10.vaja: Strojno učenje

1. naloga

Usposobite regresijski model za napovedovanje molekularnih lastnosti organskih spojin, kot so tališče, vrelišče in topnost v vodi. Podatke pridobite iz javno dostopnih baz, kot je ChemSpider, pri čemer mora biti v naboru vsaj 1000 različnih organskih spojin, za katere so znani vsi trije ciljni podatki. S pomočjo knjižnice RDKit (Pythonova knjižnica za kemoinformatiko) izračunajte ustrezne kemijske deskriptorje (npr. molekulska masa, LogP, polarna površina, število rotacijskih vezi ipd.). Nato v Fortran ali C++ naredite multiplo regresijo na podlagi 80 % naključno izbranih spojin. Preostanek (20 %) uporabite za validacijo modela in analizo njegove napovedne uspešnosti (npr. z R^2 in RMSE).

D1. 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.

D2. 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.