# *Gručenje in razlaga gruč*

Začnimo s primerom. Imam podatke o večini držav sveta, njih skoraj 200, ki jih popišemo s socioekonomskimi značilkami. Značilk imamo, recimo, veliko, vsaj nekaj deset. Vzorec takih podatkov je na primer v tabeli spodaj, a dejansko je naša podatkovna matrika veliko večja.

| Država | Pričakovana življenjska doba | Povprečno število let šolanja | BDP na prebivalca | Delež urbanega prebivalstva | Necepljeni dojenčki (ošpice) |
|---|---|---|---|---|---|
| Avstralija | 82.5 | 13.2 | 42 822.0 | 89.4 | 7.0 |
| Čile | 82.0 | 9.9 | 21 665.0 | 89.5 | 6.0 |
| Kuba | 79.6 | 11.8 | 7455.0 | 77.1 | 1.0 |
| Danska | 80.4 | 12.7 | 44 519.0 | 87.7 | 10.0 |
| Grčija | 81.1 | 10.5 | 24 808.0 | 78.0 | 3.0 |
| Slovenija | 80.6 | 12.1 | 28 664.0 | 49.7 | 6.0 |
| Švica | 83.1 | 13.4 | 56 364.0 | 73.9 | 7.0 |
| Venezuela | 74.4 | 9.4 | 15 129.0 | 89.0 | 11.0 |

Tabela 5: Primer profiliranja držav s socioekonomskimi podatki.

Nad takimi podatki lahko izvedemo gručenje, torej postopek, kjer podatke razdelimo v skupine (gruče) glede na njihovo podobnost. V tem poglavju bomo postopke gručenja skušali predstaviti sistematično, a šele kasneje, saj jih je, prvič, bralec skozi svoje dosedanje šolanje najbrž že spoznal, in drugič, ker se bomo najprej posvetili postopkom razlage gruč. Tretjič pa zato, ker smo nekatere postopke za iskanje gruč že spoznali v prejšnjem poglavju, ko smo gradili podatkovne karte oziroma smo podatke skušali predstaviti v dvorazsežnem razsevnem diagramu. Primer takšne podatkovne karte je prikazan na sliki 12.

Na podatkovni karti s slike 12 lahko razberemo nekaj skupin. Izbrali smo eno izmed njih, z osmimi državami, in želimo vedeti, kaj je tem državam skupnega. Imamo nekaj možnosti:

- Imena držav lahko izpišemo. To bo za osem držav najbrž prva stvar, ki jo lahko naredimo, problem pa bi bil, če bi bila izbrana

Slika 12: Države v grafu t-SNE. Za gradnjo grafa smo uporabili podatkovni nabor *Human Development Index* iz leta 2014 z 50 značilkami. Med državami na grafu smo izbrali manjšo skupino.

skupina večja in manj pregledna.

- Pogledamo, kje so te države na zemljevidu. Zemljevid nam ponuja dodatno informacijo, ki sicer ni vsebovana v podatkih, a nam pride prav, sploh če nam geografija ni tuja.
- Lahko si pomagamo s podatki o skupinah držav, na primer mediteranske, azijske, južnoameriške, in potem ugotovimo, ali večina držav, ki smo jih izbrali, pripada kakšni od teh skupin.
- Razmišljamo lahko, v katerih značilnostih so izbrane države različne od vseh ostalih. Še najbolj prav bi nam tu prišel urejen seznam atributov, od atributov, kjer se izbrane države najbolj ločijo od vseh ostalih, do atributov, kjer je ta ločitev slabša ali pa je ni. Seveda nam bodo tu pomagala socioekonomska znanja za razumevanje tega urejenega seznama.
- Če imamo atributov veliko, bi nam lahko pomagala ureditev atributov v skupine. Recimo, določili bi lahko atribute, ki so povezani z zdravstvom, pa s finančnimi kazalci, šolstvom in podobnim. Potem bi morda lahko za izbrane države ugotovili, da so to med najpremožnejšimi državami na svetu ali pa med državami z najbolj razvitim šolstvom in podobno.

Zgoraj našteti so različni načini razlage skupin. Pri nekaterih smo uporabili vedenje o (izbranih) primerih, pri drugih pa smo skušali izbrane primere (diferencialno) opisati z uporabo podatkov. Pri večini nekoliko boljših razlag nam je prav prišlo neko dodatno znanje oziroma informacije, ki niso bile prisotne v sami tabeli podatkov. Takemu znanju pravimo tudi domensko znanje. Pri vseh zgornjih idejah, razen morda pri prvih dveh, bomo sicer morali uporabiti neke računske postopke za rangiranje atributov ali pa ugotavljanje obogatenosti skupin. In prav o teh bo govor v nadaljevanju besedila.

## *Analiza obogatenosti skupin*

Poenostavimo zgornji primer in predpostavimo, da smo v podatkih imeli samo 12 držav in med njimi našli skupino petih, kot to kaže tabela 6. Opazimo, da je med izbranimi državami kar nekaj mediteranskih. Pravzaprav je v tabeli pet mediteranskih držav (Francija, Grčija, Hrvaška, Italija, Španija), med njimi pa so kar štiri države iz našega izbora. V našem izboru je tudi Portugalska, ki ni mediteranska država. Vprašamo se, ali bi tako veliko (ali večje) število mediteranskih držav v našem izboru dobili tudi po naključju. Torej, kakšna je verjetnost, da bi pri naključnem izboru petih držav izmed vseh dvanajstih izbrali vsaj štiri mediteranske države?

Pojavu, da se neka skupina v izboru pojavlja pogosteje, kot bi pričakovali pri naključnem izboru iz celotne množice, pravimo obogatenost skupine.

Pomagajmo si s kombinatoriko. Naši podatki vsebujejo končno množico $N = 12$ držav, med katerimi je $K = 5$ mediteranskih držav. Iz te množice smo izbrali $n = 5$ držav, pri čemer je bilo $k = 4$ mediteranskih. Verjetnost, da v izboru dobimo natanko $k$ mediteranskih držav, je enaka

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

Ta izraz je znan kot hipergeometrična porazdelitev, ki opisuje verjetnosti pri vzorčenju brez vračanja iz končne množice. V našem primeru slučajna spremenljivka $X$, ki šteje število mediteranskih držav v izboru, sledi hipergeometrični porazdelitvi s parametri $N$, $K$ in $n$.

V hipergeometrični porazdelitvi je $\binom{K}{k}$ število načinov, kako izmed vseh $K$ mediteranskih držav izberemo $k$ držav, $\binom{N-K}{n-k}$ število načinov, kako izmed vseh nemediteranskih držav izberemo preostalih $n - k$ držav, $\binom{N}{n}$ pa število vseh možnih izborov $n$ držav iz celotne množice $N$ držav. Ulomek tako predstavlja delež vseh možnih izborov, v katerih je natanko $k$ mediteranskih držav.

V našem primeru smo opazili $X = 4$, saj so med petimi izbranimi državami štiri mediteranske. Ker nas zanima, ali je to nenavadno velik delež, računamo verjetnost, da bi po naključju dobili vsaj štiri mediteranske države:

$$P(X \geq 4) = P(X = 4) + P(X = 5).$$

Dobimo

$$P(X = 4) = \frac{\binom{5}{4}\binom{7}{1}}{\binom{12}{5}} = \frac{5 \cdot 7}{792} = \frac{35}{792},$$

in

$$P(X = 5) = \frac{\binom{5}{5}\binom{7}{0}}{\binom{12}{5}} = \frac{1 \cdot 1}{792} = \frac{1}{792}.$$

Skupaj je torej

$$P(X \geq 4) = \frac{35}{792} + \frac{1}{792} = \frac{36}{792} = \frac{1}{22} \approx 0{,}0455.$$

To pomeni, da bi pri povsem naključnem izboru petih držav izmed dvanajstih dobili vsaj štiri mediteranske države z verjetnostjo približno 4,5%. Tak rezultat je torej razmeroma malo verjeten in kaže na to, da so mediteranske države v našem izboru obogatene glede na to, kar bi pričakovali po naključju.

Za dodatno orientacijo lahko pogledamo še pričakovano število mediteranskih držav v naključnem izboru petih držav. Za hipergeometrično porazdelitev velja, da je pričakovana vrednost enaka

$$E(X) = n\frac{K}{N} = 5 \cdot \frac{5}{12} = \frac{25}{12} \approx 2{,}08.$$

Tabela 6: Države in njihov izbor. Izbrane države so označene v koloni "Izbor" z 1, ostale z 0.

| Država | Izbor |
|---|---|
| Avstrija | 0 |
| Francija | 1 |
| Grčija | 1 |
| Hrvaška | 0 |
| Italija | 1 |
| Nemčija | 0 |
| Nizozemska | 0 |
| Portugalska | 1 |
| Poljska | 0 |
| Španija | 1 |
| Švedska | 0 |
| Švica | 0 |

V naključnem izboru bi torej v povprečju pričakovali nekaj več kot dve mediteranski državi, opazili pa smo kar štiri. Tudi tako vidimo, da je naš rezultat drugačen od pričakovanega pri naključnem izboru.

Seveda bi namesto mediteranske skupine lahko vzeli tudi kakšno drugo skupino držav. Recimo, države evrobmočja. Med dvanajstimi državami v naših podatkih je takih držav devet, med petimi izbranimi državami pa so vse članice evroobmočja. Če označimo z $X$ število držav evroobmočja v izboru, velja $N = 12$, $K = 9$, $n = 5$ in $k = 5$. Verjetnost, da bi pri naključnem izboru petih držav dobili same države evroobmočja, je

$$P(X = 5) = \frac{\binom{9}{5}\binom{3}{0}}{\binom{12}{5}} = \frac{126}{792} \approx 0{,}159.$$

Tak rezultat torej ni posebno malo verjeten, zato v tem primeru ne bi mogli govoriti o izraziti obogatenosti.

Obogatenost je torej pojav, ko se elementi določene skupine v izboru pojavljajo pogosteje, kot bi pričakovali po naključju. Analiza obogatenosti pa je postopek, s katerim to odstopanje kvantitativno ovrednotimo in preverimo, ali ga lahko pripišemo naključju ali pa kaže na dejanski vzorec v podatkih.

## *Analiza obogatenosti v kodi*

Z analizo obogatenosti lahko torej razložimo, kateri skupini primerov pripadajo izbrani primeri tako, da je ta pripadnost (močno) drugačne od naključne. Za to poleg podatkov pripravimo domensko znanje v obliki skupin, za naš primer na primer v zapisu, kot je prikazan na sliki 13.

Za izračun obogatenosti bomo uporabili funkcijo `hypergeom.sf` iz knjižnice `scipy.stats`, kjer je `sf` tako imenovana funkcija preživetja (angl. *survival function*) in nam vrne verjetnost, da je slučajna spremenljivka $X$ večja od določene vrednosti $k$.

Slika 13: Zapis v obliki YAML s primerom določitve skupin držav.

```
Mediteranske:
  - Francija
  - Grčija
  - Hrvaška
  - Italija
  - Španija

Balkanske:
  - Grčija
  - Hrvaška

Zahodnoevropske:
  - Francija
  - Nemčija
  - Nizozemska
  - Avstrija
  - Švica
```

```
def enrichment(all_items, selected_items, group):
    group = set(group) & all_items

    N = len(all_items)
    n = len(selected_items)
    K = len(group)
    k = len(group & selected_items)

    p_value = hypergeom.sf(k - 1, N, K, n) if K > 0 else 1.0
    expected = n * K / N if N > 0 else 0.0
    fold = (k / expected) if expected > 0 else float("nan")

    return {
```

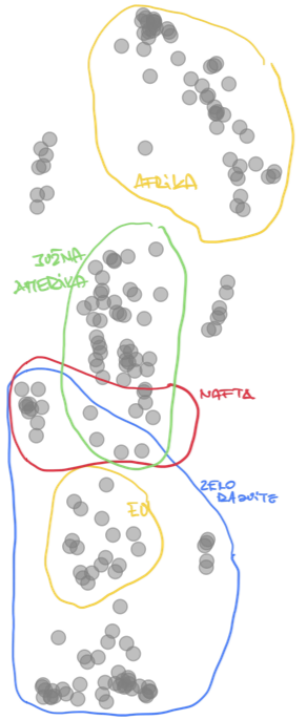
```
        "K": K,
        "k": k,
        "expected": expected,
        "fold": fold,
        "p_value": p_value,
    }
```

Vhod v našo funkcijo obogatenosti `enrichment` so množica vseh objektov, množica izbranih objektov in skupina, za katero računamo obogatenost oziroma verjetnost, da bi izbrani objekti pripadali tej skupini, če bi bil izbor naključen. Funkcija vrne parametra porazdelitve $K$ in $k$, pričakovano vrednost (koliko elementov iz dane skupine bi pričakovali med izbranimi objekti pri naključnem izboru) faktor obogatenosti (razmerje dejanske in pričakovane vrednosti), in končno še $p$-vrednost, ki ocenjuje statistično značilnost opaženega prekrivanja.

V glavnem delu naše implementacije preberemo podatke o izboru, podatke o skupinah držav, izračunamo obogatenost in izpišemo rezultate:

```
import yaml
import pandas as pd

df = pd.read_excel("izbrane-države.xlsx")

all_items = set(df["Država"].astype(str).str.strip())
selected_items = set(
    df.loc[df["Izbor"] == 1, "Država"].astype(str).str.strip()
)

with open("skupine-držav.yaml", "r", encoding="utf-8") as f:
    groups = yaml.safe_load(f)

results = []
for name, group in groups.items():
    r = enrichment(all_items, selected_items, group)
    r["group"] = name
    results.append(r)

results.sort(key=lambda x: x["p_value"])

for r in results:
    print(
        f"{r['group']+":":20s} "
        f"k={r['k']:2d}, K={r['K']:2d}, "
        f"E={r['expected']:.2f}, "
        f"fold={r['fold']:.2f}, "
        f"P={r['p_value']:4f}"
    )
```

Statistična značilnost opisuje, kako verjetno je, da bi opažen rezultat nastal zgolj zaradi naključja. To ocenimo s $p$-vrednostjo: manjša kot je, manj verjetno je, da je rezultat posledica naključnega izbora.




Slika 14: Primer možne anotacije t-SNE razporeditve držav v dvodimenzionalno karto. Bi znali tako razlago karte sprogramirati?

Analiza obogatenosti za naš izbor držav in skupine držav, s katerimi predstavimo domensko znanje, torej vrne po $p$-vrednosti urejen seznam skupin:

```
Mediteranske:        k= 4, K= 5, E=2.08, fold=1.92, P=0.0455
Evroobmočje:         k= 5, K= 9, E=3.75, fold=1.33, P=0.1591
Obmorske:            k= 5, K=10, E=4.17, fold=1.20, P=0.3182
Balkanske:           k= 1, K= 2, E=0.83, fold=1.20, P=0.6818
Zahodnoevropske:     k= 1, K= 5, E=2.08, fold=0.48, P=0.9735
```

Pravzaprav nas zanimajo samo skupine, ki so statistično značilne in kjer je njihova naključna verjetnost zelo majhna. Za to pa rabimo določiti prag, običajno označen z $\alpha$ (npr. $\alpha = 0{,}05$), pod katerim $p$-vrednost štejemo kot dovolj majhno. Skupine z $p < \alpha$ tako obravnavamo kot obogatene, ostale pa kot posledico naključnega izbora.

Parameter $\alpha$ določa verjetnost napake tipa I, torej da napačno razglasimo rezultat za statistično značilen, čeprav je v resnici posledica naključja.

Kako se potem lotim razlage morebitnih skupin s slike 12? Lahko z izdelavo interaktivne vizualizacije, kjer bi uporabniku omogočili izbor točk oziroma držav iz vizualizacije in potem prikazali obogatene skupine. Še boljša pa bi bila avtomatična prepoznava gruč točk na vizualizaciji in potem anotacija teh gruč z imeni obogatenih skupin. Nekaj podobnega (sicer izmišljeni) razlagi s slike 14.

## *S skupinami povezane zvezne značilke*

Vrnimo se k našemu izboru skupine držav s slike 12, a se tokrat osredotočimo na njihove lastnosti. Radi bi ugotovili, v katerih značilkah se izbrana skupina razlikuje od vseh drugih. V podatkih, iz katerih smo zgradili vizualizacijo, je bilo 50 značilk, zato bi bilo smiselno dobiti urejen seznam, kjer so na vrhu tiste, ki so z našim izborom najbolj povezane. Potrebujemo torej neko cenilko povezanosti skupine z informativnostjo značilke, ki zavzame večje vrednosti za značilke, pri katerih se izbrane države najbolj razlikujejo od preostalih.

Začnimo sicer s hipotetičnim primerom in za tri značilke, imenovali smo jih kar $A$, $B$, in $C$, prikažimo, kakšna je njihova porazdelitev vrednosti v izbrani množici držav in kako so te vrednosti porazdeljene v vseh ostalih državah (slika 15). Katera značilka je najbolj informativna? Katera torej najbolje loči primere iz izbrane gruče od vseh ostalih primerov? Porazdelitvi sta ločeni pri značilkah $A$ in $B$, a je prekrivanje pri značilki $B$ manjše. Prekrivanje je veliko pri značilki $C$.

Očitno nam o izbrani skupini največ informacij poda značilka $B$. Intuitivno lahko rečemo, da so boljše tiste značilke, pri katerih sta porazdelitvi čim bolj odmaknjeni druga od druge in hkrati čim manj razpršeni. Odmaknjenost lahko ocenimo z razliko med povprečnima vrednostma obeh skupin, razpršenost pa z varianco. Ker želimo eno samo cenilko, ki upošteva oboje, lahko uporabimo $t$-statistiko. Ta meri razliko med povprečjema glede na razpršenost podatkov v obeh

Statistiko $t$ je uvedel William Sealy Gosset leta 1908. Objavil jo je pod psevdonimom Student v članku "The probable error of a mean", saj je delal za pivovarno Guinness, ki svojim zaposlenim ni dovoljevala objavljanja znanstvenih del pod lastnim imenom. Zato danes govorimo o Studentovi t-porazdelitvi in t-testu.

skupinah. Za značilko $X$ jo zapišemo kot

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

kjer sta $\bar{x}_1$ in $\bar{x}_2$ povprečji v izbrani in preostali skupini, $s_1^2$ in $s_2^2$ varianci, $n_1$ in $n_2$ pa velikosti obeh skupin. Večja absolutna vrednost $t$ pomeni večjo ločljivost med skupinama in s tem večjo informativnost značilke.

Za rangiranje značilk bo absolutna vrednost $t$ dovolj, če pa želimo oceniti tudi statistično značilnost opažene razlike, $t$ pretvorimo v $p$-vrednost. To dobimo iz $t$-porazdelitve kot verjetnost, da bi ob predpostavki, da razlike med skupinama ni, dobili tako veliko ali še večjo vrednost $|t|$. Pri tem predpostavimo, da $t$ sledi $t$-porazdelitvi z ustreznim številom prostostnih stopenj (npr. $\nu \approx n_1 + n_2 - 2$) in izračunamo repno verjetnost. Za dvostranski test velja

$$p = 2 \cdot P(T \geq |t|).$$

Manjša kot je $p$-vrednost, manj verjetno je, da je opažena razlika posledica naključja. V Pythonu to izračunamo kot:

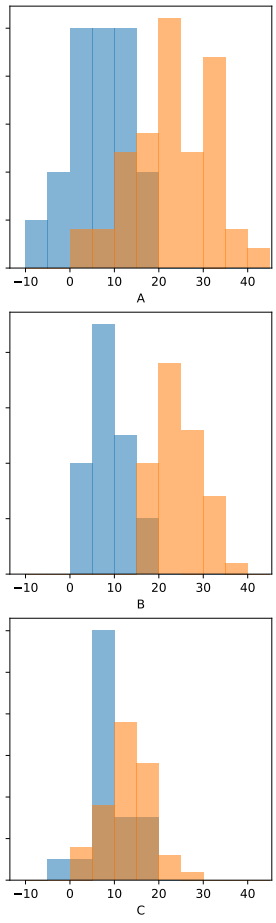
```
from scipy.stats import t
p_value = 2 * t.sf(abs(t_stat), df)
```

S kodo, ki bi primerno prebrala podatke, nam omogočila izbor neke skupine držav in potem izračunala $t$-statistiko oziroma njeno pripadajočo $p$-vrednost bi lahko dobili izpis, katerega primer je na primer spodaj:

```
0.0000 v Neenakost v življenjski dobi (%)
0.0000 v Umrljivost dojenčkov (na 1000 rojstev)
0.0006 v Mladi brez šole in zaposlitve (%)
0.0012 v Dojenčki izključno dojeni (%)
0.0014 v Neenakost v dohodku (%)
0.0294 ^ Delež žensk v parlamentu (%)
0.0769 ^ Plačan porodniški dopust (dni)
```

Ta poleg $p$-vrednosti tudi pove, ali je vrednost značilke v izbrani skupini manjša ali večja. Seveda bi bilo zelo primerno, če bi za tako analizo razvili primerni uporabniški vmesnik, ki bi nam na enostaven način omogočil raziskovanje podatkov in njihovih skupin.

Slika 15: Porazdelitev spremenljivk $A$, $B$ in $C$ v izbrani skupini primerov (oranžna) in v vseh ostalih primerih (modra).




### *Kaj pa, če so spremenljivke kategorične?*

Zgornje predpostavlja, da so vse značilke zvezne. Kaj pa, če so te kategorične (recimo spol, regija ali tip države)? V tem primeru lahko

namesto cenilke $t$ uporabimo $\chi^2$-statistiko, ki meri odstopanje med opaženimi in pričakovanimi frekvencami v kontingenčni tabeli.

Predpostavimo, da je med 12 državami 8 članic OECD, 4 pa niso, in da tabela podatkov (glej tabelo ) vsebuje značilko, ki o tem članstvu poroča. Naj bodo v izbrani skupini petih držav, kjer so 4 članice OECD in 1 ni. Ta razmerja lahko predstavimo v kontingenčni tabeli 7.

Tabela 7: Kontingenčna tabela članstva v OECD.

| | Izbrane | Ostale |
|---|---|---|
| OECD | 4 | 4 |
| Ne-OECD | 1 | 3 |
| Skupaj | 5 | 7 |

Če članstvo v OECD z izborom ne bi bilo povezano, bi pričakovane frekvence izračunali kot

$$E_{ij} = \frac{(\text{vsota vrstice}) \cdot (\text{vsota stolpca})}{\text{skupna vsota}}.$$

Tako za članice OECD v izbrani skupini dobimo

$$E = \frac{8 \cdot 5}{12} = 3{,}33,$$

za nečlanice OECD pa

$$E = \frac{4 \cdot 5}{12} = 1{,}67.$$

Odstopanje med opaženimi in pričakovanimi frekvencami bomo kvantitativno izmerili s statistiko $\chi^2$, ki združuje prispevke vseh celic kontingenčne tabele. Statistiko $\chi^2$ izračunamo kot

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

Statistiko $\chi^2$ je uvedel Karl Pearson leta 1900 in jo predstavil v članku *On the criterion that a given system of deviations...* kot metodo za preverjanje ujemanja med opaženimi in pričakovanimi frekvencami.

kjer je $O_{ij}$ opažena, $E_{ij}$ pa pričakovana frekvenca. V našem primeru dobimo

$$\chi^2 = \frac{(4-3{,}33)^2}{3{,}33} + \frac{(4-4{,}67)^2}{4{,}67} + \frac{(1-1{,}67)^2}{1{,}67} + \frac{(3-2{,}33)^2}{2{,}33} \approx 0{,}69.$$

Večja kot je vrednost $\chi^2$, večje je odstopanje med opaženimi in pričakovanimi frekvencami, in s tem močnejša povezanost med kategorično značilko in izbrano skupino. Če želimo oceniti še statistično značilnost, vrednost $\chi^2$ pretvorimo v $p$-vrednost s porazdelitvijo $\chi^2$ z ustreznim številom prostostnih stopenj, a to načeloma smemo storiti le, če so pričakovane frekvence v vseh celicah dovolj velike (praviloma vsaj okoli 5).

V Pythonu se za izračun našega primera poslužimo naslednje kode:

```
import numpy as np
from scipy.stats import chi2_contingency

O = np.array([
```

```
    [4, 4],
    [1, 3]
])

chi2_stat, p_value, df, E = chi2_contingency(O, correction=False)
print(f"chi2 = {chi2_stat:.4f}")
print(f"p    = {p_value:.4f}")
```

Funkcija `chi2_contingency` vrne vrednost $\chi^2$, njeno pripadajočo $p$-vrednost, število prostostnih stopenj ter matriko pričakovanih frekvenc. Program vrne,

```
chi2 = 0.6857
p    = 0.4076
```

in lahko sklepamo, da članstvo v OECD ni značilka, ki bi bila povezana z našim izborom držav.

## *Tudi spremenljivke lahko razvrstimo v skupine*

Razlaga skupine z rangiranjem spremenljivk zahteva seveda interpretacijo. Dobro moramo poznati pomen spremenljivk in vedeti, na kaj se te nanašajo ter interpretirati rezultate skladno z $p$-vrednostmi in vrstnim redom spremenljivk v rangu. Huh, že prejšnji stavek je precej dolg in morda kompliciran. Interpretacija torej ni enostavna.

Kaj pa, če si tudi tu pomagamo z domenskim znanjem, in na primer spremenljivke uredimo v skupine? Na primer tako, kot je to prikazano na sliki 16. Naš cilj je tu ugotoviti, ali so za skupino izbranih držav značilne značilke, ki pripadajo določeni skupini značilk, in potem poročamo o obogatenosti teh skupin. Pri tem združimo vse, kar smo že uvedli v tem poglavju, torej rangiranje značilk, njihov izbor (glede na neko zgornjo mejo $\alpha$ za $p$-vrednost), in izračun obogatenosti skupin značilk. Izhod postopka so torej verjetnosti, da so izbrane skupine značilk v opazovani gruči zastopane bolj, kot bi pričakovali po naključju.

V zgornjem postopku smo uvedli nov parameter, mejo $\alpha$. Temu se lahko izognemo. Namesto izračuna obogatenosti na zgornji način lahko za dano skupino značilk opazujemo, kje v seznamu rangiranih značilk se pojavljajo značilke iz te skupine. Zanimajo nas primeri, kjer je večina značilk iz skupine pri vrhu urejenega seznama (prisotnost) ali pa pri dnu (odsotnost).

Cenilka, ki jo uvedemo za tak izračun, se imenuje Mann–Whitneyjeva statistika, ki je sicer sorodna iz strojnega učenja bolj znani površini pod ROC krivuljo (t. i. *area under the curve*, oz. AUC). Ta meri, kako pogosto so značilke iz dane skupine uvrščene višje od značilk izven skupine, in jo lahko interpretiramo kot verjetnost, da bo naključno

Slika 16: Socioekonomske značilke razvrščene v skupine.

```
Otroci in mladina:
  - Otroško delo (%)
  - Mladi brez šole in zaposlitve (%)
  - Podhranjenost otrok(%)
  - Umrljivost otrok do 5 let (na 1000)
  - Dojenčki izključno dojeni (%)
  - Dojenčki brez cepljenja (DTP) (%)
  - Dojenčki brez cepljenja (ošpice) (%)

Zdravje in neenakost:
  - Neenakost v življenjski dobi (%)
  - Indeks življenjske dobe

Gospodarska struktura:
  - Zaposleni v kmetijstvu (%)
  - Zaposleni v storitvah (%)
  - Delež mestnega prebivalstva (%)
```

Mann–Whitneyjev test in pripadajočo statistiko sta leta 1947 predlagala Henry B. Mann in Donald R. Whitney kot neparametrično alternativo $t$-testu, ki ne predpostavlja normalnosti podatkov in temelji na rangih.

izbrana značilka iz skupine rangirana višje od naključno izbrane značilke izven skupine. Vrednosti statistike blizu 1 pomenijo koncentracijo značilk na vrhu seznama, vrednosti blizu 0 na dnu, vrednost okoli 0,5 pa ustreza naključni razporeditvi.

Poglejmo kratek primer. Recimo, da imamo osem rangiranih značilk,

$$x_1,\ x_2,\ x_3,\ x_4,\ x_5,\ x_6,\ x_7,\ x_8,$$

urejenih od najbolj do najmanj informativne. Imejmo skupino značilk

$$G = \{x_1, x_3, x_4\},$$

preostale značilke pa označimo z $\bar{G} = \{x_2, x_5, x_6, x_7, x_8\}$.

Mann–Whitneyjeva statistika primerja vse pare $(g, o)$, kjer je $g \in G$ in $o \in \bar{G}$, ter šteje, kolikokrat je značilka iz skupine uvrščena višje (ima manjši rang) kot značilka izven skupine. V našem primeru je takih parov $3 \cdot 5 = 15$. Število "zmag" skupine je

$$U = 5 + 4 + 4 = 13,$$

saj je $x_1$ pred vsemi petimi, $x_3$ pred štirimi (ne pa pred $x_2$), in $x_4$ prav tako pred štirimi. Število zmag normaliziramo z številom vseh možnih parov, in dobimo

$$\text{AUC} = \frac{U}{|G| \cdot |\bar{G}|} = \frac{13}{15} \approx 0{,}867.$$

To pomeni, da je verjetnost, da bo naključno izbrana značilka iz skupine rangirana više od naključno izbrane značilke izven skupine približno 86,7%, kar kaže na izrazito prisotnost skupine na vrhu seznama. Implementacija AUC ima linearno kompleksnost (sprehod po urejenem seznamu), če ne upoštevamo sortiranja, lahko pa uporabimo tudi tudi implementacijo iz knjižnice `scipy.stats`:

```
from scipy.stats import mannwhitneyu

group_ranks = [1, 3, 4]
other_ranks = [2, 5, 6, 7, 8]

U, p = mannwhitneyu(group_ranks, other_ranks, alternative="less")
auc = 1 - U / (len(group_ranks) * len(other_ranks))

print(f"U = {U}")
print(f"AUC = {auc:.3f}")
print(f"p = {p:.4f}")
```

Zgornja koda vrne tudi $p$-vrednost, ki pove, kako verjetno je, da bi tako ugoden razpored značilk dobili po naključju.

Možen izhod take analize bi bil seznam, kot je prikazan v tabeli 8. Ta na primer pokaže, da so izbrane države posebne z vidika zdravja,

Tabela 8: Obogatenost skupin značilk v socioekonomskih podatkih.

| AUC | $p$ | Skupina |
|---|---|---|
| 1.000 | 0.0008 | zdravje_in_neenakost |
| 0.738 | 0.0434 | neenakost |
| 0.721 | 0.0321 | demografija |
| 0.678 | 0.0846 | izobraževanje |

neenakosti, demografije in izobraževanja, kar je sicer precej abstraktna, a morda primerna razlaga. Za njeno razumevanje pa vsekakor potrebujemo dodatno znanje, predvsem o tem, ali so ta področja zastopana pozitivno ali negativno in na kakšen način.

Analiza te vrste bi morala vključevati tudi informacijo o zaželeni smeri značilk (npr. višji BDP je boljši, nižja stopnja brezposelnosti je boljša ipd.), vse to pa bi morali smiselno vključiti v uporabniški vmesnik, ki omogoča sledljivost oziroma "vrtanje v globino" — od abstraktnih rezultatov do konkretnih podatkov, s katerimi lahko pojasnimo, zakaj in kako smo do teh rezultatov prišli.

Tak pristop je skladen z načeli FAIR (angl. *Findable, Accessible, Interoperable, Reusable*), ki poudarjajo preglednost, dostopnost in ponovno uporabnost podatkov ter analiz. Več o FAIR v Wilkinson MD in sod. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*.

## *Kaj pa gručenje?*

Zgoraj smo se razpisali o razlagah gruč in malce tudi o za to potrebnih uporabniških vmesnikih. Za slednje naj sicer pripomnimo, da so taki, ki bi zares dovoljevali odlično razlago gruč, precej redki in jih najdemo bolj v specialističnih orodjih, torej teh, ki so namenjeni analizam podatkov iz izbranih domen. A nazaj na gručenje. Predpostavili bomo, da bralec področje dobro pozna, tudi iz prejšnjih predmetov, in tu samo omenili ključne pristope. Zato tu samo podamo tabelo glavnih pristopov, in opišemo nekaj izbranih metod.

| Skupina | Metode | Prednosti | Slabosti |
|---|---|---|---|
| **Particijske** | k-means, k-medoids | hitre, preproste, lahko za zelo veliko primerov | določitev $k$, občutljive na inicializacijo, sferične gruče |
| **Hierarhične** | aglomerativno, delitveno | vizualizacija z dendrogramom | počasne, primerne samo za manjše podatke |
| **Gostotne** | DBSCAN, OPTICS | gruče so poljubne oblike, obravnavajo šum, izločijo osamelce | močno občutljive na meta-parametre metode |
| **Modelne** | Gaussove mešanice | probabilistične, primeri so lahko uvrščeni v več gruč | predpostavke porazdelitve podatkov, lokalni optimumi |
| **Na omrežjih** | Louvain, Leiden | primerne za omrežja, skupnosti | odvisne od konstrukcije grafa |
| **Projekcijske / vgraditvene** | t-SNE, PCA, spektralno gručenje | zaznajo lahko kompleksno strukturo, vizualizacija | niso neposredno namenjene iskanju gruč, potrebna je dodatna analiza |

Tabela 9: **Glavne skupine metod gručenja.** Pregled ključnih pristopov z značilnimi metodami ter njihovimi osnovnimi prednostmi in slabostmi.

## *Hierarhično gručenje*

Hierarhično gručenje iz podatkov zgradi hierarhijo gruč: ob inicializaciji predpostavi, da je vsak primer svoja gruča, nato pa algoritem postopno združuje najbolj podobne gruče, dokler ne ostane

ena sama. Vhod v metodo je množica primerov, opisanih z atributi, ter izbira mere razdalje med primeri (npr. evklidska, manhattanska, kosinusna) in mere razdalje med gručami (npr. *single*, *complete*, *average linkage*, *Ward*). Pri *single linkage* gledamo najmanjšo razdaljo med elementoma dveh gruč, pri *complete linkage* največjo, pri *average linkage* povprečje vseh parov, pri Wardovi metodi pa povečanje znotrajgručne variance po združitvi. Rezultat je možno všečno prikazati kot dendrogram, ki je drevesni prikaz zaporedja združevanj; če dendrogram "prerežemo" na izbrani višini, dobimo konkretno razdelitev primerov v več gruč. Problem tega prikaza je, da ga lahko z vrtenjem posameznih vej prikažemo na različne načine.

Hierarhično gručenje praviloma ne optimizira ene same globalne cenilke za celotno rešitev, ampak gradi zaporedje lokalno najboljših združitev glede na izbrano mero razdalje. Glavna cenilka v postopku je torej razdalja med primeri oziroma gručami, pri Wardovi metodi pa povečanje vsote kvadratov odklonov od centroida. Prednost metode je, da ne potrebujemo vnaprej določiti števila gruč in da rezultat dobimo v pregledni hierarhični obliki, slabost pa računska zahtevnost ter občutljivost na izbiro razdalje, načina povezovanja in skaliranje atributov.

### *Metoda voditeljev*

Metoda voditeljev (k-means) je particijska metoda gručenja, ki podatke razdeli v vnaprej določeno število $K$ gruč. Vsako gručo predstavlja njen voditelj, to je centroid oziroma povprečni vektor primerov v gruči. Vhod v metodo je množica primerov, opisanih s številskimi atributi, izbrano število gruč $K$ in mera razdalje (najpogosteje evklidska). Algoritem začne z izbiro začetnih voditeljev, nato pa iterativno ponavlja dva koraka: vsak primer priredi najbližjemu voditelju, nato pa voditelje posodobi kot centroidne vrednosti pripadajočih gruč. Postopek se konča, ko se razbitje ne spreminja več oziroma se spremembe ustalijo.

Izhod metode je razbitje primerov na $K$ gruč in pripadajoči voditelji. Metoda optimizira kompaktnost gruč, to je vsoto kvadratov razdalj primerov do njihovih voditeljev (SSE):

$$\sum_{i=1}^{K} \sum_{x \in C_i} \|x - v^{(i)}\|^2,$$

pri čemer so optimalni voditelji enaki centroidom gruč. Algoritem praviloma hitro konvergira, vendar le do lokalnega optimuma, zato je rezultat odvisen od začetne izbire voditeljev.

Prednost metode je njena učinkovitost in primernost za velike podatkovne množice, slabosti pa so potreba po vnaprejšnji določitvi $K$,

občutljivost na inicializacijo in mera razdalje ter dejstvo, da metoda najbolje deluje za približno kroglaste in po velikosti primerljive gruče.

### *Silhueta*

Silhueta je cenilka kakovosti razbitja podatkov na skupine, ki hkrati upošteva kohezijo znotraj gruč in ločljivost med njimi. Uporabimo jo lahko za oceno kakovosti razvrstitve pri danem številu skupin $K$, ali pa za izbor ustreznega $K$ pri metodah, kot sta metoda voditeljev ali hierarhično gručenje. Vhod v metodo je razbitje podatkov na skupine in izbrana mera razdalje med primeri, izhod pa je silhuetni koeficient $s$, ki zavzame vrednosti približno med $-1$ in 1.

Silhueto izračunamo za vsak primer $x^{(i)}$ iz povprečne razdalje do vseh primerov v njegovi skupini, $a_i$, ter najmanjšo povprečno razdaljo do primerov v kateri koli drugi skupini, $b_i$. Silhueta primera je nato določena kot

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}.$$

Vrednosti blizu 1 pomenijo, da je primer dobro umeščen v svojo skupino, vrednosti okoli 0 nakazujejo na prekrivanje skupin, negativne vrednosti pa na napačno razvrstitev.

Silhueta razbitja je povprečje silhuet vseh primerov.

Pri izbiri števila skupin izračunamo silhueto za različne vrednosti $K$ in izberemo tisto, ki daje največjo vrednost $s$. Uporabimo jo lahko tako pri hierarhičnih metodah kot pri metodi voditeljev. Pristop ne optimizira razbitja neposredno, temveč služi kot zunanja cenilka kakovosti. Njena prednost je intuitivna interpretacija in upoštevanje dveh ključnih vidikov gručenja, slabost pa občutljivost na izbrano mero razdalje ter manjša zanesljivost v primerih, kjer so razlike med možnimi razbitji majhne ali kjer gruče niso jasno ločene.

### *DBSCAN*

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) je metoda gručenja, ki temelji na gostoti podatkov in išče gruče kot območja z visoko gostoto, ločena z območji nizke gostote. Vhod v metodo je množica primerov, mera razdalje ter dva parametra: $\varepsilon$, ki določa radij okolice, in $minPts$, ki določa minimalno število primerov v tej okolici. Algoritem razvrsti primere v tri tipe: jedrne (z dovolj sosedi), robne (blizu jedrnih) in šum (osamelci). Izhod metode je razbitje na gruče in množica točk, označenih kot šum.

Algoritem začne z izbiro neobdelanega primera in preveri njegovo okolico. Če je primer jedrni, iz njega razširi gručo tako, da rekurzivno vključuje vse dosegljive primere v njegovem $\varepsilon$-okolju. Ta

postopek se ponavlja, dokler niso obdelani vsi primeri. DBSCAN ne zahteva vnaprejšnje določitve števila gruč in lahko odkrije gruče poljubnih oblik.

Metoda ne optimizira eksplicitne globalne cenilke, temveč temelji na lokalnem kriteriju gostote. Njena prednost je robustnost na šum in sposobnost odkrivanja kompleksnih struktur, slabosti pa so občutljivost na izbiro parametrov $\varepsilon$ in *minPts* ter slabša učinkovitost pri podatkih z zelo različno gostoto.

### *Gručenje na omrežjih*

Pri gručenju na omrežjih podatke predstavimo kot graf, kjer vozlišča predstavljajo primere, povezave pa podobnosti med njimi. Ena najpreprostejših metod je razširjanje oznak (angl. label propagation), kjer vsakemu vozlišču najprej dodelimo lastno oznako, nato pa v iteracijah vsako vozlišče prevzame najpogostejšo oznako med svojimi sosedi. Postopek ponavljamo do konvergence. Rezultat so skupnosti vozlišč, ki so med seboj gosto povezane. Metoda je zelo hitra in ne zahteva vnaprejšnje določitve števila skupin, vendar je lahko nestabilna in občutljiva na vrstni red posodabljanja.

Bolj robusten in pogosto uporabljen pristop je metoda Louvain, ki temelji na optimizaciji modularnosti, to je mere, ki ocenjuje, kako močno so povezave znotraj skupin gostejše od pričakovanih v naključnem grafu. Algoritem poteka v dveh korakih: najprej lokalno premika vozlišča med skupinami tako, da povečuje modularnost, nato pa zgradi nov, manjši graf, kjer so skupine združene v nadvozlišča, in postopek ponovi. Gručenje po Louvainu je popularno zaradi svoje učinkovitosti na velikih omrežjih, dobre kakovosti odkritih skupnosti, število skupin pa je lahko (močno) odvisno od metaparametrov.

Atributne podatke lahko za tak pristop uporabimo tako, da iz njih najprej zgradimo graf podobnosti, na primer s povezovanjem najbližjih sosedov ali z uporabo praga razdalje. Uteži povezav lahko določimo glede na podobnost med primeri. Nato na tako dobljenem grafu uporabimo metode za gručenje na omrežjih, kot sta razširjanje oznak ali Louvain. Na ta način združimo prednosti atributnega opisa podatkov in strukturne analize omrežij