negativno	2.409344	2.3040
Pozitivno	3.218692	3.5126

Tabela 3: logP (lipofilnost)

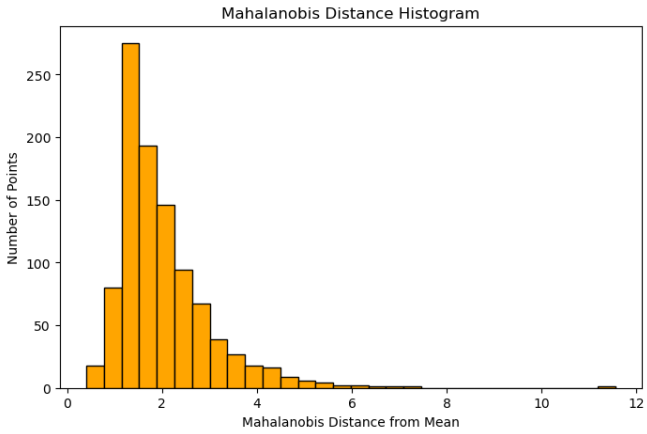
Označeni osamelci imajo višjo vrednost logP parametra, kar nakazuje, da so osamelci bolj lipofilni. Izbor osamelcev je 5% najbolj oddaljenih.

Mahalanobis osamelec	Povprečje	Mediana
Negativno	12.306253	0.00
Pozitivno	27.405200	35.53

Tabela 4: TPSA (polarnost)

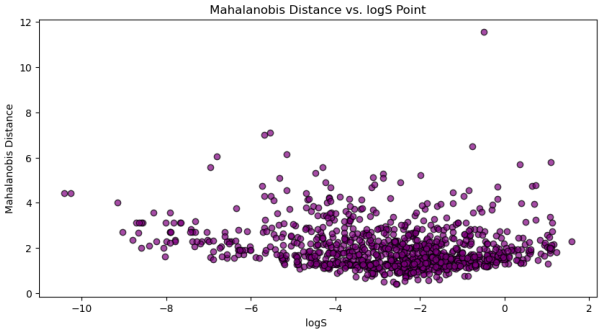
Osamelci so tudi bolj polarni od preostalih spojin. Povsem enako analizo sem opravil tudi za set spojin z danim tališčem, se rezultati ne razlikujejo dosti, zato jih nisem vključil.

Pri setu podatkov s podano topnostjo pride do veliko večjih razdalj od povprečja, analiziran trend za osamelce je tudi drugačen.



Graf 16:Mahalanobisova razdalja od povprečja

Porazdelitev je tukaj zelo enakomerna, se pa pojavijo zelo očitni osamelci na veliki razdalji.



Graf 17:Mahalanobisova razdalja v odvisnosti od topnosti

Tudi iz tega prikaza je vedina dobra porazdelitev podatkov z nekaj osamelci. Za treniranje na teh podatkih bi vse z razdaljo nad 5 izrezali, potem bi imeli zelo dober set podatkov.

Mahalanobis osamelec	Povprečje	Mediana
Negativno	2.409344	2.3040
Pozitivno	3.218692	3.5126

Tabela 5: logP (lipofilnost)

Pir topnosti so osamelci povprečno manj lipofilni od povprečja, je pa mediana večja, to nakazuje na nekaj slabih podatkov, ki so pokvarili našo bazo.

Mahalanobis osamelec	Povprečje	Mediana
Negativno	12.306253	0.00
Pozitivno	27.405200	35.53

Tabela 6: TPSA (polarnost)

Osamelci so v povprečju bolj polarni, a tu je tudi velika težava, saj je za večino podatkov mediana 0.