ANALIZA ENVIRONMENTAL DATA: CLUSTERING

- PROJEKAT IZ PREDMETA „TEHNIKE I METODE ANALIZE PODATAKA“ U SARADNJI SA FIRMOM NISSATECH-

Student: Ivana Milivojević

Mentor: Bratislav Trojić

Niš, april 2024.

Izveštaj

Zadatak: Tema projekta obuhvata primenu klasterizacije nad „environmental“ podacima (prikupljenim iz realnog industrijskog pogona) i analizu dobijenih rezultata.

Implementacija: Projekat je realizovan u okviru Jupyter Notebook-a kroz pet glavnih koraka: istraživanje podataka, preprocesiranje podataka, upoznavanje sa algoritmom hijerarhijske klasterizacije, određivanje optimalnog broja klastera, i na kraju primena klasterizacije sa odabranim brojem klastera.

1. Istraživanje podataka

Na početku su učitani podaci iz CSV fajla sa 10 kolona, od kojih prva predstavlja vremenski trenutak očitavanja odgovarajućih parametara sa senzora, a preostalih 9 numeričke vrednosti (float64) tih parametara. Dostupno je ukupno 345 151 instanci podataka za period od 07.11.2022. do 10.11.2022. Takođe, može se primetiti da za atribut *CO2* ima manje vrednosti nego za ostale (344 264), kao i da nisu snimljene vrednosti svih parametara za ukupno 448 sekundi (skup sadrži podatke za 4 dana, tako da bi trebalo da ima 345 599 podataka). Na slici br. 1 je prikazano kako izgleda prvih pet podataka iz skupa, pri čemu se *timestamp* kolona koristi za indeksiranje.

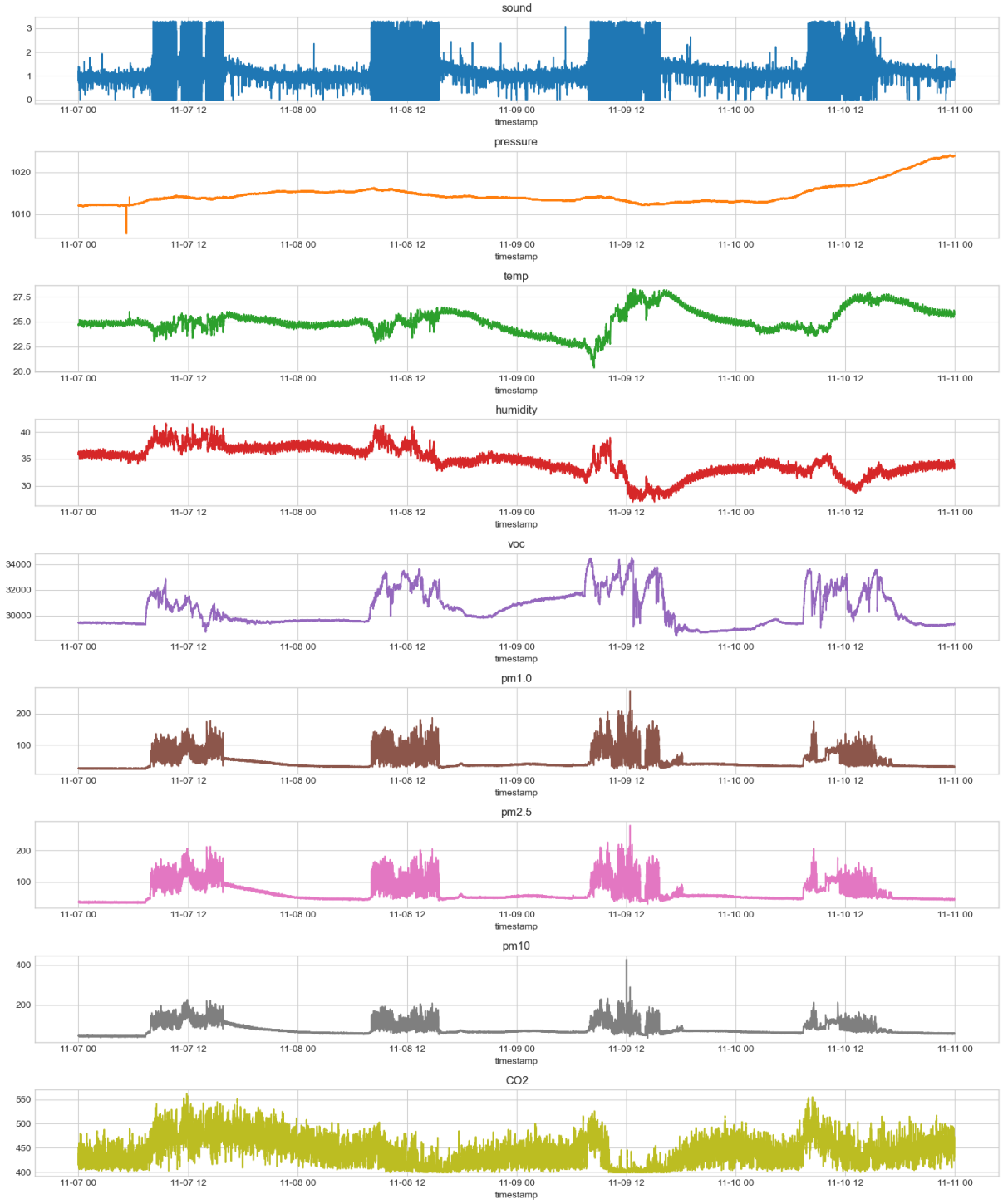
timestamp	sound	pressure	temp	humidity	voc	pm1.0	pm2.5	pm10	CO2
2022-11-07 00:00:01	0.881076	1012.037290	24.777757	35.996349	29466.0	26.0	36.0	45.0	421.0
2022-11-07 00:00:02	0.863325	1012.025698	24.762371	36.002125	29467.0	26.0	36.0	45.0	419.0
2022-11-07 00:00:03	0.908509	1012.083523	24.767500	36.013384	29459.0	26.0	36.0	45.0	418.0
2022-11-07 00:00:04	0.924645	1012.091207	24.767500	36.013384	29454.0	26.0	36.0	44.0	418.0
2022-11-07 00:00:05	0.896406	1012.091207	24.764935	36.013405	29458.5	26.0	36.0	44.0	416.0

Slika 1: Prvih pet uzoraka iz skupa podataka

Atributi u skupu podataka su:

1. sound – nivo buke [dB]?
2. pressure – vazdušni pritisak [mb]
3. temp – temperatura vazduha [°C]
4. humidity – vlažnost vazduha [%]
5. voc – koncentracija isparljivih organskih jedinjenja (*Volatile Organic Compounds*) [µg/m³]?
6. pm1.0 – koncentracija PM (*Particulate Matter*) 1.0 čestica, koje imaju prečnik manji od 1 µm [µg/m³]
7. pm2.5 – koncentracija PM 2.5 čestica, koje imaju prečnik manji od 2.5 µm [µg/m³]
8. pm10 – koncentracija PM 10 čestica, koje imaju prečnik manji od 10 µm [µg/m³]
9. CO2 – koncentracija ugljen-dioksida [ppm]

U okviru deskriptivne analize je primećeno da parametri *pm1.0, pm2.5* i *pm10* imaju veliku standardnu devijaciju i „repove“ na desnoj strani raspodele, odnosno da kod njih postoje vrednosti koje su značajno veće od ostalih. Takođe, za parametar *pressure* se može primetiti da je većina vrednosti bliska središnjoj, ali da postoje i vrednosti koje su povišene.



Slika 2: Vremenska raspodela vrednosti parametara

Na dijagramima vremenske raspodele vrednosti parametara (slika br. 2), može se primetiti da svi parametri sem *pressure* i *CO2* imaju jasnu periodičnost. Svakog dana je u periodu od 08:00 do 16:00h bilo neke aktivnosti, pri čemu se za parametar *pressure* primećuju varijacije

vrednosti 07.11. oko 05:00 h (nagli pad na 1005 mb), kao i porast tokom 10.11. Takođe, za dan 09.11. se primećuju nešto veće vrednosti parametara *temp*, *voc*, *pm1.0*, *pm2.5* i *pm10*, i manje vrednosti za *humidity*.

Na osnovu matrice korelacije (slika br. 3) može se primetiti da postoji jaka korelacija između atributa *pm1.0*, *pm2.5* i *pm10*, tako da bi se potencijalno mogla odraditi redukcija dimenzionalnosti izbacivanjem suvišnih atributa (ostaviti npr. atribut *pm2.5*).

Slika 3: Matrica korelacije

2. Preprocesiranje podataka

Utvrđeno je da za atribut CO2 postoji 887 nedostajućih (NaN) vrednosti i da u skupu podataka postoje 24 duplikata kada je u pitanju *timestamp*. Nedostajuće vrednosti zamenjene su srednjom vrednošću za taj parametar, a duplikati su uklonjeni.

Nakon toga je pozvana *resample* funkcija kako bi skup podataka imao vrednosti za svaku sekundu posmatranog intervala, a zatim je izvršena standardizacija podataka. S obzirom na to da su se javili problemi sa memorijom prilikom izvršavanja hijerarhijskog klasterovanja nad celim skupom podataka odmeravanim na 1s, proces je nastavljen sa skupom odmeravanim na 30s (slika br. 4), nakon čega ostalo 11 519 instanci.

timestamp	sound	pressure	temp	humidity	voc	pm1.0	pm2.5	pm10	CO2
2022-11-07 00:00:30	0.956112	1012.062426	24.739293	36.041862	29462.5	25.0	36.0	45.0	421.5
2022-11-07 00:01:00	0.964988	1011.991563	24.623900	36.195291	29463.0	26.0	38.0	47.0	440.0
2022-11-07 00:01:30	0.995648	1012.033767	24.670057	36.177992	29451.0	27.0	38.0	45.0	419.0
2022-11-07 00:02:00	0.914963	1012.072187	24.711086	36.149430	29450.0	26.0	37.0	47.0	435.0
2022-11-07 00:02:30	0.935941	1012.041327	24.741857	36.115291	29450.0	25.0	37.0	44.0	408.0

Slika 4: Podaci nakon odmeravanja na 30s

U ovom slučaju nije bilo potrebe za redukcijom dimenzionalnosti, s obzirom na to da skup sadrži svega 9 atributa, ali zbog postojanja jake korelacije ustanovljene u prethodnom koraku, izbačeni su parametri *pm1.0* i *pm10*.

3. Primena hijerarhijskog klasterovanja

Klasterizacija je tehnika nenadgledanog učenja kojom se podaci dele u grupe (klastere) na osnovu njihove sličnosti, pri čemu objekti treba da budu sličniji objektima iz istog klastera, nego objektima iz drugih klastera. Hijerarhijska klasterizacija je metoda kojom se kreira hijerarhija klastera tako što se rekurzivno vrši podela/udruživanje entiteta po top-down (*divisive hierarchical clustering*) ili bottom-up (*agglomerative hierarchical clustering*) principu.

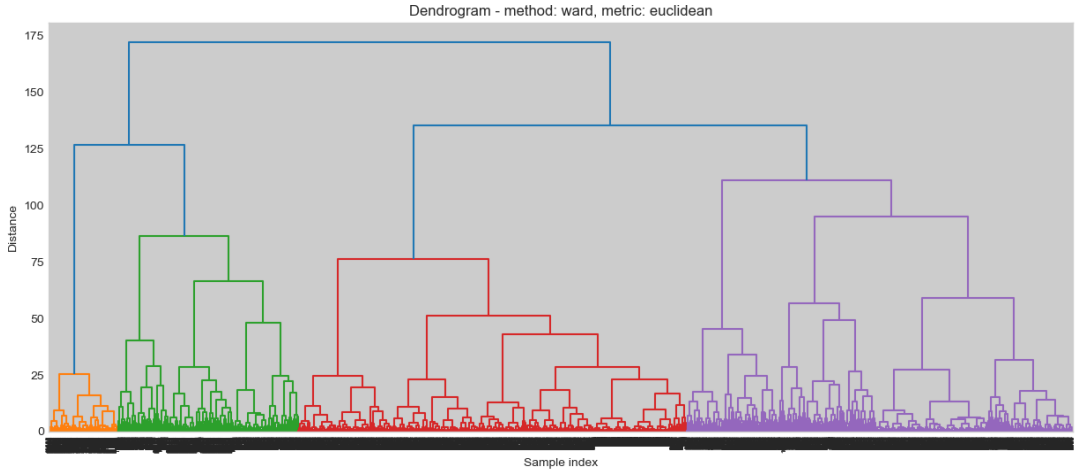
Najčešće se primenjuje *agglomerative hierarchical clustering*, koji je iskorišćen i u ovom projektu. Algoritam funkcioniše tako što se na početku svaki objekat posmatra kao pojedinačni klaster (list stabla), a zatim se u svakom koraku udružuju parovi najsličnijih klastera dok se ne dostigne jedinstveni klaster koji obuhvata sve objekte (koren stabla). Kao rezultat tog procesa se dobija stablo, poznato kao dendrogram. Najvažniji parametri koje treba zadati pre primene ovog algoritma su mera udaljenosti između objekata i način udruživanja klastera (odnosi se na odabir udaljenosti koju parovi klastera treba da minimizuju da bi bili odabrani za udruživanje). Takođe, potrebno je odrediti na kom mestu treba preseći stablo, tj. u koliko klastera treba podeliti podatke.

U ovom koraku je algoritam primenjen bez specificiranja broja klastera, jer je fokus bio na kreiranju dendrograma i razmatranju različitih metrika. Dendrogram je kreiran uz pomoć odgovarajućeg modula SciPy biblioteke (scipy.cluster.hierarchy). SciPy nudi dosta opcija za parametar *metric*: „braycurtis“, „canberra“, „chebyshev“, „cityblock“, „correlation“, „cosine“, „dice“, „euclidean“, „hamming“, „jaccard“, „jensenshannon“, „kulczynski1“, „mahalanobis“, „matching“, „minkowski“, „rogerstanimoto“, „russellrao“, „seuclidean“, „sokalmichener“, „sokalsneath“, „sqeuclidean“ i „yule“, od kojih su isprobane: „cityblock“, „euclidean“, „cosine“ i „mahalanobis“. Dostupne su različite vrednosti i za parametar *method*: „single“, „complete“, „average“, „weighted“, „centroid“, „median“ i „ward“ (pri čemu su „centroid“, „median“ i „ward“ ispravno definisane samo ako se koristi euklidsko rastojanje), a u ovom projektu su isprobane sve navedene mogućnosti.

Povezanost klastera je evaluirana korišćenjem *cophenet correlation* tehnike i vizuelnom analizom dendrograma. Za svaku razmatranu kombinaciju je računata *cophenet* mera korelacije između udaljenosti tačaka u prostoru atributa i udaljenosti na dendrogramu, pri čemu je klasterovanje bolje što je ona bliža jedinici (slika br. 5). Na početku su eliminisane kombinacije kod kojih gotovo svi entiteti na dendrogramu pripadaju jednom klasteru, jer to ne donosi novu informaciju o sistemu. Kao najbolja kombinacija odabran je slučaj kada se koriste „euclidean“ rastojanje i „ward“ metod (slika br. 6), gde *cophenet* korelacija iznosi 0,62.

Metric/Method	single	complete	average	weighted	centroid	median	ward
euclidean	0.603	0.677	0.795	0.681	0.78	0.643	0.618
cityblock	0.56	0.606	0.748	0.576	nan	nan	nan
cosine	0.35	0.532	0.673	0.528	nan	nan	nan
mahalanobis	0.624	0.579	0.777	0.705	nan	nan	nan

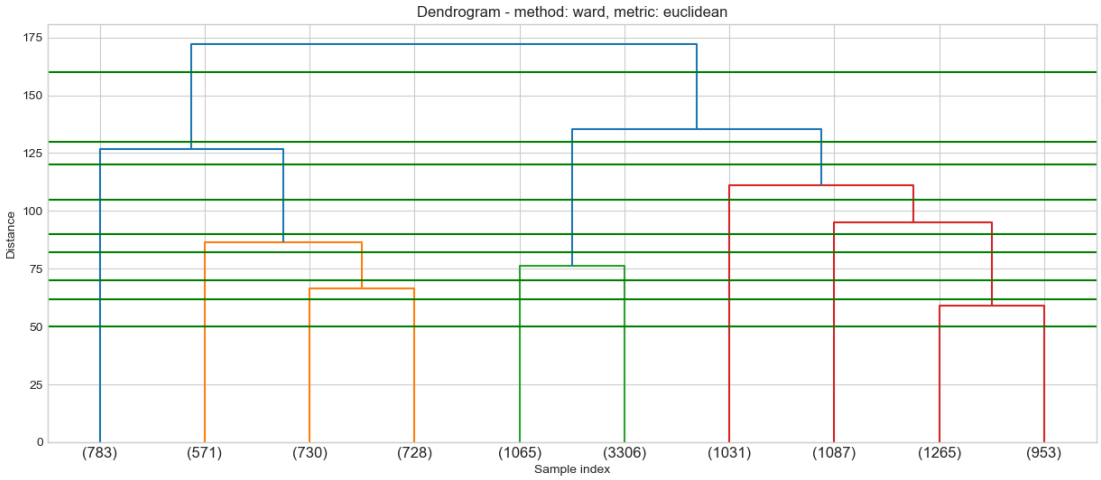
Slika 5: Vrednosti *cophenet* mere



Slika 6: Dendrogram koji odgovara odabranim parametrima

4. Određivanje optimalnog broja klastera

Dendrogram je ponovo vizuelizovan, sa izdvojenih „najviših“ 10 klastera (slika br. 7). U okviru ovog koraka primenjena su dva različita načina za određivanje optimalnog broja klastera: Silhouette score i Calinski-Harabasz score. Izračunate su njihove vrednosti za opseg od 2 do 10 klastera (slika br. 7), pri čemu je za dalju analizu odabrana podela na 4 klastera (odgovara odsecanju na visini 120).



Slika 7: Visine odsecanja dendrograma za 2-10 klastera

Score/Clusters	2	3	4	5	6	7	8	9	10
Silhouette	0.289	0.187	0.223	0.237	0.238	0.253	0.261	0.272	0.263
Calinski-Harabasz	2591.98	2435.72	2525.93	2589.97	2590.63	2606.03	2593.31	2550.78	2409.17

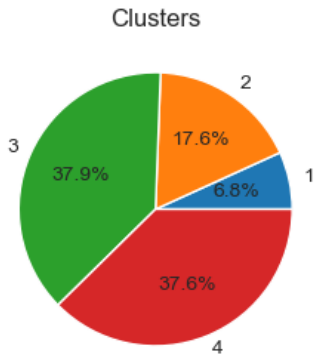
Slika 8: Evaluacija broja klastera

5. Primena i evaluacija algoritma

Izvršena je hijerarhijska klasterizacija na 4 klastera, pri čemu su dobijene evaluacione metrike:

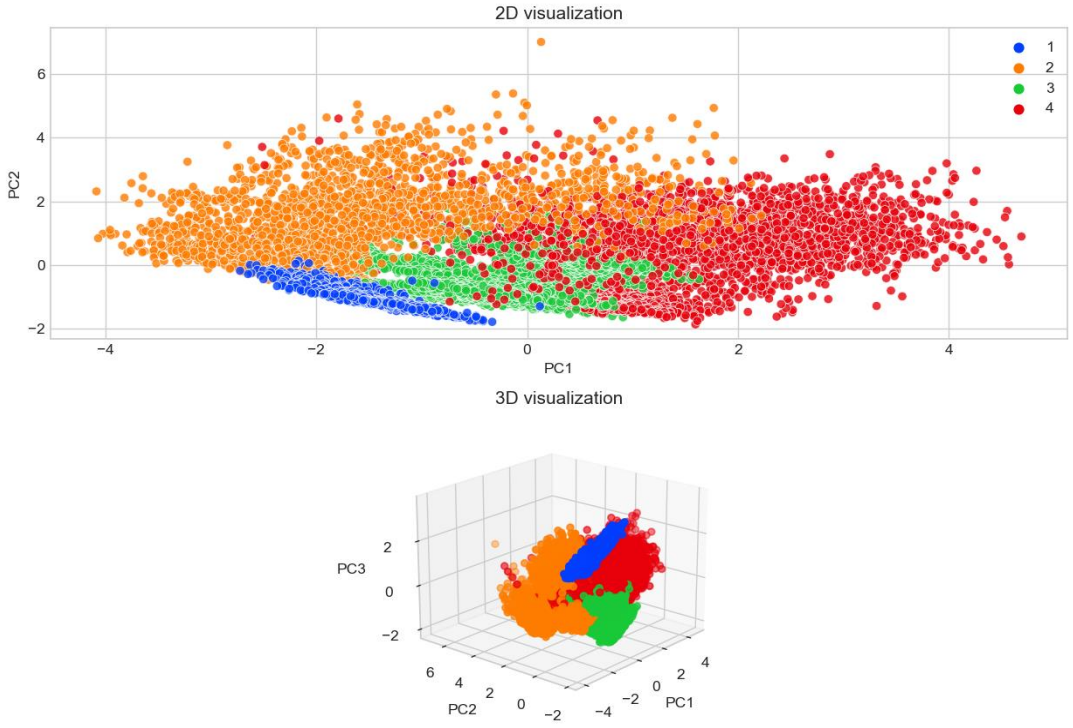
- Silhouette score: 0.22
- Calinski-Harabasz score: 2525.93
- Davies-Bouldin score: 1.47

Klasteru 1 pripada 783 entiteta, klasteru 2 pripada 2029, klasteru 3 pripada 4371, dok se u klasteru 4 nalazi 4336 entiteta (slika br. 9).



Slika 9: Raspodela podataka po klasterima

Za 2D i 3D vizuelizaciju je iskorišćena PCA metoda, kako bi se skup redukovao na 2, odnosno 3 dimenzije, što je prikazano na slici br. 10.



Slika 10: Vizuelizacija klastera

Nakon klasterizacije je primenjena LASSO metoda za određivanje značaja atributa, tj. uticaja parametara na pripadnost klasteru. Ustanovljeno je da najviše utiču vlažnost, pritisak, temperatura, koncentracija ugljen-dioksida i koncentracija PM čestica, redom.

- Klaster 1: Povećan pritisak, povišena temperatura, blago povećana koncentracija CO2, blago povećana buka
- Klaster 2: Smanjena vlažnost, povišena temperatura, povećana koncentracija PM čestica, blago povecana koncentracija voc
- Klaster 3: Normalno stanje sistema sa prosečnim vrednostima parametara
- Klaster 4: Povećana vlažnost, povećana koncentracija CO2, povećana koncentracija PM čestica

Primećeno je da period povišenog pritiska i povišene temperature predstavlja posebno stanje sistema u kome su blago povećane koncentracija CO2 i buka (klaster 1), da posebno stanje sistema predstavlja period povećane koncentracije CO2 i PM čestica u uslovima povećane vlažnosti (klaster 4), kao i da je u uslovima povišene temperature i smanjene vlažnosti povećana koncentracija PM čestica i lako isparljivih jedinjenja (klaster 2).