

**Analiza Environmental data: Clustering**

- Projekat iz predmeta „tehnike i metode analize podataka“ u saradnji sa firmom nissatech-

**Student:** Ivana Milivojević

**Mentor:** Bratislav Trojić

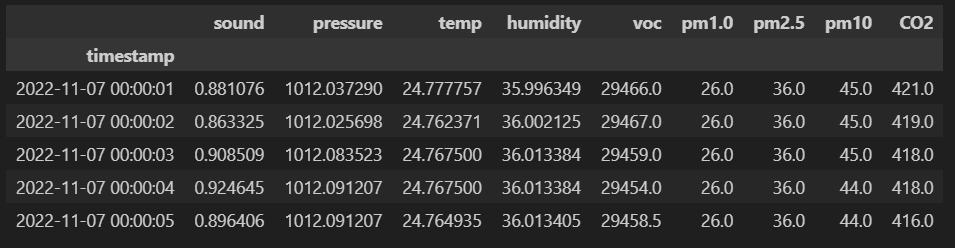
Niš, april 2024.

# **Izveštaj**

Zadatak: Tema projekta obuhvata primenu metoda 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 ovog algoritma sa odabranim brojem klastera i evaluacija rezultata.

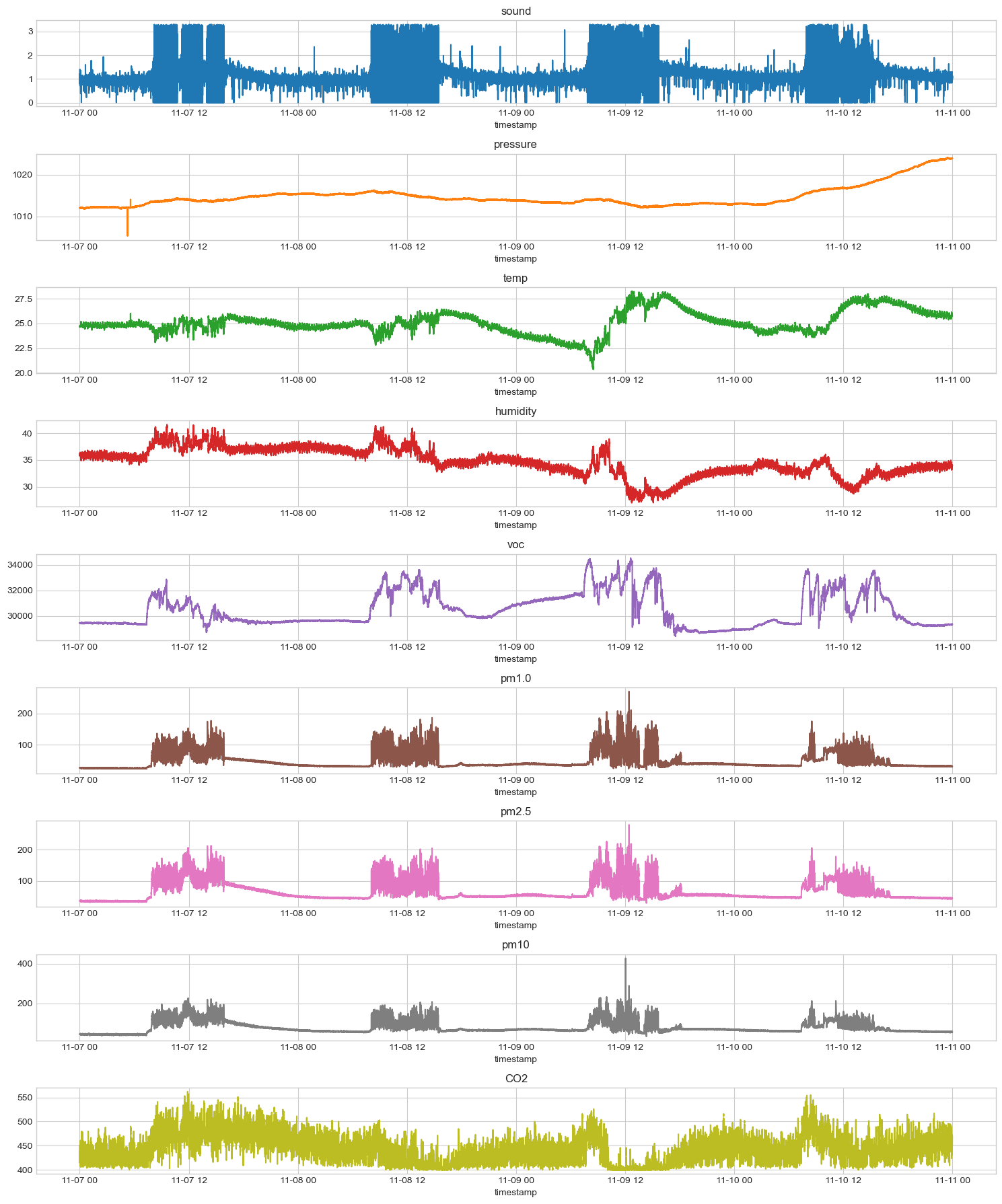
1. **Istraživanje podataka**

Na početku implementacije 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 oko 450 sekundi (skup sadrži podatke za 4 cela dana, što je 345 600 sekundi). Na slici br. 1 je prikazano kako izgleda prvih pet podataka iz skupa, pri čemu se *timestamp* kolona koristi za indeksiranje.

Slika 1: Prvih pet podataka iz skupa podataka

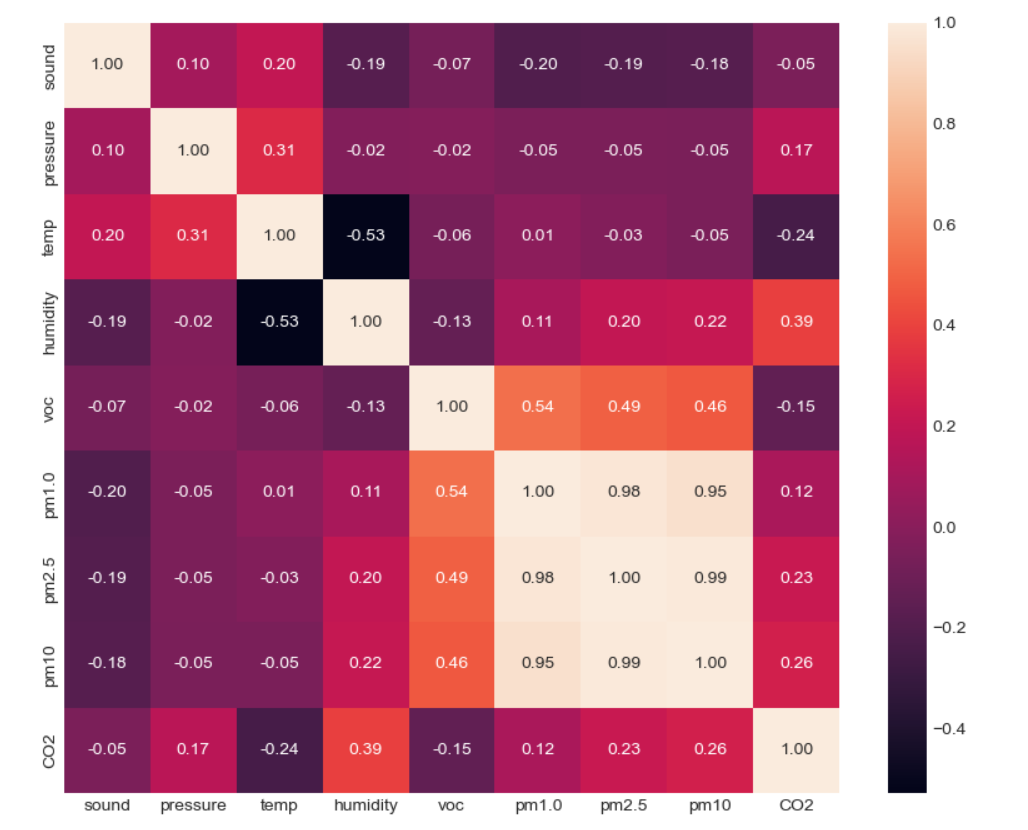
Atributi u skupu podataka su:

1. sound – nivo buke [dB]?
2. pressure – atmosferski 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 16:00 h 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 veoma jaka linearna korelacija između atributa *pm1.0*, *pm2.5* i *pm10*, tako da bi se mogla odraditi redukcija dimenzionalnosti izbacivanjem suvišnih atributa (ostaviti npr. atribut *pm10*).

Slika 3: Matrica korelacije

1. **Preprocesiranje podataka**

Utvrđeno je da za atribut CO2 postoji 887 nedostajućih (NaN) vrednosti i da u skupu podataka postoji 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, pokušan je isti proces sa skupom odmeravanim na 30s.

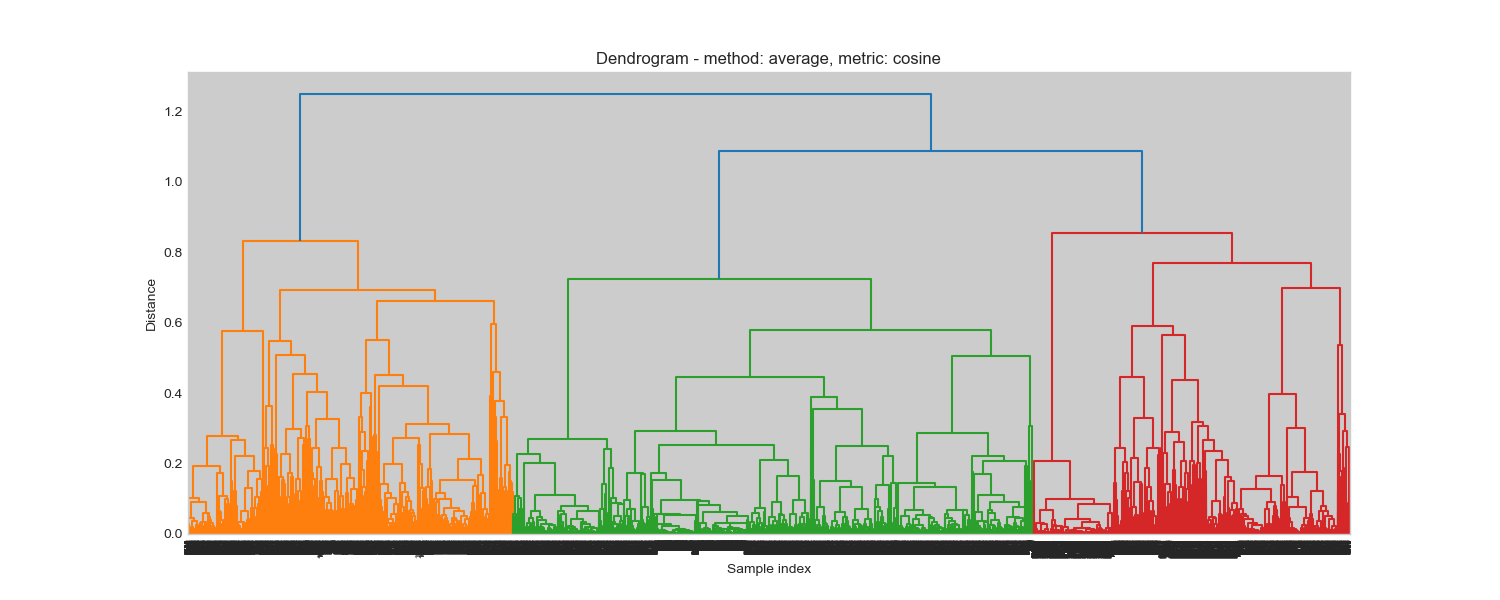
U ovom primeru nije bilo potrebe za redukcijom dimenzionalnosti, s obzirom na to da skup sadrži svega 9 atributa. Međutim, pošto je utvrđeno da postoji jaka korelacija između parametara *pm1.0*, *pm2.5* i *pm10*, razmotreno je izbacivanje *pm1.0* i *pm2.5*.

1. **Primena hijerarhijskog klasterovanja**

Klasterizacija je tehnika nenadgledanog učenja kojom se podaci dele u grupe (klastere) na osnovu njihove sličnosti, pri čemu su objekti 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 sličnosti (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“. U ovom projektu 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. U ovom projektu su isprobane sve mogućnosti.

Povezanost klastera je evaluirana korišćenjem cophenet correlation tehnike. 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. 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 „cosine“ rastojanje i „average“ metod (slika br. 4), gde cophenet korelacija iznosi 0,72. Alternativno se mogu razmotriti kombinacije koje koriste „euclidean“ rastojanje i „complete“ ili „ward“ metod, gde su dobijeni cophenet koeficijenti 0.69 i 0.70, respektivno.

Slika 4: Dendrogram koji odgovara najboljim parametrima

1. **Određivanje optimalnog broja klastera**

U okviru ovog koraka je 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, pri čemu su obe metrike imale najbolje vrednosti za 2, 3 i 4 klastera. Klasterizacija je u sledećem koraku pokušana sa 3 i 4 klastera.

1. **Primena i evaluacija algoritma**