Tema A4 Analiza Environmental data: Clustering

Tema obuhvata primenu *unsupervised multivariate clustering* metoda nad *environmental parameters data*. Podaci su prikupljeni iz realnog industrijskog pogona uz pomoć environment senzora i pouzdanog *data collection pipeline*. Cilj je primeniti poznate tehnike nad datim setom podataka, a potom protumačiti ceo proces zaključivanjem o dodatoj vrednosti iz izlaza.

Proces:

1. **Prvi korak** je istraživanje podataka. Pre svega je bitno razumeti kontekst i priču iza podataka da bi output bilo koje analize mogao da bude validan. Zato vršimo početno istraživanje podataka, koje služi kao određivanje osnovnih podataka o samim podacima, kao što je tip podataka, vrednosti, jedinice, a onda i dao analitike, gde spada vizualizacija, posmatranje *mean*, *median values*, *variation* i *standard deviation*, *histogram*, *distribution* podataka.
2. **Drugi korak** je preprocesiranje podataka. Nakon istraživanja podataka, već imamo bolji uvid u kontekst, pa lakše možemo odrediti i metode preprocesiranja. U osnovne metode spada pronalaženje i ispunjavanje missing values (NaN) ukoliko ih ima, otklanjanje duplikata, resampling (poželjno na frekvenciji od 1 Hz), normalizacija/standardizacija i na kraju potencijalno *dimensionality reduction*.
3. **Treći korak** je primena algoritama, u našem slučaju je to hierarhijsko klasterovanje. Cilj je proučiti, istražiti i primeniti ovu metodu nad sada prečišćenim podacima. Primenu za početak ostvariti bez specificiranja broja klastera, vizualizovati, ispitati dendrogram i distance. Isprobati više različitih metrika za povezivanje podataka, kao što su *euclidean*, *mahalanobis*, i dr. Evaluirati povezanost klasterovanja uz pomoć *cophenet correlation* tehnike (ili neke druge).
4. **Četvrti korak** je odrediti optimalan broj klastera. Metode za to nisu striktno određene, tako da imate mogućnost eksperimentisanja više različitih načina, kao na primer *Silhouette Score*, *Calinski-Harabasz*, nešto treće ili eventualno i neka vrsta agregacije.
5. **Peti korak** je primena algoritma sa dobijenim brojem klastera i evaluacija algoritama. Kad su u pitanju *unsupervised* metode kao klastering, obzirom da nemamo *true data values*, ovaj korak je malo teže precizno definisati. Osnovna metoda je vizualizacija, pored toga postoji više različitih metrika koje smisleno daju drugačije rezultate. Potrebno je primetiti neki obrazac po kome se podaci definišu i dele, pa kao takav opravdati u ponašanju realnih parametara, primer: primećeno je da period visoke temperature i niskog pritiska predstavlja posebno stanje sistema.

   Dodatno je da se Tema B4 bavi istim realnim procesom, istim metodama, nad istim vremenskim periodom, samo nad drugim setom podataka (energetski podaci), te je moguće uporediti rezultate.

Krajnji rezultat se očekuje u kao .csv fajl koji ima 2 kolone. Prva kolona predstavlja vremena (timestamp), dok su druge dve kolone niz brojeva u skladu sa tim koliko klastera je detektovano i kome koji podatak pripada.

Pored ovog dokumenta, u zip fajlu se nalazi environmental dataset par radova kao propratna literatura na temu konkretnih metoda za analizu koji će biti od pomoći za bolje razumevanje algoritma.

Ceo proces je poželjno dokumentovati na par strana (**dužina dokumenta nije bitna**) i dostaviti skriptu u kojoj su izvedeni svi ovi koraci.