



# 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.

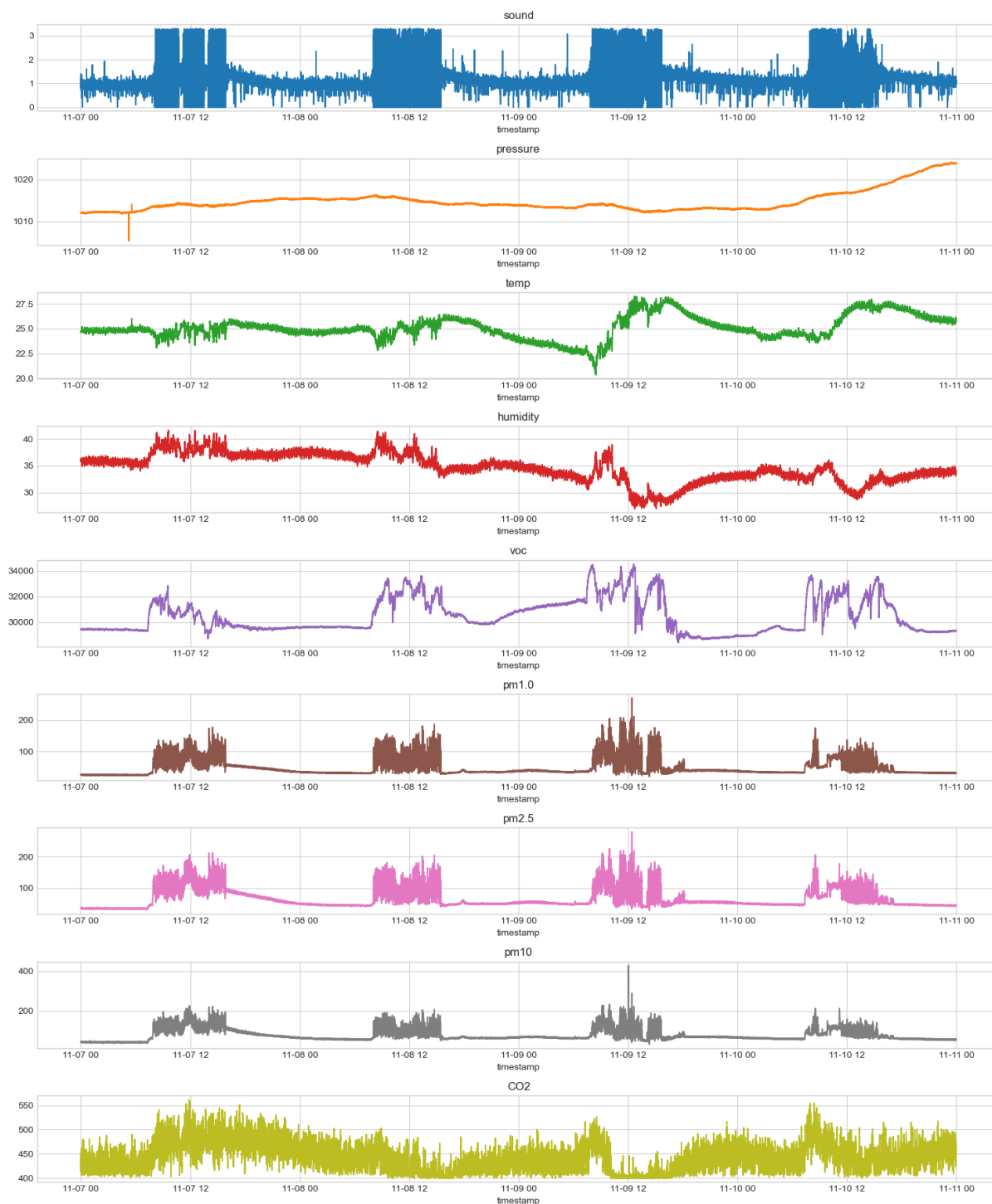
|                     | sound    | pressure    | temp      | humidity  | voc     | pm1.0 | pm2.5 | pm10 | CO2   |
|---------------------|----------|-------------|-----------|-----------|---------|-------|-------|------|-------|
| timestamp           |          |             |           |           |         |       |       |      |       |
| 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 [V]
2. pressure – vazdušni pritisak [mb]
3. temp – temperatura vazduha [°C]
4. humidity – vlažnost vazduha [%]
5. voc – koncentracija isparljivih organskih jedinjenja (*Volatile Organic Compounds*) [ppm]
6. pm1.0 – koncentracija PM (*Particulate Matter*) 1.0 čestica, koje imaju prečnik manji od 1  $\mu\text{m}$  [ $\mu\text{g}/\text{m}^3$ ]
7. pm2.5 – koncentracija PM 2.5 čestica, koje imaju prečnik manji od 2.5  $\mu\text{m}$  [ $\mu\text{g}/\text{m}^3$ ]
8. pm10 – koncentracija PM 10 čestica, koje imaju prečnik manji od 10  $\mu\text{m}$  [ $\mu\text{g}/\text{m}^3$ ]
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 su zamenjene primenom linearne interpolacije 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.

|                     | sound    | pressure    | temp      | humidity  | voc     | pm1.0 | pm2.5 | pm10 | CO2   |
|---------------------|----------|-------------|-----------|-----------|---------|-------|-------|------|-------|
| timestamp           |          |             |           |           |         |       |       |      |       |
| 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 (klaster) 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.

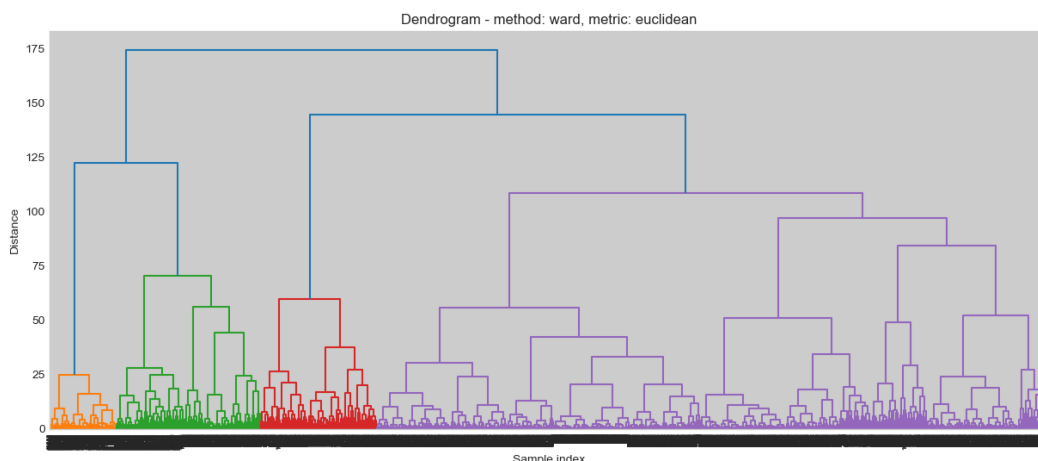
Najčešće se primenjuje *agglomerative hierarchical clustering*, koji je iskorišćen i u ovom projektu. Algoritam funkcioniše tako što se na početku svaki objekat posmatra kao pojedinačni klaster (list stabla), a zatim se u svakom koraku udružuju parovi najbližijih klastera dok se ne dostigne jedinstveni 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 je isprobano nekoliko vrednosti koje najčešće daju dobre rezultate („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 iako su imale visoke vrednosti za *cophenet* koeficijent (preko 0.7) – 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 0,68 (ova kombinacija daje „jasne“ klasterne na dendrogramu).

| Metric/Method | single | complete | average | weighted | centroid | median | ward  |
|---------------|--------|----------|---------|----------|----------|--------|-------|
| euclidean     | 0.604  | 0.612    | 0.794   | 0.66     | 0.801    | 0.58   | 0.676 |
| cityblock     | 0.56   | 0.687    | 0.748   | 0.604    | nan      | nan    | nan   |
| cosine        | 0.35   | 0.568    | 0.664   | 0.608    | nan      | nan    | nan   |
| mahalanobis   | 0.624  | 0.575    | 0.761   | 0.598    | nan      | nan    | nan   |

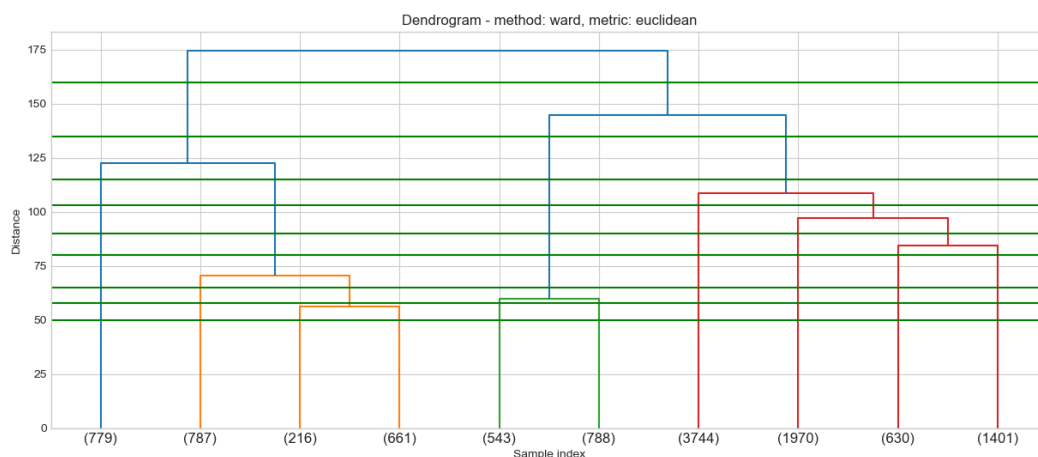
Slika 5: Vrednosti *cophenet* mere



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 primenjene su dve različite metrike za određivanje optimalnog broja klastera: Silhouette score i Calinski-Harabasz score (njihove vrednosti za opseg od 2 do 10 klastera prikazane su na slici br. 8). Odabrana je podela na 4 klastera, pri čemu je prvenstveno uzeta u obzir vrednost za Silhouette score – koja je najbolja za 4 klastera, dok Calinski-Harabasz score ima najbolju vrednost za 6 klastera, ali i vrednost dobijena za 4 klastera se ne razlikuje značajno od najbolje.



Slika 7: Visine odsecanja dendrograma za 2-10 klastera

| Score/clusters    | 2       | 3       | 4      | 5      | 6       | 7       | 8       | 9       | 10      |
|-------------------|---------|---------|--------|--------|---------|---------|---------|---------|---------|
| Silhouette        | 0.3     | 0.301   | 0.322  | 0.217  | 0.242   | 0.269   | 0.279   | 0.28    | 0.284   |
| Calinski-Harabasz | 2683.01 | 2693.54 | 2689.2 | 2711.8 | 2743.91 | 2737.05 | 2667.84 | 2567.95 | 2489.36 |

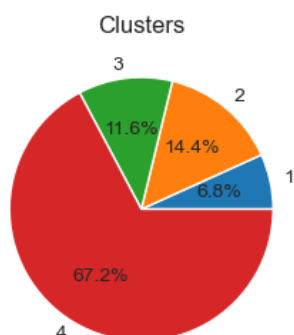
Slika 8: Evaluacija broja klastera

## 5. Primena i evaluacija algoritma

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

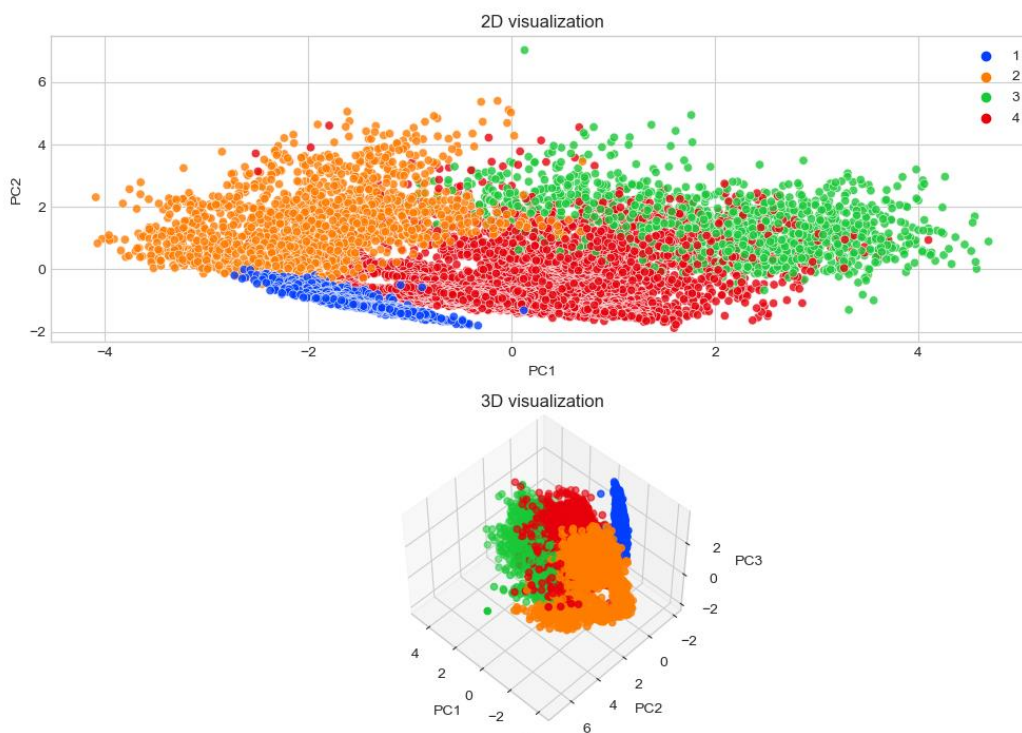
- Silhouette score: 0.32
- Calinski-Harabasz score: 2689.2
- Davies-Bouldin score: 1.27

Klasteru 1 pripada 779 entiteta, klasteru 2 pripada 1664, klasteru 3 pripada 1331, dok se u klasteru 4 nalazi 7745 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 Random Forest tehnika za određivanje značaja atributa, tj. uticaja parametara na pripadnost klasteru. Ustanovljeno je da najviše utiču buka, pritisak, temperatura, vlažnost, koncentracija PM čestica, koncentracija isparljivih organskih jedinjenja, pa koncentracija ugljen-dioksida.

- Klaster 1: Povećan pritisak, povišena temperatura, blago povećana koncentracija CO<sub>2</sub>
- Klaster 2: Povećana buka, povišena temperatura, smanjena vlažnost, blago povećana koncentracija PM čestica, blago povećana koncentracija isparljivih organskih jedinjenja
- Klaster 3: Značajno smanjena buka, povećana vlažnost, povećana koncentracija PM čestica, blago povećana koncentracija isparljivih organskih jedinjenja, blago povećana koncentracija CO<sub>2</sub>
- Klaster 4: Normalno stanje sistema bez značajnih promena parametara

Primećeno je da period povećanog pritiska i povišene temperature predstavlja posebno stanje sistema u kome je blago povećana koncentracija ugljen-dioksida (klaster 1), da posebno stanje sistema predstavlja period povišene temperature i smanjene vlažnosti u kome je povećana buka i blago su povećane koncentracije PM čestica i isparljivih organskih jedinjenja (klaster 2), kao i da postoji period povećane vlažnosti u kome je značajno smanjena buka, značajno je povećana koncentracija PM čestica i blago je povećana koncentracija isparljivih organskih jedinjenja i ugljen-dioksida (klaster 3).