Klasterska analiza

Sekulić, Sven

SVEUČILIŠTE U RIJECI

TEHNIČKI FAKULTET

Prijediplomski sveučilišni studij elektrotehnike

Završni rad

KLASTERSKA ANALIZA

Rijeka, rujan 2023.

Sven Sekulić
0069087051

SVEUČILIŠTE U RIJECI

TEHNIČKI FAKULTET

Prijediplomski sveučilišni studij elektrotehnike

Završni rad

KLASTERSKA ANALIZA

Mentor: izv. prof. dr. sc. Ivan Dražić

Komentor: doc. dr. sc. Angela Bašić-Šiško

Rijeka, rujan 2023.

Sven Sekulić
0069087051

SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET
POVJERENSTVO ZA ZAVRŠNE ISPITE

Rijeka, 13. ožujka 2023.

Zavod: **Zavod za matematiku, fiziku i strane jezike**
Predmet: **Inženjerska matematika ET**
Grana: **1.01.07 primijenjena matematika i matematičko modeliranje**

ZADATAK ZA ZAVRŠNI RAD

Pristupnik: **Sven Sekulić (0069087051)**
Studij: Sveučilišni prijediplomski studij elektrotehnike

Zadatak: **Klasterska analiza**

Opis zadatka:

U radu je potrebno objasniti pojam klasterske analize i temeljne pojmove vezane uz provođenje klasterske analize. Potrebno je objasniti različite modele klasteriranja kao što su primjerice hijerarhijsko i centroidno klasteriranje, a detaljno opisati nekoliko najčešće korištenih modela.

U praktičnom dijelu rada potrebno je odabrati nekoliko modela klasteriranja i provesti ih na različitim skupovima realnih podataka. Dobivene rezultate potrebno je usporediti i protumačiti. U završnom dijelu rada klastersku analizu potrebno je staviti u kontekst primjene u inženjerstvu i elektrotehnici.

Rad mora biti napisan prema Uputama za pisanje diplomskih / završnih radova koje su objavljene na mrežnim stranicama studija.

Zadatak uručen pristupniku: 20. ožujka 2023.

Mentor:

Izv. prof. dr. sc. Ivan Dražić

Doc. dr. sc. Angela Bašić-Šiško (komentor)

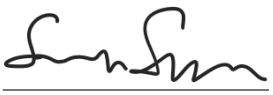
Predsjednik povjerenstva za završni ispit:

Prof. dr. sc. Dubravko Franković

IZJAVA

Sukladno članku 7. stavku 1. Pravilnika o završnom radu, završnom ispitu i završetku sveučilišnih prijediplomskih studija Tehničkog fakulteta Sveučilišta u Rijeci od 4. travnja 2023., izjavljujem da sam samostalno izradio završni rad prema zadatku preuzetom dana 20. ožujka 2023.

Rijeka, 10. rujna 2023.



Sven Sekulić

Zahvaljujem se mentoru izv. prof. dr. sc. Ivanu Dražiću na izrazitoj podršci i pomoći pri pisanju završnog rada. Veliko hvala za utrošeno vrijeme i strpljenje koje je mentor iskazao tijekom kolegija, izrade projekta i pisanja završnog rada.

Također bi se htio zahvaliti profesorima Tehničkog fakulteta u Rijeci na znanju i vještinama koje su mi prenijeli. Isto tako htio bi se zahvaliti kolegama koji su mi bili podrška u najtežim trenucima studija.

Na posljetku, htio bi se zahvaliti roditeljima i obitelji na potpori tijekom studija.

Sadržaj

1. Uvod

Klasterska analiza dio je šireg statističkog područja klasifikacije koji se bavi identifikacijom grupa u određenom skupu podataka na temelju sličnosti atributa ili karakteristika elemenata. Ideju pronalaska uzoraka i sličnosti između podataka prvi istražuje K. Pearson[1] u članku *"On Lines and Planes of Closest Fit to Systems of Points in Space"* iz 1901. godine [1]. Što nam kasnije i postaje podlogom za klastersku analizu.

Sam pojam klasterske analize dolazi iz djela Drivera[2] i Kroebera[3] iz 1932. godine koji ga koriste u sklopu antropologije [2], a kasnije ga Joseph Zublin[4] uvodi u psihologiju 1938. godine [3]. Prvi spomen same riječi klasterska analiza dolazi iz djela R. Tryona[5] iz 1939. godine [4], a S. C. Johnson[6] je prvi koji u jednom djelu opisuje klastersku analizu i njenu metodu 1967. godine [5] dok prvi sveobuhvatni uvod u klastersku analizu nalazimo u djelu P. Rousseeuw-a[7] iz 2009. godine [6].

U današnje vrijeme razvoj metoda grupiranja se više fokusira na razradu algoritma za specifične skupove podataka, te se može reći da je razrada novih algoritama za opće svrhe stagnirala. No, to nikako ne umanjuje značaj grupiranja kao metode koja je sve više zastupljena. Naime, u novije se vrijeme barata sve većom količinom podataka koje je potrebno pravilno procesirati kako bi bili korisni, a grupiranje je u tom kontekstu ključni korak jer nam je u većini slučajeva gdje baratamo sa većom količinom podataka gotovo nemoguće analizirati skup bez neke prethodne analize i pojednostavljenja podataka. To nam je zapravo glavna zadaća klasterske analize, da skup podataka pojednostavi i olakša nam daljnju analizu. Današnja se klasterska analiza zbog često velikog obujma podataka i jednostavnosti gotovo isključivo provodi na računalima uz pomoć programa kao što su $Stata$. Klasterska analiza kao metoda grupiranja sve se više koristi pa je tako susrećemo u područjima poput:

- Medicine - primjerice za analizu PET skeniranja gdje se koristi u svrhu razlikovanja različitih tkiva, za analizu otpora bakterija na antibiotike, za grupiranje dijelova ljudske genetike kako bi se izradila populacijska struktura ili pak za određivanje funkcionalnosti pojedinih gena s obzirom na njihovu sličnost nekim genima čiju funkciju znamo,
- Ekonomije i marketinga - najviše se u ovoj struci koristi za analizu multivarijatnih podataka iz anketa, za grupiranje stanovništva u svrhu boljeg razumijevanja određenih grupa i njihove

[1] Karl Pearson, 1857.-1936., engleski matematičar i biostatičar
[2] Harold Edson Driver, 1907.-1992., američki kulturni antropolog
[3] Alfred Louis Kroeber, 1876.-1960., američki kulturni antropolog
[4] Joseph Zublin, 1900.-1990., američki psiholog
[5] Robert Choate Tryon, 1901.-1967., američki bihevioralni psiholog
[6] Stephen Curtis Johnson, 1944., američki informatičar
[7] Peter J. Russeeuw, 1956., belgijski statičar

međusobne interakcije kao i za efikasnije i bolje marketinške kampanje koje ciljaju točno određenu skupinu ljudi,

- Optimizacije tražilica - za prikazivanje relevantnijih rezultata na stranicama kao što su primjerice *Google* ili *YouTube*,
- Internetske sigurnosti - uočavanje parametara ponašanja koji odskaču od uobičajnih kako bi pravovremeno spriječili zlonamjeran proboj,
- Računalne znanosti - za prepoznavanje objekata na slici primjerice kod umjetne inteligencije ili automonme vožnje.

Ovdje smo naveli samo jedan mali broj primjena klasterske analize. Ona nam u suštini omogućava da bolje iskoristimo raspoložive podatke te da lakše uočimo sličnosti u podacima. Rezultati klasterske analize mogu nam pomoći i kod predikcije budućih vrijednosti nekog promatranog elementa s obzirom na to kojoj bi grupi mogao pripadati te gledajući vrijednost te grupe.

U elektrotehnici klastersku analizu možemo iskoristiti za primjerice dijagnostiku mreže u svrhu njena poboljšanja, odnosno otklanjanja problema. Najčešće to radimo na način da analiziramo električne signale kao što su napon, struja i snaga te kako se ponašaju u normalnim uvjetima za specifičan način rada. Ukoliko postoji problem, element unutar skupa ima devijaciju u odnosu na svoju predviđenu grupu i time nam pokazuje da postoji problem u mreži. Koristimo ga također i za predviđanje tereta na mrežu, na primjer ako nam grupe predstavljaju teret mreže na određeni dan ili za određene klijente možemo se adekvatno unaprijed pripremiti te tako maksimalno optimizirati opskrbu mreže električnom energijom odnosno opskrbljivati mrežu s točno onoliko energije koliko smo predvidjeli da bi nam trebalo. Kao i kod prepoznavanja objekta na slikama u računarstvu, klastersku analizu možemo koristiti i u elektrotehnici za procesiranje termalnih slika i tako odrediti zagrijavanje pojedinih elemenata u mreži te odrediti postoje li kakve devijacije u odnosu na uobičajan rad [7].

U ovom radu bavit ćemo se najčešćim modelima i algoritmima klasterske analize te njihovom primjenom. Prvo ćemo objasniti teorijski dio i metodologiju klasterske analize, a zatim ćemo klastersku analizu primijeniti na stvarnim podacima.

Pokazati ćemo kako se klasterska analiza može koristiti za grupiranje studenata s obzirom na njihov pristup učenju tijekom semestra. To nam može pomoći kako bi otkrili neke uzorke ponašanja u svrhu individualizacije pristupa određenoj grupi studenata.

Također ćemo provesti klastersku analizu i kroz realni primjer u elektrotehnici na način da ćemo analizirati potrošnju električne energije te njenu efikasnost za proizvodnju jednog dobra kao i udio obnovljivih izvora energije u ukupnoj potrošnji da vidimo postoje li neke korelacije između efikasnosti i korištenja obnovljivih izvora te koje države imaju najamnji utjecaj na zagađenje.

2. Definicija klasterske analize

Klasterska analiza je tehnika statističke analize čiji je zadatak grupiranje određenog skupa podataka na način da su elementi iste grupe slični, a elementi različitih grupe različiti. S obzirom na raspodjelu elemenata grupiranje možemo podijeliti na:

- Čvrsto grupiranje - gdje svaki element pripada isključivo jednoj grupi,
- Meko grupiranje - gdje jedan element može pripadati u više grupa istovremeno.

Razlikujemo puno različitih algoritama i pristupa klasterskoj analizi s obzirom na to s kakvim skupom podataka radimo. Neke najpoznatije ćemo detaljnije obraditi u sklopu ovog rada.

U sljedećim potpoglavljima proći ćemo osnovne pojmove vezane za klastersku analizu kako bi je bolje razumjeli i kako bi razumjeli od čega se ona sastoji. Nenadzirano učenje nam je bitno za proći kako bi shvatili klastersku analizu jer klasterska analiza upravo i spada u algoritam nenadziranog učenja, kod kojeg na početku ne znamo rezultate te to zapravo predstavlja najveći izazov grupiranja.

Ovo poglavlje obrađeno je prema [11], [12], [13], [14], [15] i [16].

2.1. Nenadzirano i nadzirano učenje

Nenadzirano učenje je vrsta strojnog učenja kod kojeg na početku ne znamo rezultate koje bi trebali dobiti. S druge strane, nadzirano učenje je strojno učenje kod kojeg znamo koje rješenje trebamo dobiti te nam zapravo najviše služi za primjerice učenje agenta[1] u sklopu umjetne inteligencije kako bi obavio zadatak koji smo mu zadali. Klasterska analiza spada u nenadzirano učenje, pa ćemo se zbog toga na tu vrstu učenja i fokusirati. Moramo grupirati podatke u grupe a da ne znamo kako bi optimalno grupe trebale izgledati. No, prvo moramo vidjeti kako ćemo skup i njegove elemente označavati.

2.1.1. Matematički zapis problema grupiranja

Skup n elemenata x svrstan u m grupa y možemo zapisati kao:

$$D = \{(x^{(i)}, y^{(j)})\}_{i=1}^{n}, 1 < j < m \tag{2.1}$$

gdje je D skup elemenata, a x i-ti element j-te grupe y. Navesti ćemo kratki primjer da formula bude jasnija.

[1] Agent u umjetnoj inteligencijij predstavlja program ili algoritam koji percipira svoje okruženje i donosi odluke s obzirom na to koji cilj nastoji ostvariti

Primjer 2.1. *Ako imamo tri elementa koja trebamo smjestiti u dvije grupe na način da dva elementa idu u prvu grupu, a jedan element u drugu grupu, onda skup D možemo zapisati kao:*

$$D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(1)}), (x^{(3)}, y^{(2)})\}. \tag{2.2}$$

Dakle, skup D sastoji se od dvije grupe y i tri elementa x na način da su dva elementa u prvoj grupi, a jedan element u drugoj grupi.

Međutim, kako se radi o nenadziranom učenju, mi ne znamo koji element skupa x pripada kojoj grupi y te takav skup negrupiranih primjera možemo zapisati kao:

$$D = \{x^{(i)}\}_{i=1}^{n}. \tag{2.3}$$

Sad dolazimo do problema. Kako ćemo znati da je rezultat koji smo dobili optimalan ako nemamo označenih primjera? Drugim riječima kako ćemo znati da je grupacija koju smo dobili algoritmom optimalna i je li uopće ispravna? Također se otvara pitanje koliko je dobro algoritam odredio sličnosti (i razlike) elemenata?

Jedna od metoda za osiguranje kvalitete algoritma je da algoritam primjenimo s unaprijed valoriziranim primjerima te da ga usavršavamo sve dok ne dobijemo rezultate koji želimo, odnosno rezultate koje smo zadali.

Primjerice, recimo da želimo razviti algoritam za grupiranje studenata s obzirom na njihovu završnu ocjenu analizirajući njihov pristup kolegiju tijekom semestra. Možemo uzeti prošlogodišnje rezultate na kojima već imamo završnu ocjenu te ubaciti u algoritam sve osim završne ocjene. Nakon što dobimo rješenje možemo onda prilagoditi algoritam na način da korektno grupira studente prema njihovim ocjenama. Kada smo prilagodili algoritam na ispravno grupira studente prema ocjenama za dodatnu sigurnost možemo algoritam provesti i na rezultatima od pretprošle godine te opet vidjeti hoće li algoritam dobro grupirati. Ako nije, prilagodimo ga još više, a ako je onda znamo da će nam za iste podatke koje smo mu ubacili na prve dvije analize točno grupirati studente prema završnim ocjenama.

Grupiranje se također može koristiti i u nadziranom učenju, no više u svrhu smanjenja dimenzionalnosti podataka čime efektivno pojednostavljujemo primjere. Drugim riječima, klasterska analiza nam može pomoći pojednostaviti skup i bolje razdvojiti elemente tako da nam je na kraju jednostavnije analizirati rezultate. To radi na način da efektivno izbacuje elemente iz grupe koji su možda više različiti od ostatka i tako nam smanjuje različitosti elemenata u grupi i time olakšava analizu.

Grupiranje ne predstavlja jedan jedinstveni algoritam, već problem koji je potrebno riješiti. Unutar grupiranja postoji više tipova algoritama za svrstavanje podataka te odabir specifičnog algoritma ovisi o vrsti podataka s kojim radimo.

Kako smo prije naveli, grupiranje spada u nenadzirano učenje te čak algoritmi specifično prilagođenih za naše potrebe mogu na istom skupu podataka davati različite rezultate. Ovo se najčešće

događa jer je redoslijed uzimanja podataka najčešće nasumičan. Kako bi znali koji element ćemo pridodati kojoj grupi, moramo imati neku mjeru udaljenosti koja bi u našem slučaju predstavljala međusobnu sličnost elemenata te ćemo zbog toga u sljedećem potpoglavlju navesti neke osnovne metrike udaljenosti koje imamo u klasterskoj analizi te kako se one koriste.

2.2. Metrike u statističkoj analizi

Metrika je funkcija koja definira kolika je udaljenost dvaju međusobnih elemenata u nekom skupu podataka. U statističkoj analizi koristi se za definiranje koliko su dva elemenata međusobno slična prema tome kolika im je međusobna udaljenost pa nam je kao takva vrlo korisna za naš zadatak grupiranja podataka prema sličnosti. Također se često primjenjuje i za izračun pogreške ili nagrade[2] u strojnom učenju. Elementi čiju udaljenost računamo ne moraju nužno biti brojevi već to mogu biti i vektori, matrice ili arbitrarni objekti. Postoje dvije skupine metrika koje najčešće koristimo, a to su:

- Diskretne metrike - udaljenosti između dva elementa mogu poprimiti samo određene vrijednosti,
- Neprekidne metrike - udaljenosti mogu poprimiti bilo koju vrijednost.

Izbor odgovarajuće metrike izuzetno je bitan, a za to je potrebno poznavati kako se koja metrika primjenjuje. Naime, u slučaju lošeg odabira metrike nećemo dobiti rezultate koje želimo odnosno koji su optimalni.

U sljedećih nekoliko poglavlja dajemo pregled najčešće korištenih metrika.

2.2.1. Euklidska metrika

Euklidska[3] metrika najčešća je mjera udaljenosti (ujedno i najjednostavnija) te predstavlja klasičnu geometrijsku duljinu segmenta koji spaja dva elementa, a uglavnom se koristi kad imamo podatke niskih dimenzija.

Definicija 2.1. *Neka su zadana dva vektora* $x = (x_1, x_2, ..., x_n)$ *i* $y = (y_1, y_2, ..., y_n)$. *Njihova Euklidska udaljenost računa se formulom*

$$D(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \tag{2.4}$$

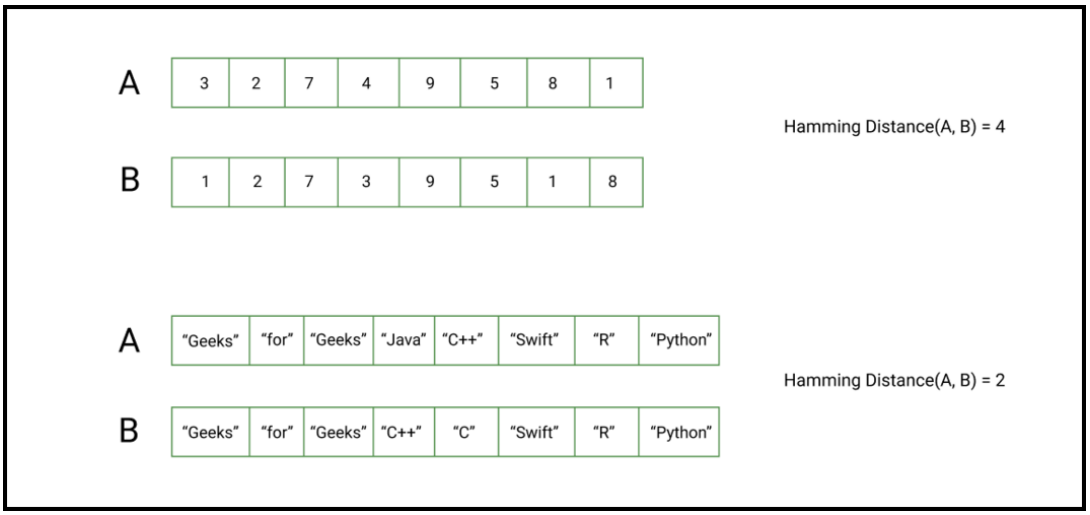
[2]Nagrada u kontekstu strojnog učenja predstavlja numeričku vrijednost koja je posljedica neke akcije koju je agent izveo, te zapravo predstavlja vrijednost koju agent unutar strojnog učenja pokušava maksimizirati.

[3]Euklid, starogrčki matematičar

Problematika korištenja Euklidske metrike očituje se kod višedimenzionalnih prostora kod kojih više nije u potpunosti jasna njena geometrijska interpretacija, a često je prije njenog korištenja neophodna i normalizacija podataka[4], što dodatno otežava izračun [23].

2.2.2. Hammingova metrika

Hammingova[5] metrika predstavlja broj različitih elemenata između dva vektora (najčešće se radi o bitovima unutar neke binarne riječi) te se najčešće koristi za otkrivanje i ispravljanje grešaka unutar binarnih kodova koji se koriste u računalima, a znamo je često susretati i u elektrotehnici u području digitalne elektronike. No, Hammingova metrika može služiti za usporedbu bilo koja dva skupa te nije ograničena samo na bitove. Česta je i kod ugradbenih sustava gdje se koristi upravo za otkrivanje i ispravljanje pogrešaka prilikom prijenosa podataka.



Slika 2.1. Primjer prikaza Hammingove metrike prilikom usporedbe dvaju skupova [8]

2.2.3. Manhattan metrika

Manhattan metrika dobila je ime prema središnjem dijelu New Yorka koji se naziva Manhattan. Urbanistička karakteristika Manhattana je matematički precizan raspored ulica u vidu pravilne pravokutne mreže. Drugim riječima, da bi se došlo od točke A do točke B potrebno je hodati po pravokutnoj mreži i udaljenost od točke A do točke B može se definirati brojem vodoravnih i okomitih segmenata koje je na tom putu potrebno prijeći.

Formalno se Manhattan metrika može definirati na sljedeći način.

[4]Normalnizacija podataka je obrada podataka na način da su svi dostupni podaci prikazani kroz isti standardizirani format, to nam je nužno za analiziranje međusobnih korelacija između podataka te za analizu podataka.

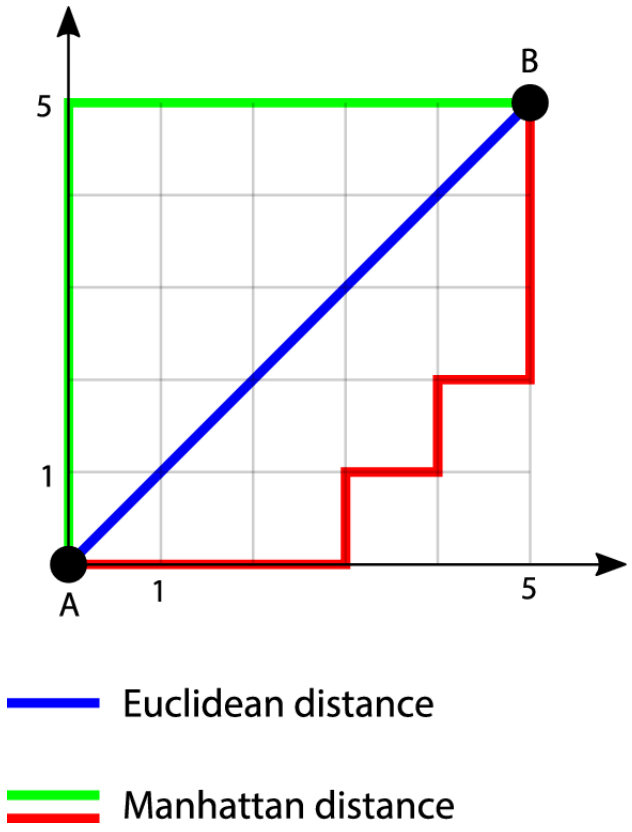
[5]Richard Wesley Hamming, 1915.-1998., američki matematičar

Definicija 2.2. *Neka su zadana dva vektora* $x = (x_1, x_2, ..., x_n)$ *i* $y = (y_1, y_2, ..., y_n)$. *Njihova udaljenost u Manhattan metrici računa se formulom*

$$D(x, y) = \sum_{i=1}^{n} |x_i - y_i|. \tag{2.5}$$

Ako nam vektori x i y predstavljaju dvodimenzionalne točke, jasno je da smo ovim izrazom dobili netom opisanu udaljenost.

Manhattan metrika vrlo često se koristi kod diskretnih i binarnih atributa, kao i kod podataka visokih dimenzija. Svakako treba naglasiti da se često događa da je udaljenost dobivena Manhattan metrikom veća od one dobivene Euklidskom metrikom, što je i vidljivo na slici *2.2.*.



Slika 2.2. Usporedba Euklidske i Manhattan udaljenosti [9]

Kako se moramo kretati po rubovima zamišljenih kvadrata, tako Manhattan udaljenost zna biti dulja od Euklidske, ali joj upravo ta robusnost daje prednost pri analizi podataka viših dimenzija.

2.2.4. Jaccardova metrika

Jaccardova[6] metrika često se koristi kod diskretnih podataka koji se ne mogu opisati uređenim n-torkama, već skupovima. Za izračun Jaccardove metrike ključan je Jaccardov indeks koji definiramo na sljedeći način.

[6]Paul Jaccard, 1868.-1994., švicarski botaničar

Definicija 2.3. *Neka su zadana dva skupa A i B. Njihov Jaccardov indeks izračunava se formulom*

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}, \tag{2.6}$$

gdje $\cap$ označava presjek, a $\cup$ uniju skupova, dok je $|\cdot|$ kardinalni broj, odnosno broj elemenata u skupu.

Jaccardov indeks u biti je svojevrsni omjer presjeka i unije, a po svojoj definiciji poprima vrijednosti između 0 i 1.

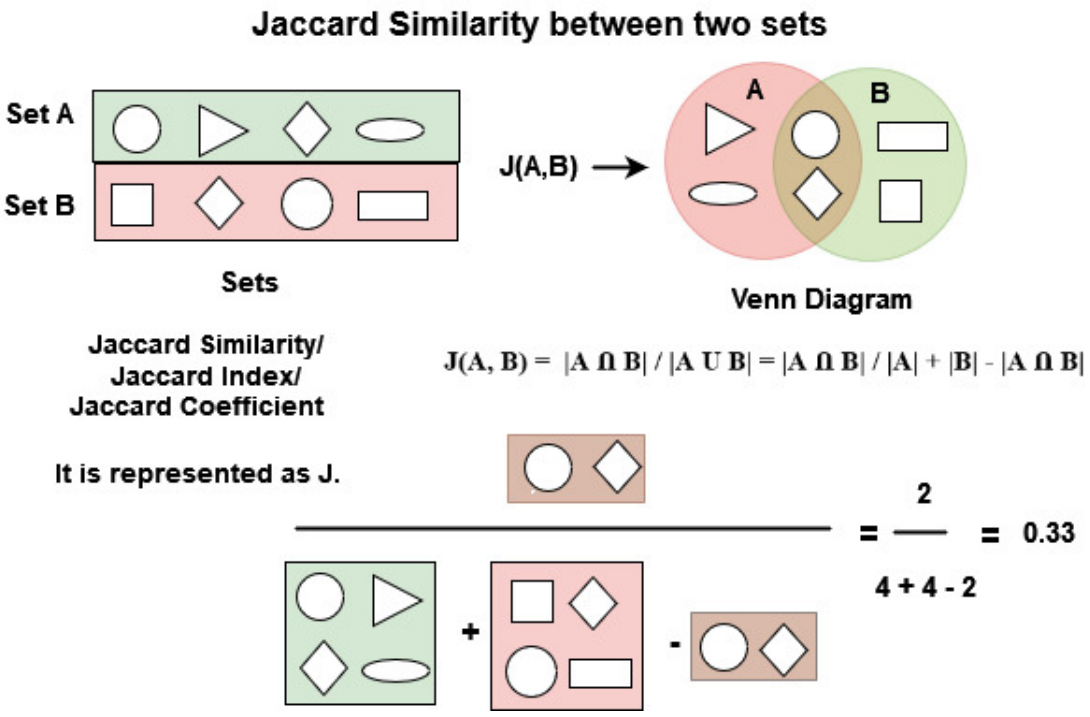
Sada možemo definirati i Jaccardovu metriku.

Definicija 2.4. *Neka su zadana dva skupa A i B. Jaccardova metrika kojom se definira udaljenost ta dva skupa dana je formulom*

$$D(A,B) = 1 - J(A,B), \tag{2.7}$$

gdje $J(A,B)$ označava upravo definirani Jaccardov indeks.

Primjer izračuna Jaccardovog indeksa prikazan je na slici *2.3.*.



Slika 2.3. Primjer izračuna Jaccardovog indeksa [10]

Jaccardovu metriku možemo naći u analizi tekstova gdje gledamo koliko je jedan tekst sličan drugom ili kod procesiranja slike u sklopu dubokog učenja.

Najveći nedostatak Jaccardovog indeksa je što veličina skupa utječe na rezultat jer se dodavanjem elemenata u skup unija povećava, a presjek ostaje isti ili vrlo sličan.

2.2.5. Kosinusna metrika

Kosinusna metrika predstavlja razliku u kutu između dvaju elemenata reprezentiranih vektorima. Geometrijski je jasno da dva elementa istih orijentacija imaju sličnost 1, dok dva elementa suprotnih vektora imaju sličnost -1. Valja napomenuti da nam duljina vektora ovdje nije bitna i to je zapravo glavna karakteristika kosinusne metrike. Kosinusna metrika često se koristi kod analiza tekstova i skupova visoke dimenzije. Kosinusna metrika dana je sljedećom definicijom.

Definicija 2.5. *Neka su zadana dva vektora x i y. Njihova kosinusna metrika dana je formulom*

$$D(x,y) = \cos(\theta) = \frac{x \cdot y}{\|x\|\|y\|}, \tag{2.8}$$

gdje $\|x\|$ označava normu vektora, a točkasti operator je skalarni produkt.

Iz same definicije kosinusne metrike je jasno da za njen izračun moramo poznavati odgovarajući skalarni produkt.

2.2.6. Chebyshevljeva metrika

Chebyshevljeva[7] metrika predstavlja maksimalnu udaljenost između dvije komponente vektora u n dimenzionalnom prostoru. Najčešće se koristi u digitalnom procesiranju signala. Chebyshevljeva metrika dana je u sljedećoj definiciji.

Definicija 2.6. *Neka su zadana dva vektora $x = (x_1, x_2, ..., x_n)$ i $y = (y_1, y_2, ..., y_n)$. Njihova Chebyshevljeva metrika računa se formulom:*

$$D(x,y) = \max_i(|x_i - y_i|), \tag{2.9}$$

dok se za dvije točke $A = (x_a, y_a)$ i $A = (x_b, y_b)$ njihova Chebyshevljeva metrika računa formulom:

$$dist(A,B) = \max(|x_a - x_b|, |y_a - y_b|). \tag{2.10}$$

Pogledat ćemo sad jedan primjer računanja Chebyshevljeve metrike.

Primjer 2.2. *Za slučaj da su nam zadane točke $A = (70, 40)$ i $B = (330, 220)$, trebamo izračunati Chebyshevljevu udaljenost. Računamo je pomoću formule za računanje udaljenosti dvaju točaka:*

$$\begin{aligned} dist(A,B) &= \max(|x_a - x_b|, |y_a - y_b|), \\ dist(A,B) &= \max(|70 - 330|, |40 - 220|), \\ dist(A,B) &= \max(|-260|, |-180|), \\ dist(A,B) &= \max(260, 180), \\ dist(A,B) &= 260. \end{aligned} \tag{2.11}$$

[7] Pafnuty Lvovich Chebyshev, 1821.-1894., ruski matematičar

2.2.7. Metrika Minkowskog

Metrika Minkowskog[8] je metrika koja u sebi sadrži ostale već ranije spomenute metrike te ona zapravo predstavlja generalizaciju tih metrika, a definira s na sljedeći način.

Definicija 2.7. *Neka su zadana dva vektora* $x = (x_1, x_2, ..., x_n)$ *i* $y = (y_1, y_2, ..., y_n)$. *Njihova metrika Minkowskog metrika računa se formulom*

$$D(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}, \tag{2.12}$$

pri čemu je $p > 0$ *slobodni realni parametar.*

Najčešće vrijednosti parametra p su:

- $p = 1$ - u tom slučaju metrika Minkowskog ponaša se kao Manhattan metrika,
- $p = 2$ - u tom slučaju metrika Minkowskog ponaša se kao Euklidska metrika,
- $p = \infty$ - u tom slučaju metrika Minkowskog ponaša se kao Chebyshevljeva metrika.

Svaka metrika, odnosno udaljenost treba ispunjavati sljedeća svojstva:

1. udaljenost elementa od samog sebe jednaka je nuli,
2. udaljenost svaka dva različita elementa je pozitivan broj,
3. udaljenost je simetrična, odnosno udaljenost od elementa A do elementa B jednaka je udaljenosti od elementa B do elementa A,
4. mora vrijediti nejednakost trokuta, odnosno za bilo koja tri elementa A, B i C vrijedi

$$d(A, C) \leq d(A, B) + d(B, C), \tag{2.13}$$

pri čemu $d(\cdot, \cdot)$ označava analiziranu metriku.

[8] Hermann Minkowski, 1864.-1909., njemački matematičar i fizičar

3. Modeli grupiranja

S obzirom na to da konstrukcija grupe može znatno ovisiti o tome s kakvim podacima radimo, postoje različite metodologije, odnosno algoritmi grupiranja. Najčešći modeli grupiranja su:

- Centroidni modeli - elementi se grupiraju s obzirom na središnji element odnosno centroid[1] (težište), a najpoznatiji centroidni model je particijsko grupiranje
- Konekcijski modeli - grupiranje elemenata se bazira na međusobnoj udaljenosti dvaju susjednih elemenata (ili grupa), a najpoznatiji konekcijski model je hijerarhijsko grupiranje
- Modeli temeljeni na gustoći - grupe se definiraju kroz gustoću distribucije elemenata na nekom području.

Valja napomenuti da postoji još mnogo modela, no većina ih se ne razlikuje puno od ovih navedenih te zapravo svi ti modeli rade na sličnom principu [13].

U nastavku poglavlja ćemo detaljnije opisati spomenute modele.

3.1. Particijsko grupiranje

Particijsko grupiranje pripada centroidnim modelima i metoda je vektorske kvantizacije kojoj je cilj veliku skupinu primjera (vektora) podijeliti u grupe na način da je kvadratna vrijednost udaljenosti svake točke određene grupe minimalna. Drugim riječima, elemente skupa pridodajemo najbližem mu centroidu koji je na početku nasumično odabran. Nakon toga premjestimo centroid na način da smanjimo varijaciju elemenata koji su mu pridodani te ga stavljamo u središte njegove grupe. Pošto se centroidi pomiču, tako se dosta često dogodi da elementi skupa promijene grupe te da im centroid susjedne grupe postane bliži od prvobitnog pa se time onda ti elementi pridodaju upravo toj novoj grupi čiji im je centroid sada bliži. Valja napomenuti da je ishod particijskog grupiranja sklon osciliranju, odnosno ne mora nužno svaki prolaz kroz algoritam dovesti do istih rezultata, što ćemo također vidjeti u sljedećem dijelu [11].

Ako neki set elemenata x želimo smjestiti u k grupa na način da minimiziramo varijaciju elemenata unutar grupe morat ćemo za svaki element računati njegovu (najčešće Euklidsku) udaljenost od centroida određene grupe te ga pridodati grupi čiji je centroid najbliži tome elementu. Moramo imati na umu da nam broj grupa ne smije biti veći od broja elemenata. Broj grupa će ovisiti koilko će rješenje biti precizno odnosno jednostavno te će veći broj grupa davati veću preciznost, no biti će kompliciraniji i teži za opisati i analizirati, a manji broj grupa veću jednostavnost

[1] Centroid (težište) predstavlja točku u kojoj je zbroj svih vektora koji počinju u njoj, a završavaju u točkama danoga skupa jednak nuli.

i zaglađenost, no postoji šansa da će nam smanjiti preciznost i iskriviti vrijednosti nekih grupa. O problemu i rješenju pronalaska optimalnog broja grupa ćemo raspravljati kasnije [12].

3.1.1. Algoritam particijskog grupiranja

Najčešći algoritam za određivanje particijskog grupiranja je Lloyd-ov algoritam koji se još naziva i "običan *k-means* algoritam" jer su se u međuvremenu razvile i puno brže metode. Odvija se u dva jedostavna koraka koji se ponavljaju sve dok nove iteracije ne daju drukčije rješenje: dodjela i ažuriranje.

No, prije nego što uđemo u petlju ova dva koraka, na početku je potrebno odrediti broj grupa k, te nasumično odrediti početni položaj centroida grupa m_i^0. Nakon što smo to dvoje odredili, prelazimo na dva glavna koraka.

U prvom koraku se svakom elementu dodjeljuje grupa čiji je centroid najbliži (čija je udaljenost najmanja vrijednost kvadrata Euklidske udaljenosti) u usporedbi sa centroidima drugih grupa, te se onda elementi dijele prema Voronoijevom dijagramu[2]. Za pridodavanje elementa x_p u k broj grupa u t iteraciji vrijedi formula:

$$S_i^{(t)} = \left\{ x_p \colon \| \, x_p - m_i^{(t)} \| \, ^2 \leq \| \, x_p - m_j^{(t)} \| \, ^2 \right\} \tag{3.1}$$

gdje $m_i^{(t)}$ predstavlja centroid i-te grupe u određenoj iteraciji t ($m_j^{(t)}$ predstavlja centroide ostalih j grupa), a x_p predstavlja element koja je pridružen točno jednoj grupi $S^{(t)}$. Dakle, naš element x_p pripast će onoj grupi čiji centroid m_i ima manju Euklidsku udaljenost u usporedbi s Euklidskim udaljenostima ostalih centroida grupa m_j.

U drugom koraku ažuriramo položaj centroida na način da usrednjimo vekotre svih elemenata grupe, a računamo ga prema formuli:

$$m_i^{(t+1)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_p \in S_i^{(t)}} x_p, \tag{3.2}$$

gdje $\left| S_i^{(t)} \right|$ predstavlja duljinu vektora i-te grupe.

Dakle, vektore svih elemenata grupe zbrajamo te dijelimo sa duljinom vektora grupe kako bi dobili središte te grupe gdje onda smještamo centroid u novoj iteraciji. Kao što smo prije rekli, premještanjem centroida moguće je da se promijene same grupe to jest moguće je da se za neki element njegov prvobitni centroid udalji, a centroid susjedne grupe približi i postane mu bliži od prvobitnog. Radi toga, ova dva koraka ponavljamo sve dok ne dođe do stagnacije promjene elemenata grupe odnosno sve dok nove iteracije ne daju promjene grupa ili daju vrlo male promijene [11].

[2] Voronoijev dijagram je podjela ravnine na područja bliska svakom od elemenata skupa, to jest podjela ravnine s obzirom na to koji centroid je najbliži na kojem dijelu ravnine

3.1.2. Prednosti i mane particijskog grupiranja

Iako je particijsko grupiranje vrlo koristan i relativno brz algoritam, u nekim je sutiacjiama vrlo nepovoljan. Primjerice: kod particijskog grupiranja pretpostavljamo da su grupe približno jednakih veličina, varijacija elemenata se koristi za mjerenje rasprostranjenosti grupe što je krajnje nepovoljno kod usko raspodijeljenih elemenata gdje bi bolja grupacija bila grupacija temeljem gustoće. Također, potrebno je odrediti broj grupa na početku provođenja algoritma te će lošim odabirom broja grupa rezultat biti izrazito ne optimalan. No, čak i kad odredimo dobar broj grupa, algoritam je potrebno više puta provesti jer se zbog nasumičnog odabira početnih centroida grupa rezultati mogu razlikovati.

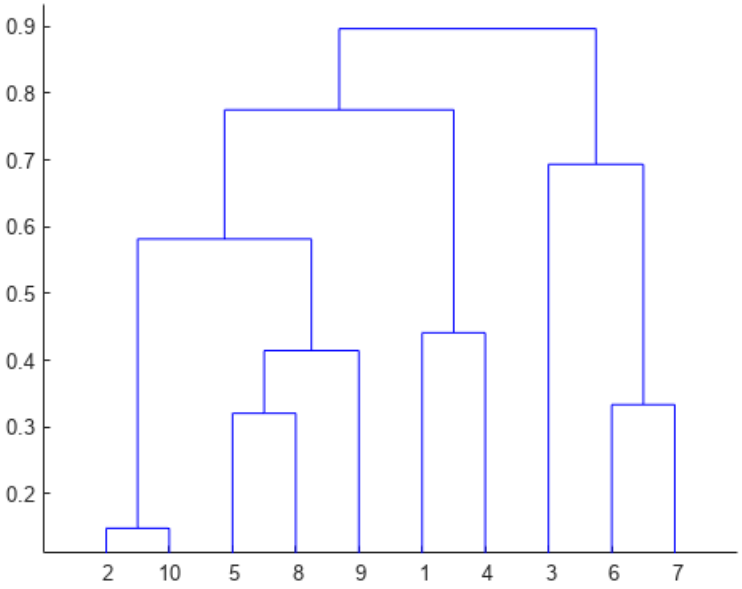
3.2. Hijerarhijsko grupiranje

Hijerarhijski algoritam grupiranja jedan je od jednostavnijih algoritama koji se bazira na izgradnji hiijerarhija između pojedinih elemenata, te se vrlo često prikazuje dendrogramom[3] gdje su najsličniji elementi prikazani na dnu dendrograma, a kako se pomičemo gore, raste različitost. Postoje dvije kategorije hijerarhijskog grupiranja:

- Aglomerativno - u kojem elementi započinju kao zasebne grupe koje se spajaju kako se krećemo gore po hijerarhiji
- Divizivno - kod kojeg krećemo sa svim elementima u jednoj grupi te ih onda razdvajamo kako se spuštamo po hijerarhiji

Standardna kategorija je hijerarhijsko aglomerativno grupiranje (HAC), te ćemo njega u nastavku opisivati. HAC se najbolje prikazuje dendrogramom. Na najnižoj razini se nalaze pojedini elementi koji predstavljaju grupu samu za sebe sa samo jednim elementom, koji se onda penjanjem po hijerarhijskim razinama prvo spajaju u grupe sa najsličnijim im susjednim elementom, zatim se ta grupa tretira kao jedan element te se spaja sa najsličnijom im grupom (ili elementom) dok ne dođemo do najviše razine gdje zapravo prikazujemo sve elemente kao jednu grupu. Možemo onda reći da visina hijerarhije predstavlja koliko su elementi međusobno slični, to jest koliko su daleko grupe spojene u nadgrupi. Ako su grane dva međusobna elementa vrlo dugačke, radi se o dva dosta različita elementa dok se kod kratkih grana elemenata radi o sličnim elementima (ili grupama). Valja napomenuti da je hijerarhijsko grupiranje velikog skupa podataka zahtjevno i sporo. [22].

[3]Dendrogram je dijagram prikazan u obliku nalik stablu gdje spojevi između elemenata (i grupa) predstavljaju novu grupu odnosno cjelinu.



Slika 3.1. Primjer dendrograma [28]

3.2.1. Povezivanje grupa

Ukoliko bi htjeli povezati grupe, moramo naći koje su međusobno najsličnije ili najbliže. Ako se u grupama nalazi samo jedan element, njihove se sličnosti ili udaljenosti vrlo lako mogu odrediti. No, ako se nalazimo više na hijerarhijskoj ljestvici, onda vrlo vjerojatno postoji više članova u jednoj grupi . Tada njihove udaljenosti ili sličnosti možemo odrediti na jedan od sljedećih načina:

- Jednostruka povezanost - gledamo najmanju udaljenost između pojedinih elemenata grupa
- Potpuna povezanost - gledamo najveću udaljenost između pojedinih elemenata grupa. Takva udaljenost naziva se potpunim jer ukoliko su najudaljeniji elementi grupa međusobno najbliži u usporedbi s elementima drugih grupa, možemo biti sigurni da će svi ostali elementi definitivno biti još bliži od tih koje smo gledali
- Prosječna povezanost - gledamo prosječnu vrijednost svih elemenata pojedinih grupa
- Centroidna povezanost - gledamo udaljenost između centroida pojedinih grupa.

Formula za jednostruko povezivanje dviju grupa, A i B definira se kao:

$$\min\{d(a,b)\colon a \in A, b \in B\}, \tag{3.3}$$

gdje $d(a,b)$ označava udaljenost elementa x sadržanog u grupi A i elementa b sadržanog u grupi B. Dakle, gledamo minimalnu vrijednost udaljenosti između elemenata dviju grupa.

Nadalje, potpuna povezanost može se onda formalno izraziti kao:

$$\max\{d(a,b)\colon a \in A, b \in B\}. \tag{3.4}$$

Dakle, gledamo maksimalnu vrijednost udaljenosti između dvaju grupa. Prosječna povezanost može se formalno izraziti kao:

$$d_{avg}(A,B) = \frac{1}{|A| \cdot |B|} \sum_{a \in A} \sum_{b \in B} d(a,b), \tag{3.5}$$

a računanje centroida, to jest centroidna povezanost može se izraziti kao:

$$d_{cent}(A,B) = \left\| \frac{1}{|A|} \sum_{c_a \in A} c_a - \frac{1}{|B|} \sum_{c_b \in B} c_b \right\| \tag{3.6}$$

gdje c_a i c_b predstavljaju centroidne (težišne) točke grupa A tj. B [23].

3.2.2. Wardova metoda

Wardova[4] metoda kriteriji je aglomerativnomg hijerarhijskog grupiranja čiji je cilj da se svakim novim korakom hijerarhijskog grupiranja grupiraju elementi (ili grupe) koje međusobnim pridodavanjem najmanje povećavaju varijaciju unutar novonastale grupe. Bazira se, dakle, na optimalnoj vrijednosti funkcije cilja gdje je funkcija cilja pogreška zbroja kvadrata Euklidske udaljenosti dvaju grupa.

Wardov kriteriji minimalne varijacije opisuje nam na koji se način grupiraju određeni elementi (ili grupe). Svaka grupa svodi se na jedinični element koji predstavlja centar te grupe te se dvije grupe međusobno grupiraju na način da je Euklidska vrijednosti što manja odnosno da se elementi buduće grupe razlikuju što manje. Dakle, Wardov kriteriji minimalne varijacije definira se prema kvadratu (najčešće Euklidske) udaljenosti i izražava se formulom:

$$d_{ij} = d(\{X_i\}, \{X_j\}) = ||X_i - X_j||^2. \tag{3.7}$$

Osim pogreške zbroja kvadrata, funkcija cilja za određivanje nove grupe može biti bilo koja funkcija koja ispunjava zadatke promatrača, no ipak u praksi se najčešće koristi upravo funkcija pogreške zbroja kvadrata radi svoje jednostavnosti [26].

3.3. Grupiranje temeljem gustoće

Grupiranje temeljem gustoće radi na principu međusobne udaljenosti elemenata od najbližih im susjednih elemenata, a elementi koji su relativno daleko udaljeni tretiraju se kao šum. Kao i kod particijskog grupiranja, ima puno metoda kojima je podloga grupiranje temeljem gustoće, no najpopularnija metoda je takozvani *DBSCAN* i njega ćemo detaljnije opisati u nastavku.

[4] Joe H. Ward, Jr., 1926.-2011., američki matematičar

3.3.1. Algoritam grupiranja temeljem gustoće

Za grupiranje temeljem gustoće potrebna su nam dva parametra: ϵ koji predstavlja radijus skupine elemenata i minimalni broj elemenata koji se mora nalaziti unutar radijusa da bi taj neki element klasificirali kao dio grupe ($minPts$). U prvom koraku započinjemo gledajući nasumični element. Nakon toga gledamo koliko se elemenata nalazi unutar radijusa odabrane točke te, ako je više od minimalne vrijednosti, taj element postaje jezgreni element i s tim elementom započinjemo skupinu to jest taj element postaje dio novonastale grupe u kojoj se nalazi samo on, a ukoliko se unutar radijusa ne nalazi dovoljno drugih elemenata, tada taj element tretiramo kao šum i odbacujemo ga od mogućnosti da bude unutar neke grupe. Ukoliko je došlo do pojave grupe, svaki njen element koji još nije posjećen se na isti se način ispituje te se pretvara u jezgreni element (dio grupe) ako je njegovim radijusom obuhvaćen minimalni broj drugih elemenata, a u protivnom se pretvara u granični element (isto dio grupe ali na njenoj granici). Ukoliko postoji još elemenata koji nisu istraženi, a koji nisu dio oktrivene grupe, nasumično se odabere jedan od njih i proces se nastavlja dalje sve dok sve točke nisu istražene.

3.3.2. Prednosti i mane grupiranja temeljem gustoće

Grupiranje temeljem gustoće ima razne prednosti u usporedbi s particijskim grupiranjem. Naime, nije mu potrebno unaprijed odrediti broj grupa već to radi automatski, oblik grupe nije toliko utjecajan na rezultate koliko je kod particijskog grupiranja što nam je kod nekih skupova jako bitno, prepoznaje šum i robustan je na granice te su potrebna samo dva parametra.

No, ipak ima neke nedostatke koji ga u određenim situacijama čine beskorisnim. Grupiranje određeno gustoćom nije potpuno deterministčko, što znači da jedna točka može završiti u raznim grupama ovisno o redoslijedu obrade skupa podataka. Isto kao i particijsko grupiranje, udaljenost se računa Euklidskom udaljenosti što kod većeg broja dimenzija stvara probleme u računanju. I na kraju, nije pogodan za obradu podataka čiji je smještaj elemenata različitih gustoća pa ih kao takve može pogrešno označiti [17].

3.4. Probabilističko grupiranje

Probabilističko grupiranje vrlo je slično po provedbi particijskom grupiranje, no spada u meko grupiranje gdje se isti element može pridodati u više od jedne grupe te se kod njega ne definira binarna vrijedost, već vjerojatnost da se određeni primjer nalazi u određenoj grupi. Bitno je naglasiti da je probabilističko grupiranje podosta teže za računati nego particijsko te ono zapravo predstavlja samo generalizaciju particijskog grupiranja. Možemo reći da je particijsko grupiranje zapravo isto kao i probabilističko, samo što smo svakom elementu pridodali vjerojatnost od 0% ili 100% da pripada nekoj grupi [24] [25].

4. Evaluacija grupiranja

Ocjena kvalitete klasterske analize izrazito je bitan korak kako bi vidjeli koliko smo optimalno uspjeli podijeliti skup podataka. Međutim, koliko je bitan korak, toliko je i težak jer nema optimalnog načina da se utvrdi kvaliteta grupiranja. Ipak, neke od češćih metoda ocjene kvalitete su: interna evaluacija, eksterna evaluacija, manualna evaluacija i indirektna evaluacija.

Ovo poglavlje napisano je prema [11].

4.1. Interna evaluacija

Interna evaluacija odvija se kad se rezultat bazira na podacima koji su grupirani. Najčešće se radi o jedinstvenom rezultatu koji daje visoku ocjenu grupaciji čiji su elementi unutar grupe vrlo slični, a grupe vrlo različite. No, sama činjenica da neki algoritam grupacije ima visoku ocjenu interne evaluacije ne znači nužno da su podaci grupirani na takav način optimalni za izvlačenje informacija iz njega. Evaluacija se najčešće radi gledajući međusobne udaljenosti elemenata i grupa, pa algoritmi koji grupiraju skup podataka prema udaljenosti će sami po sebi imati višu ocjenu interne evaluacije, što ne znači da su bolji od drugih. Interna evaluacija trebala bi se koristiti isključivo za ocjenu sličnih algoritama samo s drugačijim parametrima te ako se krivi algoritam postavi za određeni skup podataka ili ako se koristi interna evaluacija kod grupiranja za koje nije predviđena, ocjena će biti nezadovoljavajuća i beskorisna u tom slučaju. Za internu evaluaciju najčešće koristimo jednu od ovih mjera:

- Davies-Bouldinov indeks
- Dunnov indeks
- Koeficijent siluete

4.1.1. Davies-Bouldinov indeks

Idući na redu koji obrađujemo je Davies-Bouldin[1,2] indeks. Računamo ga prema formuli:

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right), j \neq i \tag{4.1}$$

gdje n predstavlja broj grupa, c_i predstavlja centroid grupe i, σ_i predstavlja srednju vrijednost odstupanja svih elemenata grupe i od centroida c_i i $d(c_i, c_j)$ predstavlja udaljenost centroida c_i i

[1]David L. Davies, američki matematičar
[2]Donald, W. Bouldin, američki matematičar

c_j. Kako tražimo grupe sa izrazito malim razlikama unutar grupe i velikim razlikama grupa, bolju će ocjenu kvalitete imati ona grupacija kod koje je Davies-Bouldinov indeks manji.

4.1.2. Dunnov indeks

Dunnov[3] indeks predstavlja omjer minimalne udaljenosti elemenata unutar grupe i maksmialne udaljenosti između grupa te ga računamo ga prema formuli:

$$D = \frac{\min d(i,j)}{\max d'(k)}, 1 \leq i < j \leq n, 1 \leq k \leq n, \tag{4.2}$$

gdje $d(i,j)$ predstavlja udaljenost između grupa i i j, a $d'(k)$ predstavlja međusobnu udaljenost elemenata unutar jedne grupe. Kao što smo vidjeli u hijerarhijskom grupiranju, udaljenost između dvije grupe možemo na više načina odrediti kao i udaljenost elemenata unutar grupe. Odabir načina na koji mjerimo udaljenosti grupa i elemenata unutar grupe ovisi o rasporedu podataka, no moramo imati na umu da jednak način primjenjujemo za cijeli skup podataka. Kod Dunnovog indeksa bolja je ona grupacija čiji je indeks viši.

4.1.3. Koeficijent siluete

Siluetu računamo prema izrazu:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, za |C_i| > 1 \tag{4.3}$$

gdje $b(i)$ predstavlja najmanju srednju udaljenost između elementa i i svih elemenata bilo koje druge grupe, a $a(i)$ srednja udaljenost između elementa i i svih drugih elemanata iste grupe. Grupa za čiji je $b(s)$ najmanji naziva se susjedna grupa te ona služi za izračun koeficijenta. Za što bolju sličnost elemenata grupe i što veću različitost grupa želimo da $a(i) << b(i)$, to jest da $s(i) \to 1$. Za veliki $b(i)$ možemo reći da je element i loše grupiran u odnosu na susjednu grupu. Ako $s(i) \to -1$, možemo reći da bi element i bolje pasao u susjednoj grupi, a za $s(i) \to 0$ govorimo da je element na granici između grupa.

Na kraju, za koeficijent siluete gledamo najveću vrijesnot $s(i)$ kroz cijeli skup podataka:

$$SC = \max_K \tilde{s}(K), \tag{4.4}$$

gdje $\tilde{s}(K)$ predstavlja vrijednost $s(i)$ kroz cijeli skup podataka za neki određeni broj grupa K.

4.2. Eksterna evaluacija

Eksterna evaluacija uspoređuje dobivene rezultate grupiranja s predodređenim rezultatom grupiranja kojeg su grupirali ljudski stručnjaci te takav predodređeni rezultat predstavlja zlatni standard prema kojem bi algoritmi trebali težiti. Eksternu evaluaciju najčešće koristimo da ocijenimo

[3]Joseph C. Dunn, američki matematičar

određeni algoritam te da ga prilagodimo na poznatim primjerima kako bi ga mogli onda koristiti na novim primjerima te da znamo njegovu pouzdanost gledajući kakve rezultate dobivamo na poznatim primjerima.

Razne su mjere uvedene kako bi se bolje mogla izračunati eksterna evaluacija. Neke od mjera uključuju:

- TP (*true positives*) - broj elemenata koji su pridodani grupi u kojoj trebaju biti
- TN (*true negative*) - broj elemenata koji nisu pridodani grupi u kojoj ne bi trebali biti
- FP (*false positive*) - broj elemenata koji su pridodani grupi u kojoj ne bi trebali biti
- FN (*false negative*) - broj elemenata koji nisu pridodani grupi, a trebali bi biti.

Sad ćemo proći neke od indeksa koji koriste ove mjere za eksternu evaluaciju.

4.2.1. Randov indeks

Randov[4] indeks prikazuje koliko su rezultirajuće grupe određenog algoritma slične mjerilu grupa te se računa pomoću formule:

$$RI = \frac{TP + TN}{TP + FP + FN + TN}, \tag{4.5}$$

čije smo parametre već ranije objasnili. Najveća mana Randovog indeksa je to što FP i FN imaju jednaku vrijednost što u nekim primjenama može biti nepoželjno. Postoji par dodataka koji rješavaju taj nedostatak, no mi ih ovdje nećemo proći.

4.2.2. Jaccardov indeks

Jaccardov indeks već smo prije spomenuli kad smo govorili o metrikama u klasterskoj analizi, no on se može primjenjivati i kod eksterne evaluacije te se kod takve primjene može zapisati kao:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP + FP + FN}. \tag{4.6}$$

4.2.3. Fowlkes-Mallowsov indeks

Fowlkes-Mallowsov[5,6] indeks još je jedan način uspoređivanja rezultirajućih grupa i mjernih grupa te kao izlaz izbacuje sličnost između ta dva skupa. Računa se pomoću sljedeće formule:

$$FM = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}}. \tag{4.7}$$

[4]Wiliam M. Rand, američki matematičar
[5]Edward Fowlkes, engleski statističar
[6]Colin L. Mallows, 1930., engleskti statističar

Viša vrijednost indeksa označava veću sličnost rezultirajućih grupa i mjernih grupa.

4.3. Manualna i indirektna evaluacija

Manualna evaluacija zapravo predstavlja ljudsku intervenciju odnosno evaluaciju od strane stručnjaka, a kako možemo pretpostaviti ona je vrlo subjektivna. No ipak valja primjeniti manualnu evaluaciju za slučaj kad ostale zakažu, odnosno kad nam ostale mjere daju dobre rezultate ali nama i dalje ne odgovara raspored elemenata po grupama.

Indirektna evaluacija je zapravo mjera koliko je određeni algoritam grupiranja koristan za određenu primjenu. Kako to dosta ovisi o samoj primjeni, nije moguće odrediti univerzalnu formulu ili izraz za njeno korištenje.

5. Odabir optimalnog broja grupa

Odabir optimalnog broja grupa vrlo je bitna zadaća klasterske analize. Trebala bi se obaviti prije detaljnije analize rezultata grupiranja kako bi bili sigurni da pravilno opisujemo rezultate. Krivim odabirom broja grupa ne možemo dobro opisati rezultate te bilo koja daljnja analiza koja se na to nadovezuje nije ispravna.

Ovo poglavlje napisano je prema [11], [13] i [14].

Odabir optimalnog broja grupa može se raditi grafički i računski. Grafički odabir optimalnog broja grupa odvija se gledajući rezultirajuće grafove (dendrograme za hijerarhijsko grupiranje) te procjenom koliki broj grupa bi najbolje opisivao elemente. Kako smo prije naveli, jedan od najvećih problema particijskog grupiranja je taj što unaprijed moramo odrediti broj grupa. To je dosta nepraktično pošto se radi o setu podataka za koji ne znamo u koliko bi se grupa optimalno mogli smjestiti. Broj grupa također ovisi o obliku i količini podataka.

Kod hijerarhijskog grupiranja kod kojeg imamo dendrogram, grafičko određivanje broja grupa gledamo prema tome gdje odlučimo presjeći dendrogram. S obzirom na to koliko grana presjećemo na dendrogramu toliko ćemo imati grupa na kraju.

Za $K \to n$, gdje n predstavlja ukupni broj elemenata cijelog skupa, a K broj grupa, imamo veću preciznost no u krajnjem slučaju gdje je $K = n$, svaki element je zasebna grupa te smo se efektivno vratili na početak problema. S druge strane, za $K \to 1$, dobivamo smanjenu rezoluciju, jednostavniji prikaz i manju grešku grupacije ali u krajnjem slučaju $K = 1$ svi su elementi smješteni u jednu grupu i nemoguće ih je razlikovati pa smo ponovo u situaciji koja nam je nepogodna. Optimalan broj K postiže balans između preciznosti (minimalne greške) i kompresije (pojednostavljenja).

Postoji nekoliko metoda za određivanje optimalnog K od kojih je najpoznatija grafička metoda "lakta", kod koje broj grupa određujemo pronalaskom lakta na grafičkom prikazu ovisnosti broja grupa o postotku varijacije pojedine grupe. "Lakat" označava broj grupa prije kojeg povećanjem broj grupa varijacija unutar grupe znatno mijenjala, a nakon kojeg se varijacija unutar grupe sporo mijenja porastom broja grupa. Lakat se najčešće očitava iz grafa ovisnosti postotka varijacije i broja grupa (koliko se novih informacija dodaje porastom broja grupa). Taj se lakat ne može uvijek lako grafički razaznati, pa se ova metoda smatra izrazito subjektivnom i nepovjerljivom.

5.1. Calinski-Harabasz indeks

Osim navedenih metoda određivanja broja grupa postoji i Calinski-Harabasz[1,2] indeks koji nam predstavlja mjeru koilko je jedan element sličan svojoj grupi u usporedbi s ostalim grupama. U ovom slučaju svakoj grupi pridodajemo centroid te se ta udaljenost uspoređuje s udaljenosti od globalnog centroida svih elemenata.

Za K grupa skupa podataka $D = d_1, d_2, ..., d_n$ formula za izračunavanje Calinski-Harabasz indeksa glasi:

$$CH = \left[\frac{\sum_{k=1}^{K} n_k||c_k - c||^2}{K-1}\right] / \left[\frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}||d_i - c_k||^2}{N-K}\right], \tag{5.1}$$

gdje n_k predstavlja broj elemenata k-te grupe, c_k predstavlja centroid k-te grupe te N predstavlja ukupni broj elemenata seta. Skraćeno, CH indeks može se zapisati i kao:

$$CH = \frac{a}{b}, \tag{5.2}$$

to jest kao omjer dvaju težina, a i b gdje a predstavlja raspršenost elemenata grupe, a b predstavlja koheziju grupe[3]. Veći rezultirajući indeks predstavlja gušće međusobno različitije grupe dok manji indeks predstavlja suportno. Gledajući kako se indeks mijenja mijenjanjem broja grupa možemo izabrati broj koji najbolje opisuje skup podataka. Valja napomenuti da ne postoji ispravan broj grupa, no ipak gledajući omjer indeksa i broja grupa ponekad se dogodi da nakon određenog broja grupa indeks počinje naglo padati. U tom nam je slučaju lako odrediti optimalan broj grupa, biramo onaj broj grupa nakon kojeg indeks počinje prvi put padati pošto nam je najčešće cilj skup podataka smjestiti u što manji broj grupa. U drugim slučajevima indeks polako raste i opada te u tim slučajevima ako odaberemo broj grupa koji je malo manji ili viši od optimalnog nećemo znatno mijenjati kvalitetu grupiranja te je tada poželjno uzeti manji broj grupa kako bi si olakšali analizu [27].

5.2. Duda-Hart indeks

Vrlo sličan Calinski-Harabasz indeksu je i Duda-Hart[4,5] indeks kojeg ćemo ovdje samo nakratko spomenuti pošto ćemo i taj indeks koristiti kasnije. Razlika između Calinski-Harabasz indeksa i Duda-Hart indeksa je u tome što Calinski-Harabasz indeks prikazuje omjer udaljenosti nekog elementa od centroida određene grupe i centroida svih elemenata zajedno dok Duda-Hart indeks prikazuje omjer međusobne udaljenosti elemenata unutar grupe i elemenata susjedne grupe.

[1] Tadeusz Caliński, 1928., poljski matematičar

[2] Jerzy Harabasz, poljski matematičar

[3] U našem slučaju raspršenost i kohezija grupe zapravo predstavljaju koliko su nam elementi grupe gusto odnosno rijetko raspoređeni

[4] Richard O. Duda, američki profesor elektrotehnike

[5] Peter E. Hart, 1941., američki informatičar i poduzetnik

Valja napomenuti da se Duda-Hart indeks može primjenjivati isključivo kod hijerarhijskog grupiranja kad uvijek promatramo (uspoređujemo) samo dva elementa (ili dvije grupe) istovremeno.

6. Primjena klasterske analize na obrazovne podatke

Klastersku analizu provodit ćemo na rezultatima iz kolegija Inženjerska matematika ET, koja se predaje na drugoj godini Tehničkog fakulteta Sveučilišta u Rijeci, a koje su studenti ostvarili akademnske godine 2021./2022. Identiteti studenata nisu bili poznati, a podatke je ustupio nositelj kolegija. Podaci koje ćemo koristiti su sljedeći:

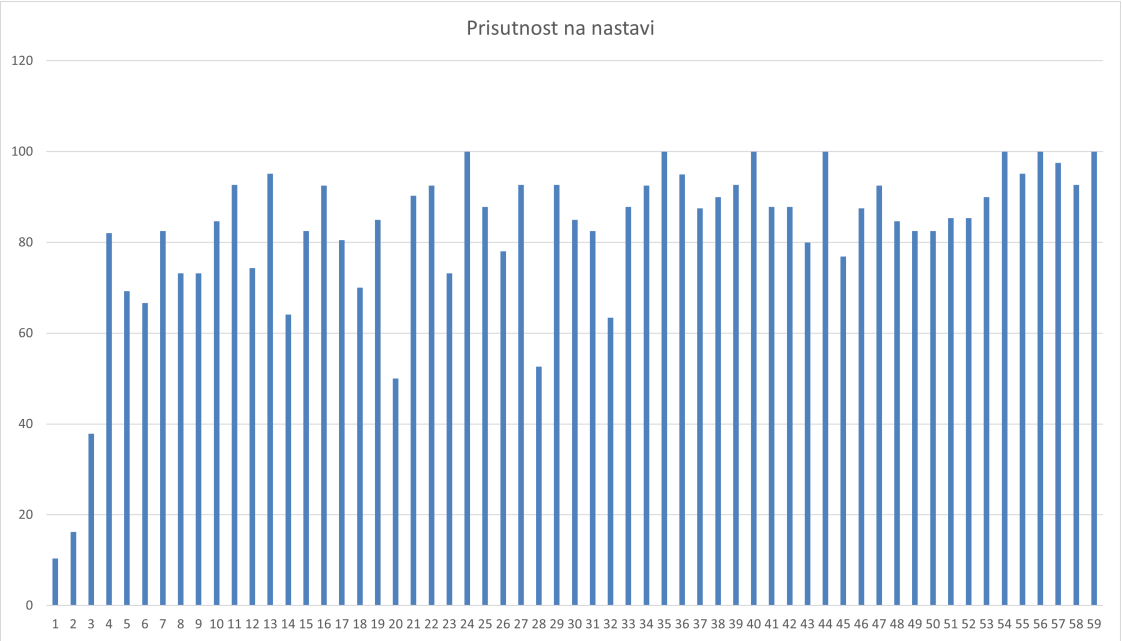
- prisustvo na nastavi (%),
- prolaz na završni ispit (oblik da/ne),
- trend rezultata domaćih zadaća (koeficijent),
- trend rezultata kontrolnih zadaća (koeficijent),
- broj domaćih zadaća kojima je student pristupio,
- broj kontrolnih zadaća kojima je student pristupio,
- broj ukupno sakupljenih bodova iz domaćih zadaća,
- broj ukupno sakupljenih bodova iz kontrolnih zadaća.

Većina ovih podataka čitljiva je direktno iz tablice, dok primjerice trend nije te ćemo njega računati. A što nama trend uopće predstavlja i kako ga računamo ćemo objasniti kasnije u ovom poglavlju. Cilj ove analize je da grupiramo studente s obzirom na njihov pristup kolegiju te ukupni ostvareni rezultat. Rezultati analize pomoći će nam da shvatimo kako pojedine varijable utječu na konačnu ocjenu kako bi mogli u budućnosti savjetovati studente kako da postignu što je bolje rezultate na kolegiju.

Sada ćemo proći neke od varijabli koje ćemo koristiti kako bi bilo jasnije o kojim se točno varijablama radi.

6.1. Prisustvo na nastavi

Prvi skup varijabli koji ćemo proći u detalje biti će prisustvo na nastavi. Prisustvo na nastavi po studentima prikazano je na slici 6.1. Kako smo studente poredali po ukupnom broju postignutih bodova, možemo vidjeti da porestom bodova koje je student sakupio raste i prisustvo na nastavi, odnosno sutdenti koji su imali veće prisustvo na nastavi su na kraju sakupili veći broj bodova. No, barem iz ovoga, nije vidljiva direktna korelacija između konačnih rezultata i ostvarenih bodova, no vidjet ćemo još hoće li nam se otvoriti nova slika kad pogledamo prisustvo po grupama.



Slika 6.1. Prisutnost studenata na nastavi

Dakle, otovo svi studenti se kreću otprilike tu negdje i nema znatnih odstupanja osim studenata 1, 2, 3, 20, 28 i 32. Vidjet ćemo kasnije hoće li ta odstupanja imati neki zančajniji utjecaj na to u koju će grupu studenti biti raspoređeni.

6.2. Trendovi

Trend u našem slučaju predstavlja kretanje rezultata kroz kontrolne zadaće, odnosno ostvaruje li student bolje ili gore rezultate na kontrolnim zadaćama kako semestar odmiče. Dakle, ako je dobiveni postotak na kontrolnim zadaćama niži svakom idućom kontrolnom zadaćom, rezultirajući trend će biti negativan, ako je dobiveni postotak na kontrolnim zadaćama sve veći na svakoj idućoj kontrolnoj zadaći rezultirajući trend će biti pozitivan, a ako su dobiveni rezultati na kontrolnim zadaćama relativno jednaki rezultirajući trend će se sve više približavati 0. Računanje trenda će nam pomoći da ustanovimo trudi li se student više ili manje kako semestar odmiče.

6.2.1. Računanje trendova

Trendove ćemo računati na sljedeći način: grafički ćemo prikazati rezultate na način da nam x-os predstavlja redni broj kontrolne zadaće, a y-os predstavlja ukupni postotak dobiven na svim kontrolnim zadaćama zajedno. Nakon što smo grafički prikazali točke koje predstavljaju rezultate, naći ćemo funkciju u obliku polinoma drugog reda koji najbolje približno opisuju kretanje rezultata kroz semestar. Za pronalazak te funkcije ćemo koristiti naredbu $FitPoly$ unutar GeoGebre koja nam za ulaz uzima listu točaka koje gledamo i red polinoma koji želimo da nam te točke opiše, a za izlaz izbacuje polinom koji najbolje opisuje zadane točke. Drugim riječima, formirali

smo regresijski polinom drugog stupnja. Nakon što smo to napravili izračunat ćemo derivaciju dobivene funkcije te ćemo gledati nagib rezultirajućeg pravca. Valja napomenuti da za funkciju $Fitpoly$ možemo odabrati da nam Geogebra točke pokaže funkcijama viših redova, no jednostavnosti radi koristit ćemo funkciju drugog reda jer znamo da derivacijom te funkcije dobivamo dobivamo polinom prvog stupnja čiji je grafički prikaz pravac te nam je puno lakše odrediti trend gledajući nagib pravca.

Sad ćemo na jednom primjeru pokazati računanje trenda.

Primjer 6.1. *Uzet ćemo jednog studenta za primjer da vidimo kako ćemo mu izračunati njegov trend tijekom semestra. Student kojeg ćemo gledati biti će student broja ID 50.*

Student ID 50 tijekom semestra postigao je sljedeće rezultate na kontrolnim zadaćama:

Tablica 6.1. Rezultati kontrolnih zadaća za studenta ID 50

Kontrlona zadaća	KZ 1	KZ 2	KZ 3	KZ 4
Dobiveni rezultat	11.25	13.5	9.25	13.75

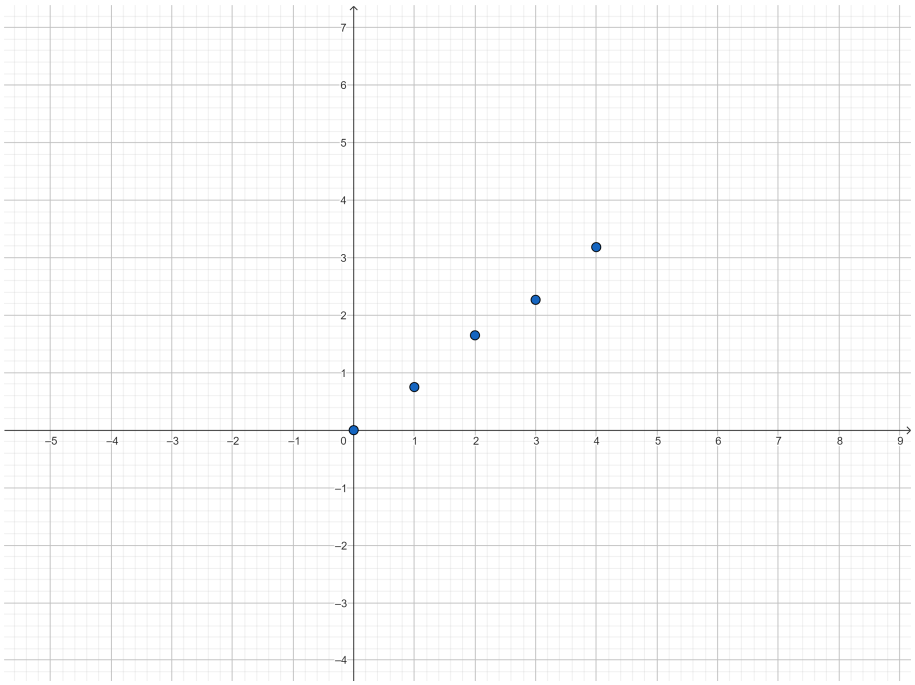
Sad ćemo za te rezultate tražiti postotak svake kontrolne zadaće na način da ćemo sve podijeliti s ukupnim brojem bodova (15). Nakon toga plotiramo dobivene postotke na graf na način da nam x-os predstavlja redni broj kontrolne zadaće, a y-os ukupni postotak na dosadašnjim kontrolnim zadaćama. Za naš primjer računanje koordinata točaka izgledalo bi ovako:

$$\begin{aligned} kz1 =& \frac{11.25}{15} = 0.75, \\ kz2 =& \frac{13.5}{15} = 0.9, \\ kz3 =& \frac{9.25}{15} = 0.62, \\ kz4 =& \frac{13.75}{15} = 0.92, \\ KZ1 =& (1, kz1) = (1, 0.75), \\ KZ2 =& (2, kz1 + kz2) = (2, 1.65), \\ KZ3 =& (3, kz1 + kz2 + kz3) = (3, 2.27), \\ KZ4 =& (4, kz1 + kz2 + kz3 + kz4) = (4, 3.19). \end{aligned} \tag{6.1}$$

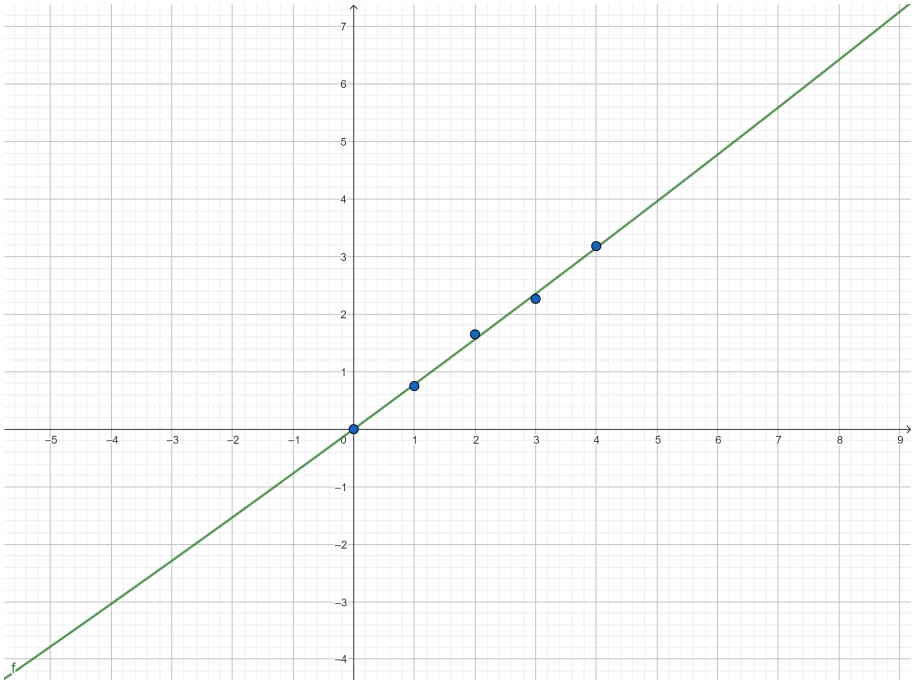
Dobiveni uređeni parovi grafički su prikazani na slici 6.2.

Sad kad smo našli uređene parove koji nam predstavljaju toče na grafu tražimo polinom drugog reda koji najbolje opisuje zadane točke. Za to koristimo se funkcijom $FitPoly$ unutar GeoGebre, a rezultat provedbe funkcije te pripadajući polinom vidljivi su na slici 6.2.

Nakon što smo i to odredili da vidimo kako se funkcija kreće kroz kontrolne zadaće računamo derivaciju dobivenog polinoma te gledamo nagib dobivenog pravca, to jest gledamo koji broj



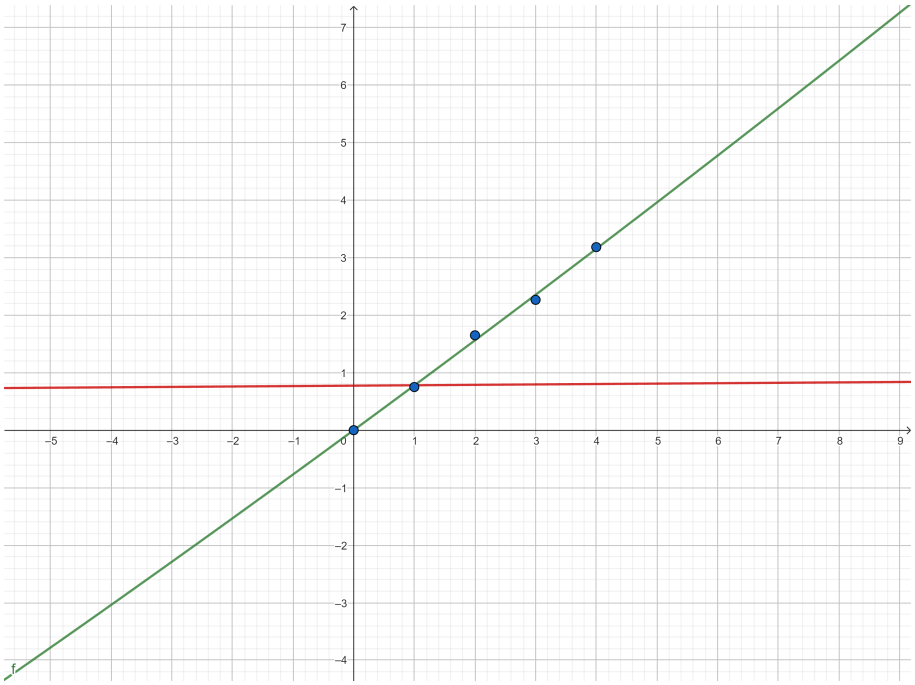
Slika 6.2. Plotirane točke za primjer rezultata kontrolnih zadaća studenta ID 50



Slika 6.3. Plotirane točke uz odgovarajući polinom za primjer rezultata kontrolnih zadaća studenta ID 50

množi x unutar dobivene funkcije. Grafički dobivena funkcija prikazana je na slici 6.3.

Sad smo izračunali sve što nam treba te samo gledamo nagib pravca, u ovom slučaju on iznosi 0.00714 što nam dakle zapravo predstavlja da se student ID 50 tijekom semestra popravlja u rezultatima pošto je nagib pozitivan. Koliko je dobar ili loš taj trend trenutno ne znamo jer ne znamo koliko trendovi iznose za ostale studente. Tek usporedbom sa ostalim studentima možemo vidjeti gdje stoji naš student ID 50.



Slika 6.4. Plotirane točke uz odgovarajući polinom i njegovu derivaciju za primjer rezultata kontrolnih zadaća studenta ID 50

Na isti način ćemo računati i trendove domaćih zadaća te ćemo onda kasnije sve te rezultate ubaciti u klastersku analizu da vidimo postoje li neke grupe prema tome koje trendove studenti imaju i koliki je njihov konačan rezultat.

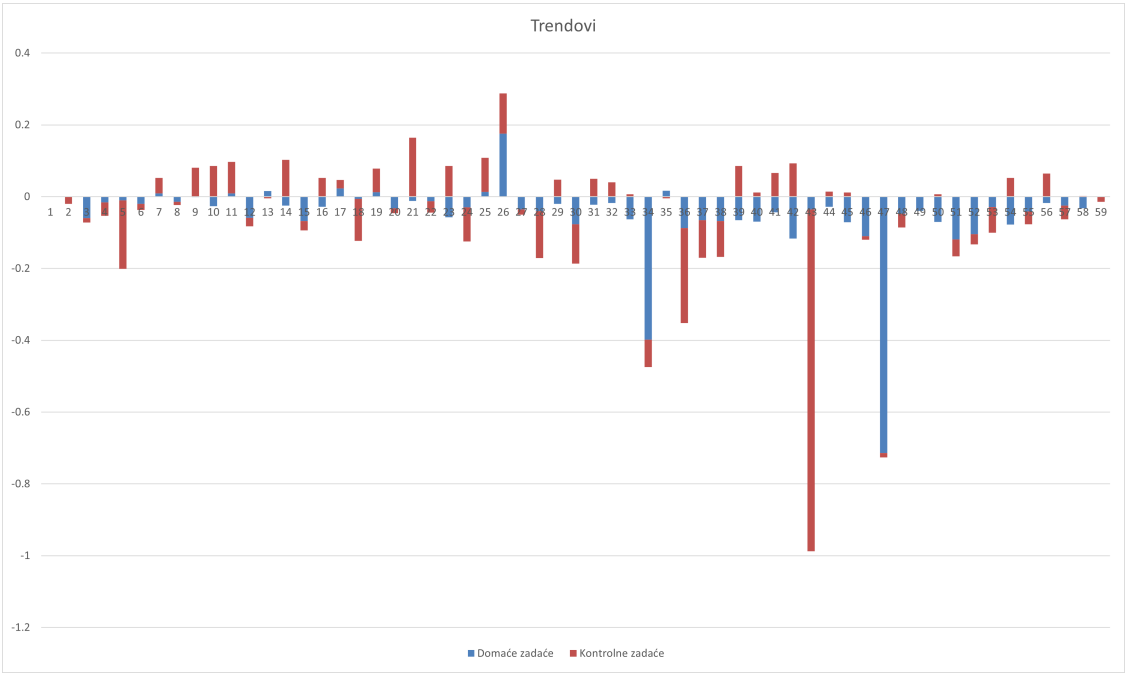
6.2.2. Rezultirajući trendovi po studentima

Sad kad smo objasnili što u našem kontekstu znači trend, kako se koristi i računa, provest ćemo ga kroz studente i vidjeti kako koji student ima trendove za kontrolne i domaće zadaće. Ukupni trendovi koje smo dobili prikazani su na slici 6.5.

Ono što je jasno vidljivo je to da studenti ispod ili blizu granice prolaza (student ID 10 je student s najnižim brojem bodova koji je i dalje prošao na završni) imaju znatno bolje trendove od studenata koji su srednje iznad granice. Već tu vidimo neku moguću korelaciju između ostvarenih bodova i trendova, no detaljniju analizu po grupama ćemo proći analizom grupa.

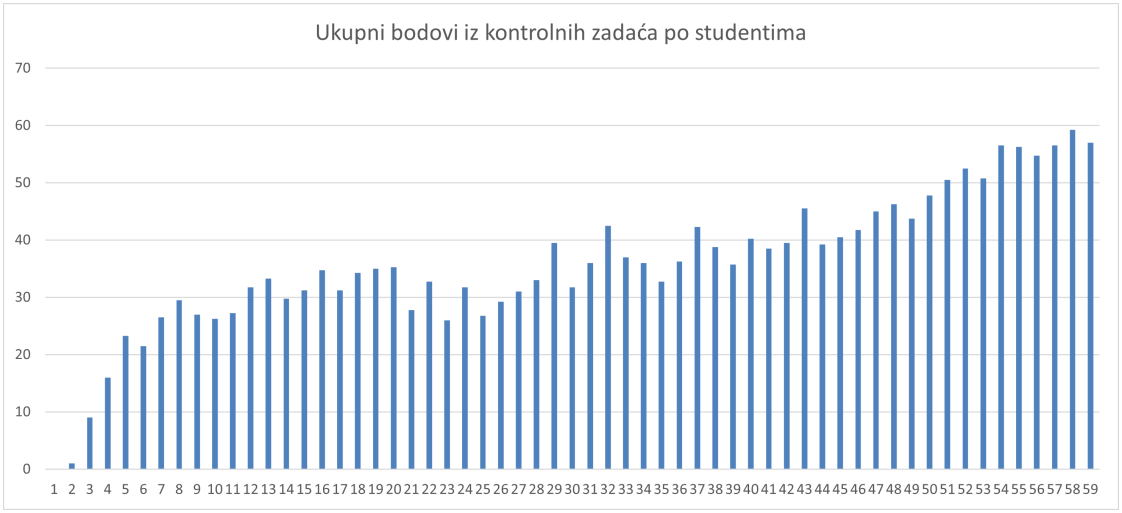
6.3. Ukupni ostvareni bodovi iz kontrolnih zadaća

Posljednji podaci koje ćemo detaljnije analizirati po grupama su ukupni ostvareni bodovi iz kontrolnih zadaća. Bodove iz bonusa kao i bodove iz domaćih zadaća nećemo detaljnije obrađivati pošto zapravo prate isti princip kao i bodovi iz kontrolnih zadaća, to jest povezanost s grupiranjem je gotovo identična kao i kod kontrolnih zadaća pa zapravo analizom rezultata kontrolnih zadaća analiziramo i domaće zadaće kao i bonuse te daljnjom analizom ta dva paramatra nećemo saznati ništa novo.



Slika 6.5. Rezultirajući trendovi po studentima gdje su plavom bojom označeni trendovi iz domaćih zadaća, a crvenom trendovi iz kontrolnih zadaća

Ukupni ostvareni bodovi iz kontrolnih zadaća po studentima mogu se vidjeti na slici 6.6.



Slika 6.6. Ukupni postignuti rezultati iz kontrolnih zadaća po studentima

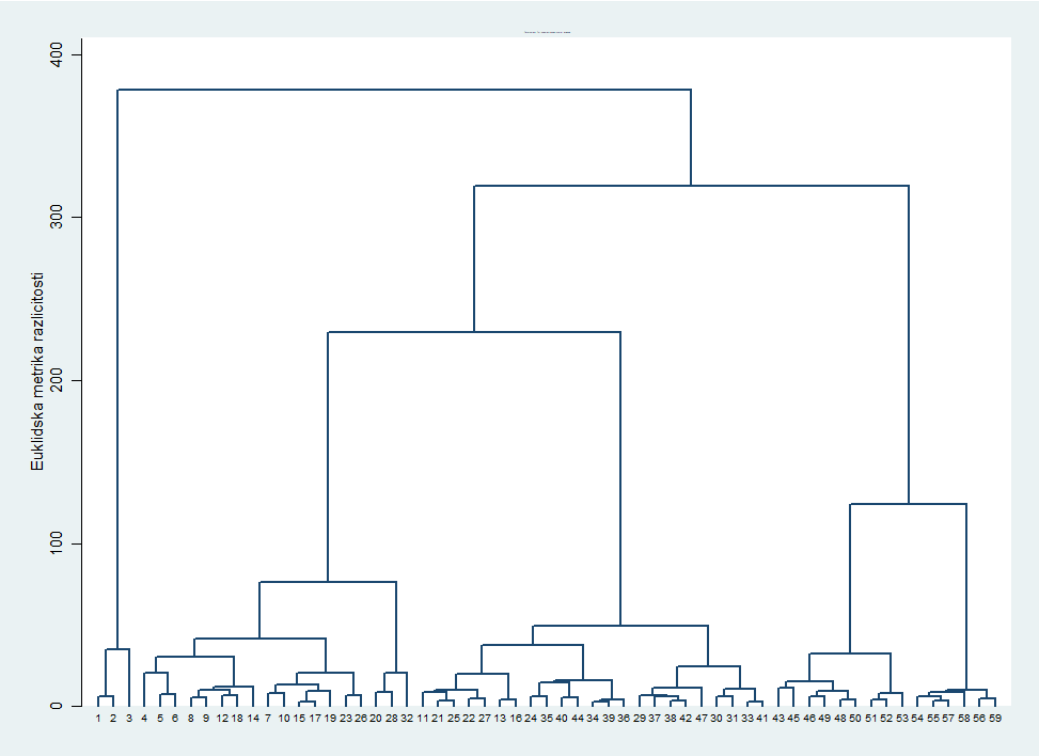
Jasno je vidljivo da što su bolji studenti to imaju više bodova iz kontrolnih zadaća, to jest uspjeh na kontrolnim zadaćama je veliki razlog zašto neki studenti uspije ili ne uspije proći na završni ispit.

Osim ovih podataka koristit ćemo još i podatke o tome je li student prošao na završni ispit kao i broj zadaća kojima je student pristupio (kao što smo naveli na početku), no ti su nam podaci više kao pomoćni kako bi napravili još veću raspodjelu između grupa te tako dobili grupe koje se mogu lakše opisivati. Te podatke nećemo detaljnije prolaziti ali je nužno napomenuti da ćemo i njih ubacivati u algoritam.

Sad kad smo vidjeli koje ćemo sve podatke koristiti i što oni predstavljaju ubacit ćemo naš skup podataka u algortiam da vidimo kako će ih algoritam grupirati. Klastersku analizu provodit ćemo u softverskom paketu *Stata 13*.

6.4. Klasterska analiza obrazovnih podataka korištenjem hijerarhijskog grupiranja

Klstersku analizu obrazovnih podataka prvo ćemo provesti hijerarhijskim grupiranjem. Za analizu ćemo koristit Wardovu metodu i Euklidsku udaljenost, oboje smo već ranije objasnili. Nakon što dobijemo rezultate prikazat ćemo ih na dendrogramu te ćemo ih interpretirati. Korištenjem, dakle, Wardove metode i Euklidske udaljenosti dobili smo rezultate koje prikazane na dendrogramu izgledaju ovako:



Slika 6.7. Rezultirajući dendrogram

Na donjem dijelu dentograma su prikazani brojevi koji predstavljaju ID studenta odnosno njihov poredak po osvojenim bodovima tijekom semestra počevši od najnižeg prema najvišem. Prije nego što idemo dalje u analizu potrebno nam je odrediti optimalan broj grupa na koji bi mogli podijeliti studente. Gledajući dendrogram ovisno o tome gdje ga presjećemo dobit ćemo različit broj grupa. Ako ga presjećemo na najnižoj razini dobit ćemo 58 grupa, odnosno isti broj grupa koliko i ima studenata, no to ne želimo. Ako ga uopće ne presjećemo dobit ćemo jednu grupu no u tom slučaju nismo ništa postigli tako da nam ni to nije dobar odabir. Gledajući detaljnije dendrogram vidljivo je su nam solidne opcije za odabir broja grupa 3, 4 ili 7. Provest ćemo ove rezultate kroz jednu od ranije spomenutih informativnih kriterija da vidimo koje ćemo rezultate dobiti te da nam olakšaju odabir.

6.4.1. Odabir optimalnog broja grupa

Za odabir optimalnog broja grupa imamo na raspolaganju dvije metode:

- Calinski-Harabasz indeks,
- Duda-Hart indeks.

Koristit ćemo i jedan i drugi indeks i usporedit ćemo rezultate. Korištenjem Calinski-Harabasz indeksa dobili smo sljedeće rezultate:

Tablica 6.2. Rezultati računanja CH indeksa za različit broj grupa

Broj grupa	Calinski-Harabasz indeks
2	52.93
3	62.25
4	75.47
5	79.52
6	84.48
7	78.75
8	75.38
9	72.13
10	75.15
11	75.42
12	78.82
13	78.28
14	77.76
15	79.28

Gledajući rezultate vidimo da indeks počinje opadati nakon 6 grupa te nam 6 grupa zapravo predstavlja optimalan broj grupa za naš skup podataka. Pošto je razlika između manjeg i većeg broja grupa dosta velika, odabir broja grupa koji nije optimalan će nam potencionalno iskriviti sliku grupa (neki studenti koji bi inače možda bili grupa za sebe će biti pridodani jednoj od drugih grupa iako nisu posve slični) te ćemo najbolje rezultate klasterske analize prema CH indeksu dobiti analizom 6 grupa. Ako prikažemo sad indeks pomoću grafa koji prikazuje ovisnost indeksa o broju grupa jasnije postaje da je 6 grupa optimalan broj.

Vidljivo je da se na 6 grupa nalazi "lakat" odnosno funkcija ima najvišu vrijednost. Ostali broj grupa, bio on manji ili veći prema CH indeksu dosta mijenja rezultate.

6.4.2. Usporedba CH i Duda-Hart indeksa

Usporedit ćemo sad rezultate koje smo dobili CH indeksom sa Duda-Hart indeksom da vidimo hoćemo li dobiti iste rezultate ili ne. Za različit broj grupa računali smo Duda-Hart indeks i dobili sljedeće rezultate:

Slika 6.8. Promjena CH indeksa mijenjanjem broja grupa

Tablica 6.3. Rezultati računanja Duda-Hart indeksa za različit broj grupa

Broj grupa	Duda-Hart indeks
2	0.5185
3	0.5883
4	0.2530
5	0.5522
6	0.6940
7	0.6952
8	0.5617
9	0.0353
10	0.4580
11	0.5040
12	0.4632
13	0.5012
14	0.1642
15	0.1292

Vidljivo je da temeljem Duda-Hart indeksa optimalan broj grupa 3. To možemo vidjeti iz tablice na način da gledamo kad prvi put indeks kreće padati, dakle na broj grupa 3. Usporedimo li sad rezultate s CH indeksom vidimo da ne dobijemo iste rezultate. Dakle, sa CH indeksom vidimo da nam je optimalan broj grupa 4, 6 ili 7 dok sa Duda-Hart indeksom vidimo da je optimalan broj grupa 3. Na kraju je na nama da odaberemo koji ćemo broj grupa odabrati pošto znamo s kakvim skupom podataka baratamo. Oba indeksa trebala bi samo služiti kao pomagala kod odabira broja grupa ali na kraju mi biramo koji nam je broj grupa najbolji. Pošto se ovdje radi o relativno malom skupu, najbolje bi nam zapravo i bilo odabrati broj grupa preko dendrograma presjecanjem grana. Odabrat ćemo da nam je broj grupa 4, što nam je jedan od optimalnog broja prema CH indeksu.

6.4.3. Detaljnija analiza grupa

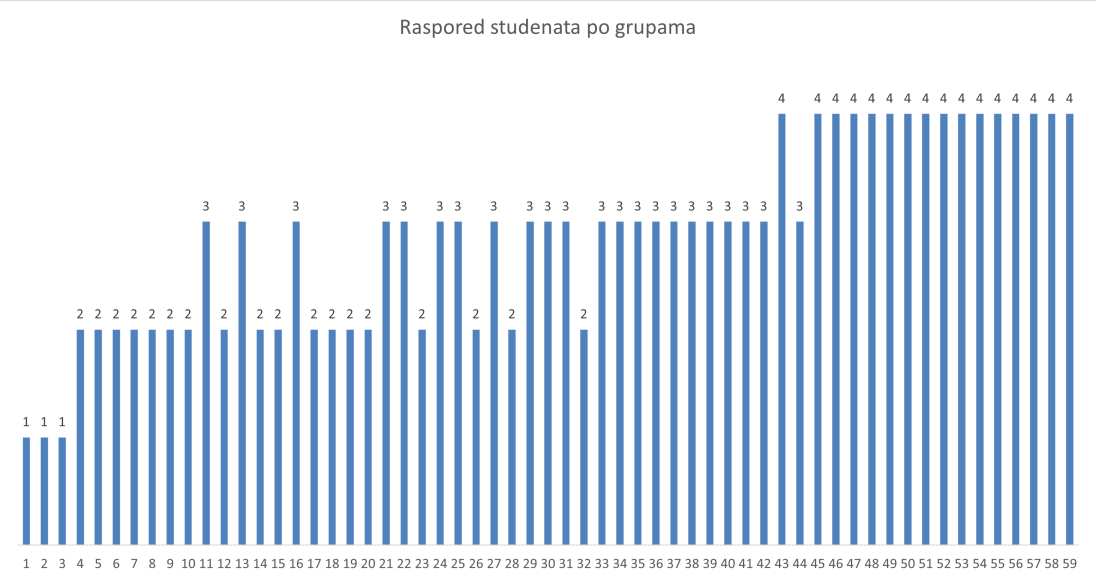
Dakle, imamo skup podataka, dobili smo grupe, odredili smo koji ćemo broj grupa odabrati, na nama je sad jedino preostalo detaljnije analizirati te grupe.

Prvo ćemo definirati 4 grupe koje imamo. Presjecanjem dendrograma na način da presječemo 4 grane dobivamo sljedeće grupe:

Tablica 6.4. Raspored studenata po grupama korištenjem hijerarhijskog grupiranja

Grupa	ID studenta
1	1, 2, 3
2	4, 5, 6, 7, 8, 9, 10, 12, 14, 15, 17, 18, 19, 20, 23, 26, 28, 32
3	11, 13, 16, 21, 22, 24, 25, 27, 29, 30, 31, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 44
4	43, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59

Grafički raspored studenata po grupama prikazan je na slici 6.9.



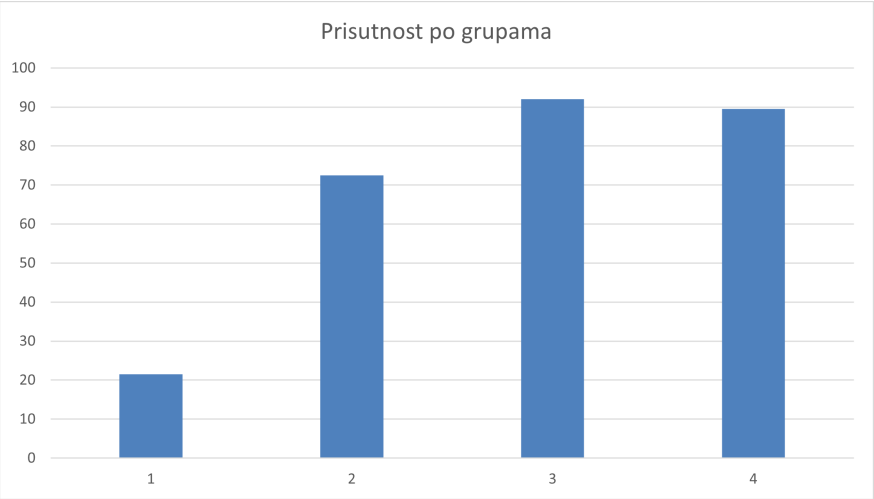
Slika 6.9. Raspored studenata po grupama

6.4.4. Prisustvo na nastavi

Prvo što bi smo mogli detaljnije analizirati je prisustvovanje nastavi. Računamo srednju vrijednost prisustva po grupi te također gledamo i standardnu devijaciju[1]. Prosjek prisustva po grupi prikazan je na slici 6.10.

Grupa 1 u prosjeku ima dolaznost od svega 21.47% uz standardnu devijaciju od 14.48%.
Grupa 2 ima prosječnu dolaznost od 72.51% uz standardnu devijaciju od 10.29%.
Grupa 3 ima najvišu prosječnu dolaznost i ona iznosi 92.04% uz standardnu devijaciju od 4.95%.

[1]Standardna devijacija je mjera koja nam govori koliko su vrijednosti podataka skupa blizu ili daleko od prosjeka

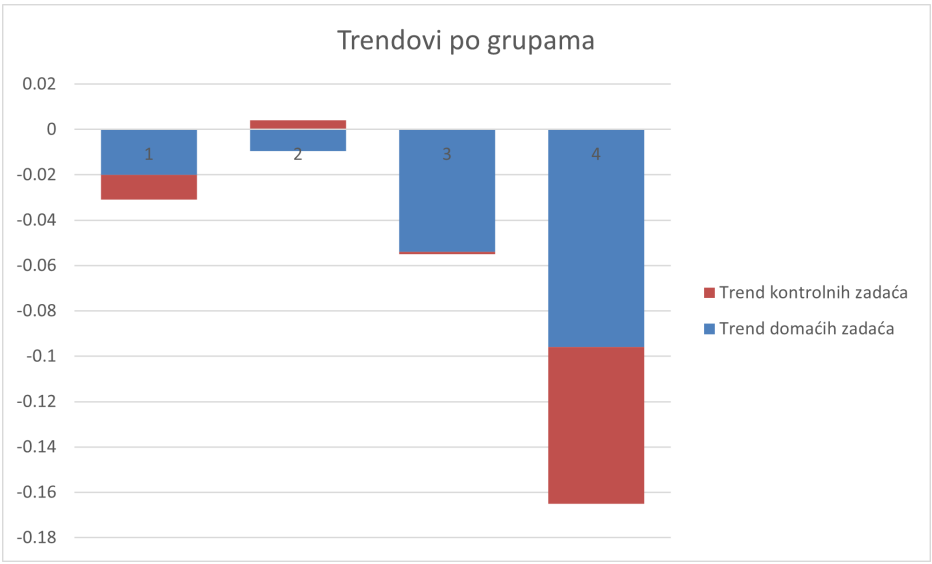


Slika 6.10. Prisutnost studenata po grupama

Grupa 4 ima nešto lošiju prosječnu dolaznost od 89.54% uz standardnu devijaciju od 7.55%.
Iz grafa je vidljivo da najbolju prisutnost ima grupa 3, dok najgoru ima grupa 1. Grupa 4 ima solidnu prisutnost no ne najbolju bez obzira što na kraju imaju najbolje rezultate te grupa 2 ima solidnu prisutnost koja je i dalje iznad obavezne prisutnosti koja je definirana pravilima kolegija. Uzevši u obzir da grupa 2 sadržava puno studenata koji nisu sakupili dovoljno bodova za proći na završni ispit, iz ovog vidimo da čak i kad studenti imaju minimalnu obaveznu prisutnost, da to ne garantira prolazak na završni. Iz ovog, dakle, možemo zaključiti da prisutnost ima korelaciju s boljim uspjehom na predmetu, no ne potpunu.

6.4.5. Trend rezultata iz domaćih i kontrolnih zadaća

Sljedeće na redu podatke koje ćemo gledati su trendovi iz kontrolnih i domaćih zadaća, prikazat ćemo ih na istom grafu koji se može vidjeti na slici 6.11.



Slika 6.11. Srednje vrijednosti trendova po grupama

Jasno je sad vidljivo kako se trendovi kreću po grupama. Trend kontrolnih zadaća grupe 1 iznosi -0.011 uz standardnu devijacjiu od 0.01, dok im je trend domaćih zadaća -0.02 uz standardnu devijaciju od 0.035. Valja napomenuti da grupa 1 i nema tako loše trendove no to je većinom zbog toga što dva studenta uopće nisu sudjelovala u nastavi pa je zbog toga njihov trend bio relativno nizak, to jest trend i ne može ni biti izrazito negativan jer nisu ni rješavali zadaće, a te koje i jesu riješili su riješili izrazito loše pa se ni ne mogu još više pogoršati.

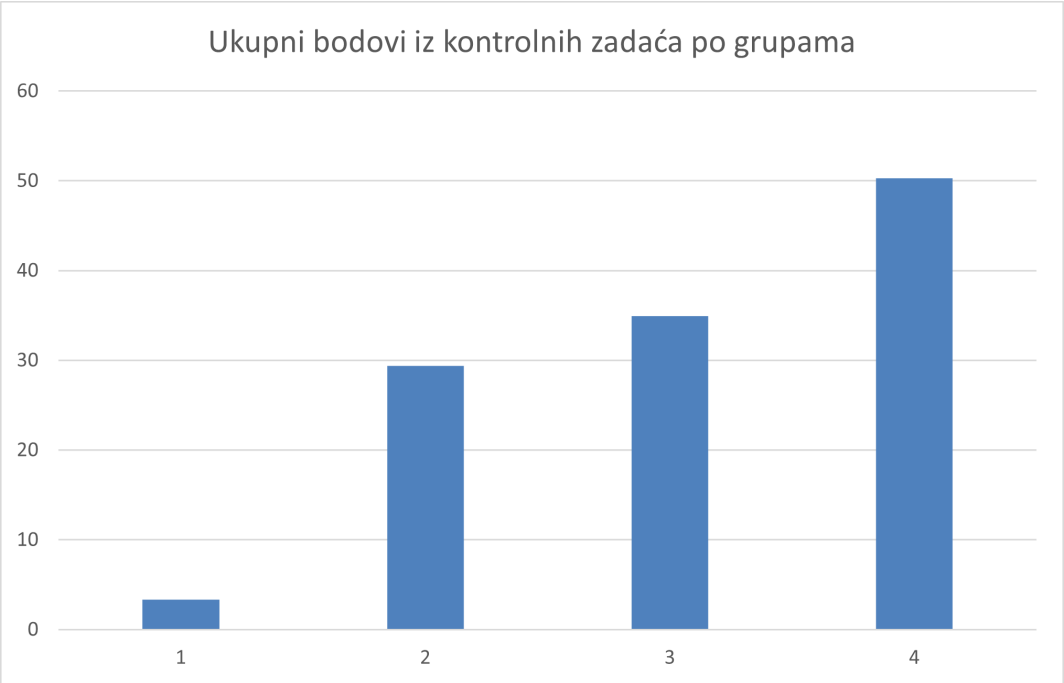
Trend kontrolnih zadaća za grupu 2 iznosi 0.004 uz standardnu devijaciju od 0.085 te je zapravo jedini pozitivan trend od svih grupa uključujući i grupu 4 koja je ostvarila najbolje rezultate na kraju. Iz toga možemo zaključiti da se zapravo studenti koji su na granici prolaska puno više trude pred kraj semestra i pretpostavka je da se smanjenjem mogućih ostvarenih bodova do kraja semestra studenti počnu puno više truditi na nastavi kako bi nadoknadili loš rad na početku semestra. Time si pokušavaju dati najbolju moguću šansu za prolaskom dalje. Trend kontrolne zadaće im je kao i kod ostalih negativan i iznosi -0.00955 uz standardnu devijacjiu od 0.0523 što nam također govori da se studenti kad im je stiska oko bodova više fokusiraju na kontrolne zadaće nego na domaće što je i razumljivo jer kontrolne zadaće donose puno više bodova od domaćih. No i dalje imaju najmanji negativni trend domaćih zadaća i to nam isto tako upućuje da osjećaju pritisak pred kraj. To nam zapravo govori da ti studenti mogu ostvariti dobre rezultate na zadaćama, no iz nekog razloga se ne aktiviraju dovoljno na početku semestra i time riskiraju prolazak na završni ispit.

Trend kontrolnih zadaća grupe 3 iznosi -0.001 uz standardnu devijaciju od 0.095. Iz ovoga je vidljivo da su studenti grupe 3 relativno konzistentni u rješavanju kontrolnih zadaća tijekom semestra. Trend domaćih zadaća grupe 3 iznosi -0.0539 uz standardnu devijaciju od 0.0848.

Trend kontrolnih zadaća grupe 4 iznosi -0.069 uz standardnu devijaciju od 0.238 te je zapravo najveći negativan trend od svih grupa. To nam govori da se studenti grupe 4 u prosjeku puno više opuste pred kraj semestra što je i razumljivo jer znaju da su sigurno prošli na završni ispit pa je pretpostavka da se pred kraj semestra krenu više fokusirati da druge predmete. Trend kontrolnih zadaća grupe 4 iznosi -0.096 uz standardnu devijaciju od 0.168. Valja napomenuti da oba trenda za grupu 4 pogoršavaju dva studenta, ID 43 za trend kontrolnih zadaća i ID 47 za trend domaćih zadaća koji imaju drastične padove u usporedi sa susjednim im kolegama te kao takvi imaju utjecaj na krajnje rezultate. Tu zapravo primjećujemo problem koji smo bili opisali još ranije kod odabira grupa. Da smo, na primjer, odabrali 6 grupa moguće je da bi upravo ti studenti bili smješteni s nekim drugim studentima ili možda čak u zasebnoj grupi, a ne sa najboljim čime bi krajnji rezultati grupa bili drastično drugačiji. U našem slučaju znamo što je problem no kod drugih analiza, ponajviše kod analiza velikog skupa podataka, taj uvid možda neće biti moguć i moguće je da samo prihvatimo rezultate kakvi jesu.

6.4.6. Ukupni ostvareni rezultati iz kontrolnih zadaća

Gledajući ukupne ostvarene bodove po grupama dobijemo graf koji je vidljiv na slici *6.12.*.



Slika 6.12. Ukupni postignuti rezultati iz kontrolnih zadaća po grupama

Iz ovog grafa vidljivo je da su studenti svake iduće grupe sve bolji u ostvarenim bodovima od prethodne.

Grupa 1 u prosjeku je ostvarila 3.33 od mogućih 60 bodova uz standardnu devijaciju od 4.93 boda. Jasno je vidljivo da se studenti grupe 1 uopće ne trude oko kontrolnih zadaća, na većinu uopće ni ne izlaze.

Grupa 2 u prosjeku je ostvarila 29.4 od mogućih 60 bodova uz standardnu devijaciju od 5.95 bodova. Dakle, studenti grupe 2 su na granici sakupljanja pola mogućih bodova tijekom semestra. Gledajući raspored studenata po grupama vidimo da je dosta studenata koji su prošli na završni ispit i dalje u grupi 2. Najčešće se radi o studentima koji se znatno zapuste kako odmiče semestar te se najčešće radi o studentima kojima je bitan samo prolaz, a ne i konačna ocjena.

Grupa 3 u prosjeku je ostvarila 34.94 od ukupno 60 mogućih bodova uz standardnu devijaciju od 4.39 bodova. Gledajući raspored grupa vidljivo je da su apsolutno svi studenti grupe 3 sakupili dovoljno bodova za izaći na završni ispit. Dakle, kao što smo mogli i predvidjeti, ukoliko u prosjeku studenti na kontrolnim zadaćama sakupe više od 50% mogućih bodova da svi unutar grupe sakupe na kraju dovoljno bodova za izaći na završni ispit.

Grupa 4 je u prosjeku ostvarila 50.28 od mogućih 60 bodova uz standardnu devijaciju od 6.05 boda. U usporedbi sa ostalim grupama imaju daleko najbolje rezultate kontrolnih zadaća što se isto tako preslikava na kako na domaće zadaće tako i na bonuse.

6.4.7. Zaključak klasterske analize provedene metodom hijerarhijskog grupiranja

Prolaskom kroz skup podataka grupirali smo studente na način da smo svakom novom grupom minimalno povećali varijaciju grupe (Wardova metoda) te smo vidjeli neke zanimljivosti vezane za grupe i njihovu aktivnost tijekom nastave. Sada kada smo prošli najbitnije varijable možemo detaljnije opisati grupe jednu po jednu te kakve studente svaka grupa sadržava.

Grupa 1 su studenti koji se od početka ne trude uopće na nastavi. Ne dolaze na predavanja, ne rješavaju zadaće bile one kontrolne ili domaće i ne pokazuju nikakvu zainteresiranost za predmet. Takvim studentima potrebna je motivacija za studiranje u cjelosti, a ne za ovaj predmet isključivo te zapravo nema jednostavnog rješenja za motivaciju tih studenata, a pogotovo ne samo za ovaj predmet.

Grupa 2 su studenti na granici ili blizu granice prolaska na završni ispit. To su studenti koji imaju dovoljnu dolaznost ali koja je na granici da bude ispod obavezne. No, vidjeli smo da bez obzira što dolaze na predavanja, nije im vjerojatan prolaz. Studenti grupe 2 nerijetko pokazuju nezainteresiranost za predmet bilo to pred kraj semestra ili početkom, no jasno se vidi povezanost unutar grupe koja pokazuje manjak motivacije tijekom semestra. Unutar grupe 2 se također nalazi i dosta studenata koji su daleko iznad granice prolaza, no i dalje su smješteni u grupu 2 iz razloga što pokazuju nemotiviranost za barem jedan od elemenata kolegija, a najčešće se radi o rješavanju domaćih zadaća. Točnije, studenti grupe 2 koji su prošli na završni ispit i dalje se nalaze u toj grupi jer pokazuju nezainteresiranost u ostale aspekte kolegija. Valja napomenuti da ipak grupa 2 od svih jedina imaju pozitivan trend rješavanja kontrolnih zadaća, a isto tako imaju najmanji negativan trend rješavanja domaćih zadaća. To može biti radi toga da im kolokviji stvaraju nekakav pritisak koji nemaju dok rješavaju domaće zadaće, no isto tako može značiti da bolje savladavaju gradivo kad se odvoji na manje dijelove te veliki obim gradiva kao kod kontrolnih zadaća im eventualno stvara problem koji je moguće riješiti i koji se iznenada riješi pred kraj semestra kad im se stvori pritisak za bodove. Isto tako imaju najmanji negativan trend domaćih zadaća vrlo vjerojatno jer im domaće zadaće predstavljaju zadnju nadu za sakupljanje dovoljno bodova za prolazak na završni ispit, a to im je zapravo pritisak koji tek osjete pred kraj te radi toga najvjerojatnije žele sakupiti što više bodova na koji god način je moguće.

Grupa 3 su studenti koji su tijekom semestra dosta konzistentni barem što se tiče rješavanja kontrolnih zadaća. Ti studenti ipak žrtvuju rješavanje domaćih zadaća kako bi postigli svoj cilj ali s druge strane imaju drugi najbolji trend rezultata iz kontrolnih zadaća. Studenti grupe 3 također imaju i najbolju dolaznost od svih grupa uključujući i grupu 4 koja ima najbolje rezultate na kraju. U prosjeku bi to bili studenti koji bi trebali predstavljati neki standard prema kojem bi pogotovo lošiji studenti trebali težiti. Lošije studente bi se trebalo motivirati da dođu do razine na kojoj su studenti grupe 3 jer studenti grupe 3 su svi prošli na završni ispit, a opet nisu toliko daleko u smislu bodova od svojih kolega iz grupe 2 te je njihov broj bodova definitivno moguć za ostvarit za svakog studenta. Bez obzira što nisu ostvarili toliko dobre rezultate kao što je i grupa 4, oni pokazuju konstantan rad tijekom cijelog semestra te ne posustaju pred kraj. Negativan trend

domaćih zadaća ipak je cijena koju plaćaju kako bi zadržali konstante rezultate tijekom semestra. Sve lošije studente trebalo bi se poticati da prate kolege koje se nalaze u grupi 3.

Grupa 4 su odlikaši koji ostvaruju najbolje rezultate iz predmeta. Iako je njihova dolaznost slabija od studenata grupe 3, ostvaruju najbolje rezultate na svim vrstama zadaća bile one kontrolne ili domaće. No, bitno je naglasiti da oni također u prosjeku imaju daleko najgori trend rezultata kako kontrolnih tako i domaćih zadaća. To je naravno za očekivati pošto dosta rano shvate da će imati dovoljno bodova za prolazak na završni te se ranije krenu fokusirati da druge predmete gdje to možda nije slučaj. Nad tim studentima nije potrebna nikakva dodatna motivacija ili podrška pošto pokazuju svojim rezultatim da shvaćaju težinu predmeta i imaju dovoljnu disciplinu učiti pravovremeno za svaku od zadaća. Bez obzira što imaju negativan trend, i dalje u prosjeku ostvaruju vrlo dobre rezultate na zadaćama. Negativan trend je u stvari samo pokazatelj koliko se na početku više trude nego pred kraj, no nikako ne umanjuje njihove rezultate na kraju koji su i dalje poželjni. Kao što smo prije rekli, ipak je bitno naglasiti da većina studenata ipak nema toliko loše trendove, no par njih koji se nalaze u grupi 4 imaju vrlo negativne trendove i time zapravo spuštaju prosjek grupe. Ako bi gledali medijan trenda umjesto prosjeka vidjeli bi drugačiju sliku gdje nam specifično ta dva studenta koji imaju vrlo loše trendove ne bi radili iskrivljenje nad prosjekom.

6.5. Klasterska analiza obrazovnih podataka korištenjem particijskog grupiranja

Osim hijerarhijskog grupiranja, analizirat ćemo rezultate klasterske analize bazirane na particijskom grupiranju te usporediti rezultate da vidimo hoće li isti studenti pripadati istim grupama, to jest kako će se grupe razlikovati u usporedbi s hijerarhijskim grupiranjem.

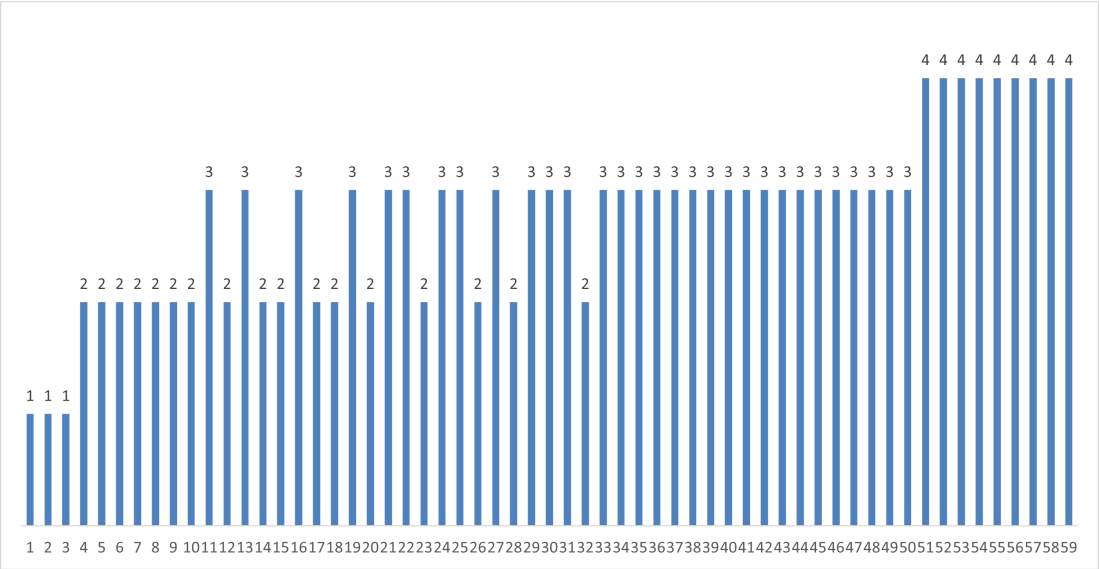
Broj grupa ćemo također odabrati kao i kod hijerarhijskog grupiranja. Bitno je za naglasiti da, za razliku od hijerarhijskog grupiranja, ovdje moramo odabrati broj grupa prije nego idemo u analizu što nije bio slučaj kod hijerarhijskog grupiranja gdje smo prvo proveli analizu pa prema rezultatima odabrali broj grupa. Dakle, ako odaberemo da nam je broj grupa 4 te provođenjem kroz algoritam particijskog grupiranje dobivamo sljedeći raspored studenata:

Tablica 6.5. Raspored studenata po grupama korištenjem particijskog grupiranja

Grupa	ID studenta
1	1, 2, 3
2	4, 5, 6, 7, 8, 9, 10, 12, 14, 15, 17, 18, 20, 23, 26, 28, 32
3	11, 13, 16, 19, 21, 22, 24, 25, 27, 29, 30, 31, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49
4	50, 51, 52, 53, 54, 55, 56, 57, 58, 59

Grafički raspored studenata po grupama vidljiv je na slici 6.13.

U usporedbi s rezultatima hijerarhijskog grupiranja vidljivo je da nema puno razlika u grupama. Grupa 1 je potpuno ista kao i kod hijerarhijskog grupiranja, grupa 2 ima samo jednu razliku i

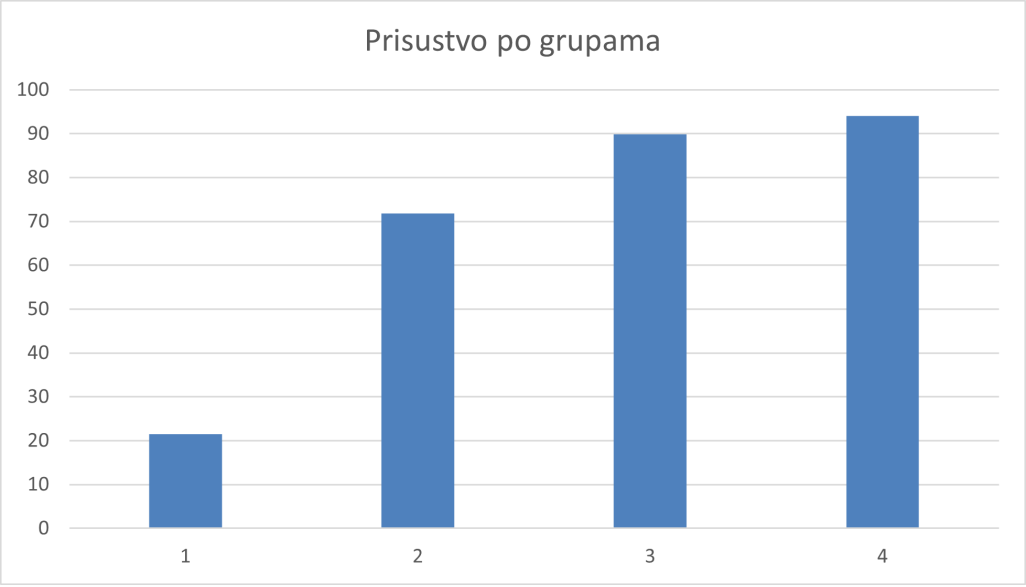


Slika 6.13. Raspored studenata po grupama korištenjem particijskog grupiranja

to je sa studentom ID 19, dok se grupa 3 i 4 ipak malo više razlikuju u usproredbi s grupama iz hijerarhijskog grupiranja. Sad ćemo detaljnije objasniti iste podatke koje smo objasnili i kod hijerarhijskog grupiranja te ćemo nakon toga usporediti rezulate da vidimo razlike između ova dva algoritma. Na kraju ćemo vidjeti koji bi nam algoritam bio korisniji za naš skup podataka.

6.5.1. Prisustvno na nastavi

Gledajući prisustvo na nastavi po grupama dobivamo graf koji je prikazan na slici 6.14.



Slika 6.14. Prisutnost studenata po grupama koristeći particijsko grupiranje

Grupa 1 je identična kao i kod hijerarhijskog grupiranja stoga nema promjene u prosječnom prisustvu koje iznosi 21.47% uz standardnu devijaciju od 14.48%.

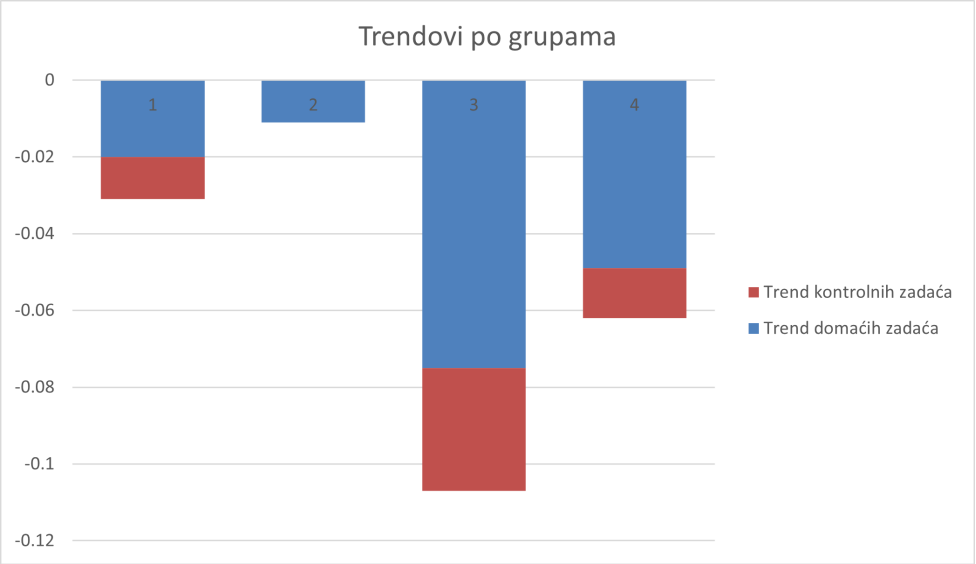
Grupa 2 u ovom slučaju ima prisustvo od 71.77% uz standardnu devijaciju od 10.1% što je nešto niže u usporedbi s hijerarhijskim grupiranjem ali i dalje u istom rangu odnosno nema veće promjene.

Grupa 3 u ovom slučaju ima prosječnu dolaznost od 89.88% uz standardnu devijaciju od 6.03%. Ono što valja napomenuti, a što je istko tako vidljivo iz grafa je da u ovom slučaju grupa 3 više nema bolju dolaznost od grupe 4 kao što je to bio slučaj kod hijerarhijskog grupiranja. Izgleda da je algoritam particijskog grupiranja rasporedio studente na način da je studente koji imaju bolju dolaznost stavio u grupu s najboljim studentima, odnosno gledajući grafički prikaz rasporedda, izgleda da je algoritam izbacio studente iz grupe 4 koji su ipak bili malo lošiji od ostalih studenata grupe te koji su izgleda rušili prosjek cijele grupe. Time je onda isto tako malo srušen prosjek grupe 3 što se tiče prisustva.

Grupa 4 je, kao što smo prije napomenuli, u ovom slučaju najbolja po dolaznosti koje iznosi 94% uz standardnu devijaciju od 6%. Grafički gledano, algoritam particijskog grupiranja je smjestio sve najbolje studente u jednu grupu i time zadržao njihovu odličnu dolaznost.

6.5.2. Trend rezultata kontrolnih i domaćih zadaća

Gledajući trend rezultata kontrolnih i domaćih zadaća dobivamo graf koji je prikazan na slici 6.15



Slika 6.15. Trendovi kontrolnih i domaćih zadaća po grupama koristeći particijsko grupiranje

Grupa 1 ima identične trendove kao i kod hijerarhijskog grupiranja i oni iznose -0.011 uz standardnu devijaciju od 0.035 za kontrolne zadaće i -0.02 uz standardnu devijaciju od 0.035 za domaće zadaće.

Grupa 2 ima trend koji iznosi -0.0004 uz standardnu devijaciju od 0.086 za kontrolne zadaće, to jest pozitivni i negativni trendovi studenata unutar grupe gotovo da se potpuno poništavaju dok

za domaće zadaće trend grupe 3 iznosi -0.011 uz standardnu devijaciju od 0.0536 i u usporedbi s hijerarhijskim grupiranjem negativni trend domaćih zadaća je nešto veći.

Grupa 3 ima trend koji iznosi -0.032 uz standardnu devijaciju od 0.1923 za kontrolne zadaće, a u usporedbi sa hijerarhijskim grupiranjem je nešto negativnija. Kao što smo i prije naveli, studenti koji su rušili trendove u grupi 4 kod hijerarhijskog grupiranja sad ruše trendove kod particijskog grupiranja za grupu 3. Trend domaćih zadaća za grupu 3 iznosi -0.075 uz standardnu devijaciju od 0.142 te je u usporedbi s hijerarhijskim grupiranjem nešto veći, to jest ovoga puta grupa 3 ima najveći negativni trend kako kontrolnih tako i domaćih zadaća. Sad tek vidimo koliko individualni studenti mogu doprinijeti potpuno drugačijoj slici.

Grupa 4 ovoga puta, dakle, nije najgora grupa što se tiče trendova te oni iznose -0.013 uz standardnu devijaciju od 0.045 za kontrolne zadaće odnosno -0.049 uz standardnu devijaciju od 0.041 za domaće zadaće.

Odmah je vidljivo da kod particijskog grupiranja niti jedna grupa nema niti jedan od trendova pozitivan, to jest svi su trendovi negativni. No, isto tako je vidljivo da su studenti koji su rušili prosjek u hijerarhijskog grupiranju za trendove grupe 4 ovdje prebačeni u grupu 3, što i ima smisla gledajući samo trendove. To jest, ima više smisla da se ti studenti koji ruše prosjek smjeste u nižu grupu, a ne da su smješteni u najbolju grupu jer zanemarivanjem jednog od aspekta kolegija ne zaslužuju biti u najboljoj grupi. Da ponovimo ono što smo rekli i kod hijerarhijskog grupiranja, to je ništa drugo nego posljedica odnosno kompromisa odabira broja grupa koji nije optimalan. Moramo negdje izgubiti na preciznosti.

6.5.3. Ukupni ostvareni bodovi iz kontrolnih zadaća

Gledajući ukupne ostvarene bodove iz kontrolnih zadaća dobivamo graf koji je prikazan na slici 6.16.

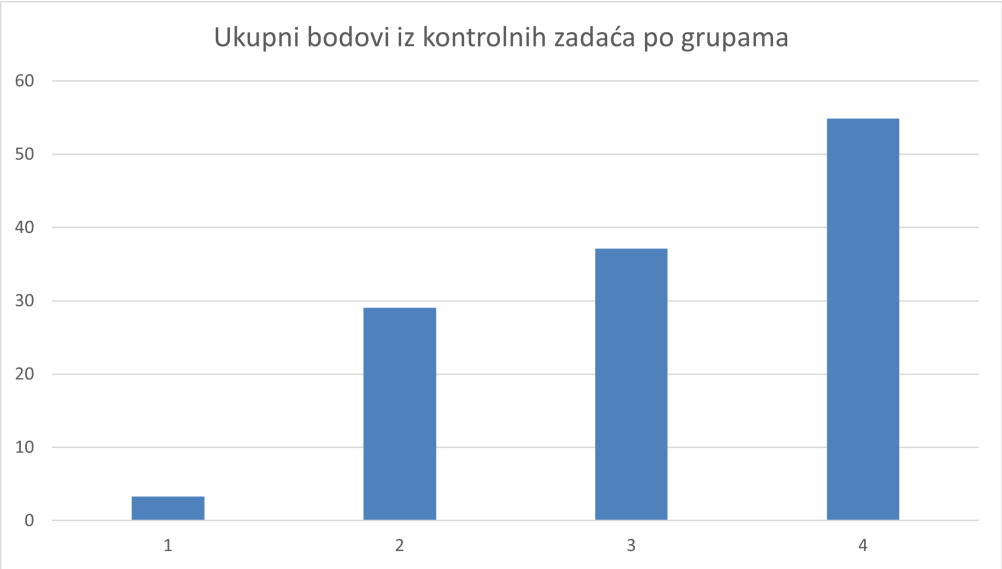
Na prvi pogled vidimo da je korelacija koju smo uočili za hijerarhijsko grupiranje ista kao i kod particijskog grupiranja.

Grupa 1 je ista kao i kod hijerarhijskog grupiranja i njeni studenti su u prosjeku ostvarili 3.33 od mogućih 60 bodova uz standardnu devijaciju od 4.93.

Grupa 2 je ostvarila u prosjeku 29.07 od mogućih 60 bodova uz standardnu devijaciju od 5.96. To je nešto manje od grupe 2 kod hijerarhijskog grupiranja ali neznatno, za samo 0.33 boda.

Grupa 3 je u prosjeku ostvarila 37.14 od mogućih 60 bodova uz standardnu devijaciju od 5.63 što je također nešto manje u usporedbi sa istom grupom kod hijerarhijskog grupiranja.

Grupa 4 je u ovom slučaju puno više odvojena od ostalih te je u prosjeku ostvarila 54.89 od mogućih 60 bodova uz standardnu devijaciju od 3.01 bod.



Slika 6.16. Ukupni ostvareni bodovi iz kontrolnih zadaća po grupama korištenjem particijskog grupiranja

6.5.4. Usporedba hijerarhijskog i particijskog grupiranja

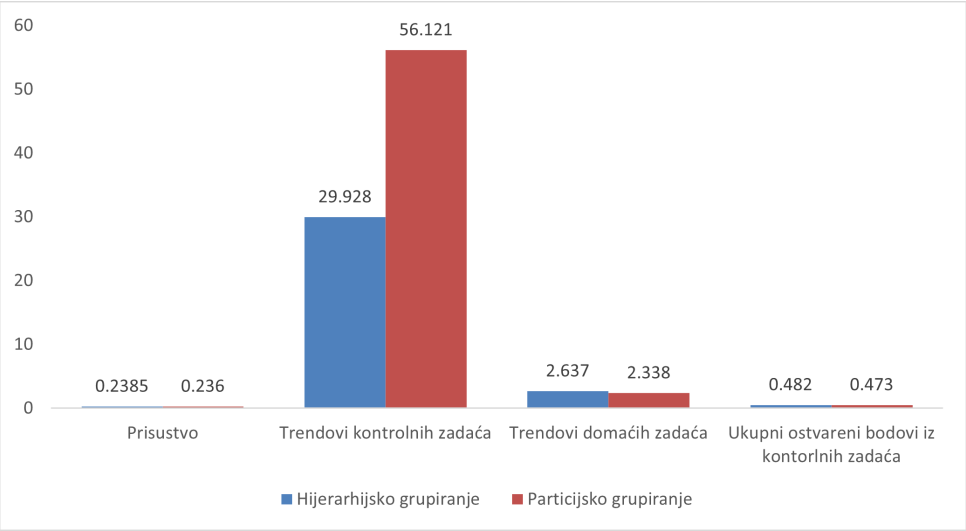
Dakle, vidjeli smo kakve grupe smo dobili particijskim grupiranjemi te kakve smo dobili sa hijerahijskim grupiranjem. Na nama je sad zapravo odluka koji ćemo koristiti. Kao što smo vidjeli u analizi, mala promjena rasporeda studenata po grupama može napraviti velike razlike u konačnim rezultatima te nam može iskriviti sliku određene grupe. No, isto tako znamo da, pogotovo za loš odabir broja grupa nikad ne možemo najbolje rasporediti studente bez obzira koji algoritam koristimo. Sad ćemo vidjeti taj kompromis gledajući koeficijent varijacije za podatke koje smo dobili. Točnije gledat ćemo rasipanje (za naš slučaj različitost studenata) unutar grupe.

Usporedbu koeficijenata varijacije[2] po podacima koje smo detaljnije analizirali možemo vidjeti na grafu prikazanom na slici 6.17.

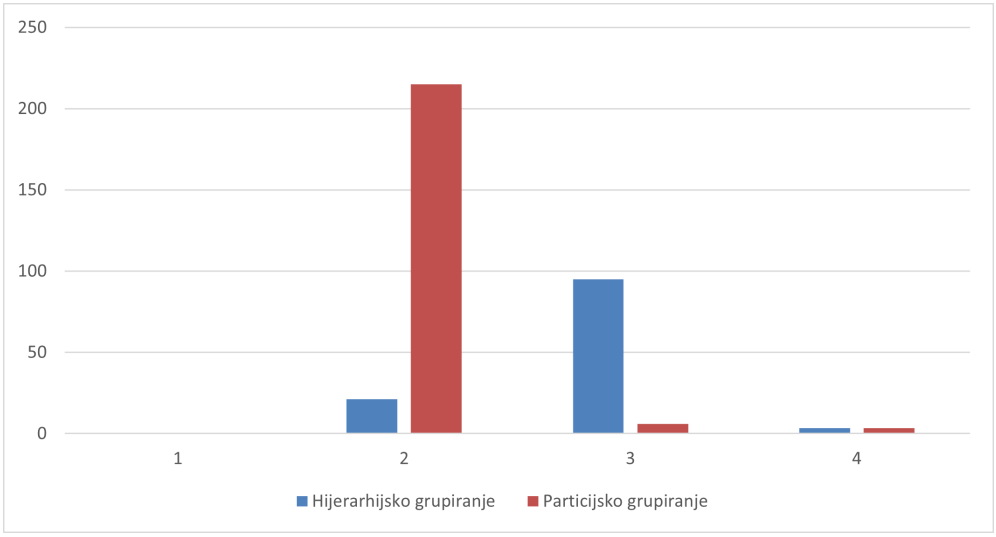
Za koeficijent varijacije nam je bitno da je što manji moguće. Vidimo da particijsko grupiranje ima puno veći koeficijent varijacije pogotovo za trendove kontrolnih zadaća dok su ostali koeficijent relativno slični. Da vidimo zašto je koeficjinet varijacije toliko velik kod trendova kontrolnih zadaća za i jedan i drugi algoritam moramo vijdeti kod koje je grupe došlo do tolikog povećanja i to možemo vidjeti na slici 6.18

Vidimo da su nam problematične grupe bile grupa 2 za particijsko grupiranje i grupa 3 za hijerarhijsko grupiranje. Tu također možemo izravno vidjeti utjecaj odabira optimalne metode za grupiranje, mi smo kod hijerarhijskog grupiranja koristili Wardovu metodu čiji je cilj zapravo smanjiti varijaciju unutar grupa svakim novim grupiranjem, dok se kod particijskog grupiranja može dogoditi situacija kao ovdje da se za naš odabir grupa krivo rasporede neki studenti i time znatno povećaju varijaciju grupe.

[2]Koeficijent varijacije je omjer varijacije i srednje vrijednosti te se koristi upravo u ovim slučajevima kad imamo podatke raznih mjera čije varijacije želimo međusobno usporediti



Slika 6.17. Usporedba koeficijenata varijacije za analizirane podatke



Slika 6.18. Usporedba koeficijenata varijacije hijerarhijskog i particijskog grupiranja po grupama za trendove kontorlnih zadaća

Particijsko grupiranje rasporedilo je studente na način da se više fokusiralo na ukupni ostvareni rezultat, pogotovo gledajući prvu i zadnju grupu. Gledajući varijacije vidi se veliki utjecaj Wardove metode kod hijerarhijskog grupiranja što mu uvelike daje prednost. U suštini, vidljivo je zapravo da nam je sveukupno bolje grupiranje odradio hijerarhijski algoritam koristeći Wardovu metodu. No, isto tako kod hijerarhijskog grupiranja smo imali izobličenje unutar grupe 4 što je jako pokvarilo vrijednosti te grupe. Prilagođavanje nastavi toj grupi ne bi bilo potpuno ispravno jer prosjek ne prikazuje dobro stanje prosječnog studenta. Kao što smo prije naveli i što ćemo ponoviti, jako je bitan odabir optimalnog broja grupa, odabirom optimalnog broja grupa prvo bi metoda bila znatno bolja. No, čak i s ovim odabirom hijerarhijsko grupiranje nama je dalo bolje rezultate. Tome pridonosi i činjenica, kao što smo i prije naveli, da rezultati particijskog grupiranja mogu varirati bez obzira što se skup podataka nije promijenio što bi značilo da niti jedan rezultat particijskog grupiranja ne mora biti potpuno objektivan i ispravan.

6.6. Zaključak grupiranja na obrazovnim podacima

Klasterska analiza na obrazovnim podacima dala nam je uvid u to kako se ponašanje studenata odrazi na njihove konačne uspjehe na kolegiju. Vidjeli smo koji nam podaci imaju visoku korelaciju sa konačnim uspjehom, a koji i nemaju baš toliku. Grupe studenata koje smo dobili bile su relativno jednostavne za opisati i sagledati. Sad imamo uvid u to kako koja grupa pristupa kolegiju i prema tome možemo prilagoditi naš pristup.

Analiza koju smo napravili može pomoći nastavnicima da uoče način na koji studenti pristupaju nastavi u svrhu prilagođavanja nastavnog procesa. Naime, pokazalo se iz analize da svaka od grupa ima svoj način učenja i pristupa kolegiju te time zahtjeva individualizaciju nastavnog procesa.

7. Primjena klasterske analize na podatke u inženjerstvu i elektrotehnici

U uvodu smo naveli dosta teorijskih primjera korištenja klasterske analize u elektrotehnici, no sad ćemo na ukratko proći i klastersku analizu na pravom primjeru u struci. Grupirat ćemo europske države prema potrošnji električne energije s naglaskom na obnovljive izvore energije. Podatke koje ćemo koristit su:

- Ukupna potrošena energija,
- Postotak obnovljivih izvora u ukupnoj potrošenoj enegiji,
- BDP uzevši u obzir i troškove života države[1],
- Korisnost potrošene energije[2].

Podaci koje ćemo koristiti preuzeti su iz [29] i [30].

7.1. Provedba klasterske analize

Kako smo vidjeli i kod analize obrazovnih podataka, bolji algoritam nam je hijerarhijski algoritam s Wardovom metodom pa ćemo i njega koristiti ovdje. Pošto se radi o hijerarhijskog algoritmu, broj grupa možemo odabrati nakon što skup podataka provedemo kroz algoritam.

Dakle, klasterskom analizom hijerarhijskog grupiranja pomoću Wardove metode i korištenjem Euklidske udaljesnosti dobili smo rezultate koji su prikazani na slici 7.1.

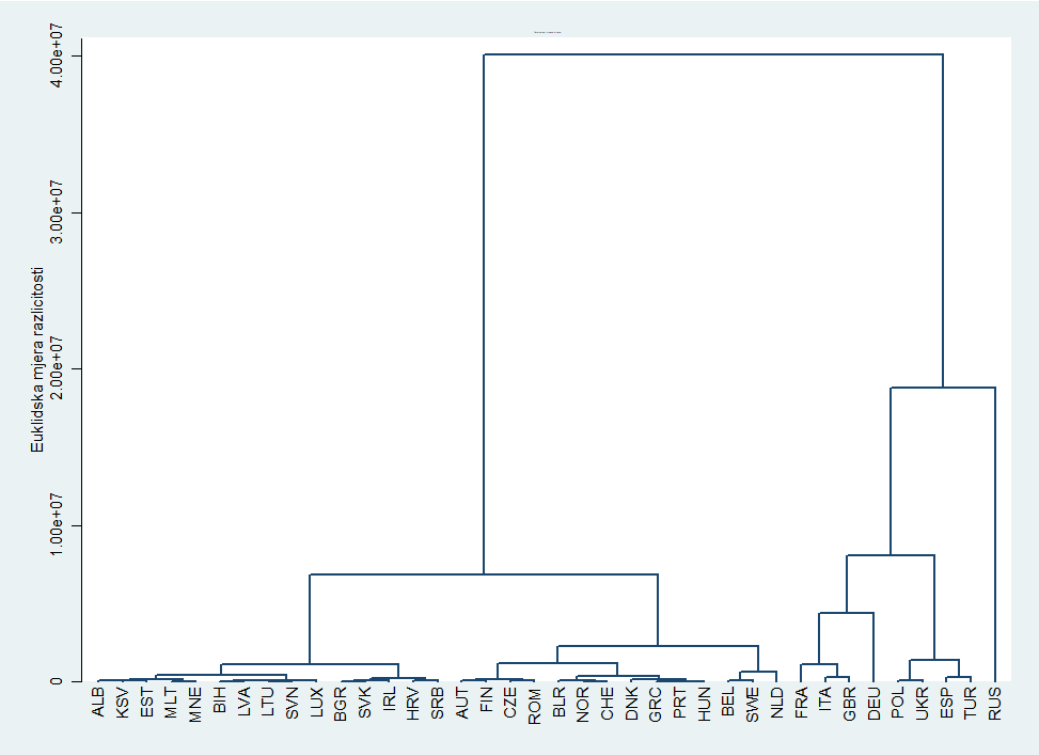
Iz grafa nam je vidljivo da imamo par opcija na raspolaganju što se tiče broja grupa. Najbolje nam se čini ili 3 ili 5 grupa. Sada ćemo provesti rezultate grupiranja krzo CH indeks da vidimo izdvaja li se neki od ponuđenih broja grupa kao najbolji. No, podsjetimo se, to nam je i dalje samo okvirna ocjena koliko bi grupa bilo dobro imati ali je na kraju ipak naš odabir koliko grupa bi htjeli opisati.

Provedbom grupa kroz CH indeks dobivamo rezultate koji su vidljivi u tablici 7.1.

Prema ovome vidimo da nam je poželjnije imati što je veći broj grupa moguće, najvjerojatnije zbog toga što su ipak države, barem prema ovim podacima, dosta različite. To jest, ako govorimo u smislu klasterske analize, elementi grupa su jako raspršeni.

[1]Formalno izraženo kao GDP, PPP te predsatvlja BDP države koji je prilagođen i za troškove života određene države te se smatra objektivnijim za uspoređivanje životnog standarda pošto uzima i troškove života u obzir koji su specifični za svaku državu i mogu uvelike izobličiti stvarnu sliku.

[2]Korisnost električne energije predstavlja koliko je električne energije (izražene u MJ) potrebno da država stvori jedno dobro. Dobro u ovom kontekstu predstavlja ekonomsku jedinicu koja se može prodati da bi se stekla financijska dobit.



Slika 7.1. Rezultirajući dendrogram analize država

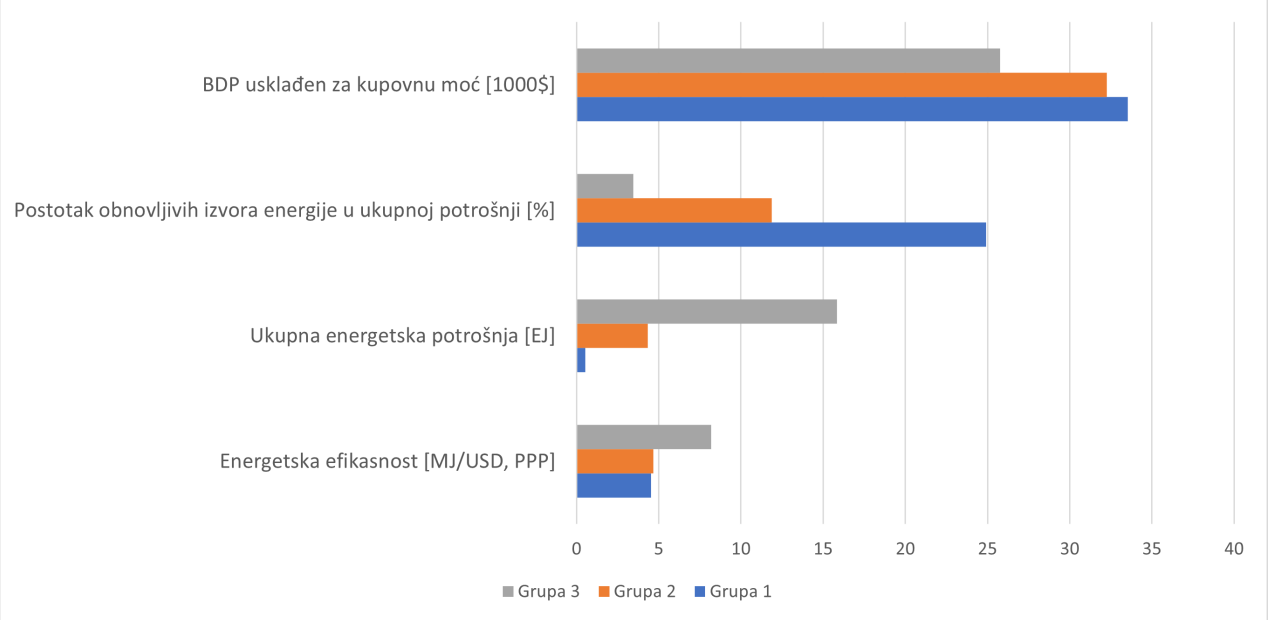
Tablica 7.1. Rezultati računanja Calinski-Harabasz indeksa za različit broj grupa

Broj grupa	Calinski-Harabasz indeks
2	42.45
3	164.41
4	251.18
5	252.80
6	692.95
7	895.60
8	1095.45
9	1127.97
10	1909.23
11	2480.50
12	4002.11
13	4276.31
14	4816.75
15	6218.74

Mi ćemo na kraju uzeti da nam je broj grupa 3, što prema ovome znači da će biti puno iskrivljenja, no prema CH indeksu izgleda da ćemo se morati ipak zadovoljiti sa iskrivljenom vrijednosti grupa.

7.1.1. Interpretacija klasterske analize

Ako sad uzmemo prosjeke ove 4 varijable koje smo koristili za klastersku analizu i prikažemo ih zajedno na graf, dobivamo graf koji je prikazan na slici 7.2.



Slika 7.2. Rezultirajući graf prosjeka podataka po grupama

Jasno je vidljiva podjela triju grupa iz grafa.

Grupa 1 ima relativno nisku energetsku efikasnost, relativno nisku ukupnu potrošnju, najveći udio obnovljivih izvora energije te relativno sličan BDP kao i druge dvije grupe. Unutar grupe 1 nalazi se jako puno država što znači zapravo da su države koje su se našle izvan te grupe po nečemu posebne. Sve te države koje se nalaze u grupi 1 pokušavaju se odmaknuti od neobnovljivih izvora energije te kako idemo udesno po dendrogramu tako se i povećava udio obnovljivih izvora u državama kao i razvijenost država. Albanija, Kosovo, Estonija, Malta i Crna Gora na jednom kraju i Belgija, Švedska i Nizozemska na drugom. No, iako su te dvije države dosta međusobno različite, i dalje su ostale grupe različitije od njih, što nam zapravo govori koliko su ostale grupe različite ali isto tako potvrđuje rezultate koje smo dobili računajući CH indeks, da bi veći broj grupa bio poželjniji.

Grupa 2, kao i grupa 1 ima relativno nisksu efikasnost, malo višu ukupnu potrošnju, manji udio obnovljivih izvora energije te sličan BPD kao i grupa 1. U grupi 2 s jedne strane nalaze se razvijenije, a s druge strane manje razvijene države. Sve ove razvijenije države koriste se neobnovljivim izvorima energije za svoju proizvodnju što može biti posljedica toga da se infrastruktura za obnovljive izvore u tim državama sporo gradi, odnosno ne prati porast proizvodnje. S druge strane, nerazvijenije države koje su u toj grupi vrlo vjerojatno jednostavno nemaju razvijenu industriju i infrastrukturu za obnovljive izvore ili jednostavno nemaju sredstava za takve projekte pošto se fokusiraju da poprave ekonomiju na koji god način je moguće bilo to obnovljivim ili neobnovljivim

izvorima.

Grupa 3, to jest jedina država koja se nalazi u grupi 3, Rusija, ima najvišu energetsku efikasnosti, puno veću ukupnu potrošnju i puno niži postotak obnovljivih izvora energije kao i niži BDP od ostalih država. Visoka potrošnja i nizak postotak obnovljivih izvora očekivani su za državu poput Rusije koja je država velike površine te je velikom većinom prekrivena tundrom, odnosno na velikom području Rusije su vremenski uvjeti jako loši. Uz to, veliki izvori plina daju malu zainteresiranost za projekte obnovljivih izvora energije.

7.2. Zaključak klasterske analize na podatke u inženjerstvu i elektrotehnici

Analiziranjem država prema potrošnji električne energije vidjeli smo kako se zapravo države dosta razlikuju po tome pitanju.

Rezultate koje smo dobili u teoriji bi mogli služiti kao smjernice za dovođenje novih odluka o načinu na koji se pristupa obnovljivim izvorima energije. Točnije, dali bi nam neki uvid u to koje bi se države na koji način trebale prilagoditi kako bi se smanjio njihov učinak na globalno zatopljenje.

Možemo vidjeti da se klasterska analiza pokazala kao izuzetno koristan alat za uočavanje neki obrazaca ponašanja koje gledajući ukupni skup nismo mogli primjetiti.

8. Zaključak

U ovom radu upoznali smo se sa klasterskom analizom. Objasnili smo njenu definiciju te vidjeli kako se ona provodi. Prikazali smo razne algoritme kojima možemo napraviti klastersku analizu te kako se ti algoritmi implementiraju. Opisali smo najčešće metrike klaterske analize te vidjeli kako se one provode. Vidjeli smo, također, i kako ocijeniti grupiranje, odnosno što grupu čini dobro ili loše grupiranom. Na kraju teorijskog dijela opisali smo kako biramo optimalan broj grupa, što smo vidjeli da nam je bitan korak klasterske analize.

Nakon što smo objasnili teoriju koja stoji iza klasterske analize, primjenili smo ju na obrazovnim podacima, točnije analizirali smo rezultate studenata iz kolegija Inženjerska matematika ET akademske godine 2021./2022. koji se provodio na Tehničkom fakultetu Sveučilišta u Rijeci. Analizirali smo podatke pomoću hijerarijskog grupiranja kako bi ustanovili postoje li neke grupe studenata s obzirom na njihov pristup kolegiju. Rezultate koje smo dobili smo potom i interpretirali na način da smo se fokusirali na potencijalne uzroke dobivenih rezultata. Nakon toga smo detaljno objasnili grupe te koji tip studenta bi koja grupa mogla sadržavati. Na posljetku analize obrazovnih podataka smo skup podataka analizirali particijskim grupiranjem te usporedili dobivene rezultate. Pokazali smo koja nam je metoda davala bolje rezultate i objasnili smo zašto.

Naš eksperiment pokazao je prednost hijerarhijskog grupiranja nad particijskim grupiranjem jer smo korištenjem Wardove metode smanjili varijaciju unutar grupe na minimum i pritom smanjili iskrivljenje vrijednosti pojedinih grupa. Doduše, pokazali smo i da će čak i takva analiza imati znatno iskrivljenje radi lošeg odabira broja grupa te smo tu zapravo vidjeli koliko nam je bitno odabrati optimalan broj grupa.

Na posljetku rada smo proveli klastersku analizu na primjeru iz struke i vidjeli praktičan primjer na koji se klasterska analiza može koristiti u elektrotehnici.

Na kraju svega možemo zaključiti da je klasterska analiza kao metoda vrlo korisna za analizu podataka kako bi iz skupa podataka izvukli neke zaključke koje ne bismo mogli zaključiti analizirajući skup bez prethodne analize. Klasterska analiza nam daje način da veliki skup podataka pojednostavimo i time olakšamo analizu dostupnih podataka.

Bibliografija

[1] Pearson, K.:"On lines and planes of closest fit to systems of points in space", Philosophical Magazine: 559-572, 1901.

[2] Driver, Harold E.; Kroeber, Alfred L.:"Quantitative Expression of Cultural Relationships": 211-256, Publikacije Sveučilišta Kalifornije iz američke arheologije i etnologije, SAD, 1932.

[3] Zubin, Joseph:"A technique for measuring like-mindedness": 508-516., 1938.

[4] Tryon, Robert C.:"Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality", Edwards Brothers, 1939.

[5] Johnson, S. C.:"Hierarchical clustering schemes": 241-254, Psychometrika, 1967.

[6] Kaufman, L.; Rousseeuw, P. J.: "Finding groups in data: An introduction to cluster analysis", John Wiley & Sons, 2009.

[7] "Electrical Peak Load Clustering Analysis Using K-Means Algorithm and Silhouette Coefficient", s Interneta, https://ieeexplore.ieee.org/document/9249773, 7.9.2023.

[8] "How to Calculate Hamming Distance in R?", s Interneta, https://www.geeksforgeeks.org/how-to-calculate-hamming-distance-in-r/, 7.9.2023.

[9] "Example of Euclidean and Manhattan distances between two points A and B", s Interneta, https://www.researchgate.net/figure/Example-of-Euclidean-and-Manhattan-distances-between-two-points-A-and-B-The-Euclidean_fig8_333430988, 7.9.2023.

[10] "How to Calculate Jaccard Similarity in R?", s Interneta, https://www.geeksforgeeks.org/how-to-calculate-jaccard-similarity-in-r/, 7.9.2023.

[11] "Cluster analysis", s Interneta, https://en.wikipedia.org/wiki/Cluster_analysis, 10.9.2022.

[12] "Grupiranje", s Interneta, https://hr.wikipedia.org/wiki/Grupiranje, 10.9.2022.

[13] Šnajder, J.:"Strojno učenje: 19. Grupiranje", UNIZG, FER, 2020.

[14] Šnajder, J.:"Strojno učenje: 20. Grupiranje II", UNIZG, FER, 2020.

[15] "9 Distance Measures in Data Science", s Interneta, https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa, 10.9.2022.

[16] "Different Types of Distances Used in Machine Learning", s Interneta, https://medium.com/swlh/different-types-of-distances-used-in-machine-learning-ec7087616442, 10.9.2022.

[17] "4 Basic Types of Cluster Analysis used in Data Analytics", s Interneta, https://www.youtube.com/watch?v=Se28XHI2_xE, 3.8.2022.

[18] "k-means clustering", s Interneta, https://en.wikipedia.org/wiki/K-means_clustering, 10.9.2022.

[19] "Voronoi diagram", s Interneta, https://en.wikipedia.org/wiki/Voronoi_diagram, 10.9.2022.

[20] "Vector quantization", s Interneta, https://en.wikipedia.org/wiki/Vector_quantization, 10.9.2022.

[21] "Euclidiean distance", s Interneta, https://en.wikipedia.org/wiki/Euclidean_distance, 10.9.2022.

[22] "Hierarichical clustering", s Interneta, https://en.wikipedia.org/wiki/Hierarchical_clustering, 10.9.2022.

[23] "DBSCAN", s Interneta, https://en.wikipedia.org/wiki/DBSCAN, 10.9.2022.

[24] "Expectation-maximization algorithm", s Interneta, https://en.wikipedia.org/wiki/Expectation–maximization_algorithm

[25] "Likelihood function", s Interneta, https://en.wikipedia.org/wiki/Likelihood_function

[26] "Ward's method", s Interneta, https://en.wikipedia.org/wiki/Ward%27s_method, 7.9.2023.

[27] "Calinski-Harabasz Index – Cluster Validity indices | Set 3", s Interneta, https://www.geeksforgeeks.org/calinski-harabasz-index-cluster-validity-indices-set-3/, 7.9.2023.

[28] "dendrogram", s Interneta, https://www.mathworks.com/help/stats/dendrogram.html, 9.9.2023.

[29] "World - Global Tracking Framework", s Interneta, https://energydata.info/dataset/world-global-tracking-framework-2017, 9.9.2023.

[30] "GDP per capita, PPP (current international $)", s Interneta, https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD, 9.9.2023.

Sažetak i ključne riječi

U ovom radu objašnjena je klasterska analiza i matematički aparat koji je potreban za njeno razumijevanje i provođenje. Definirane su različite metrike kojima se mjeri udaljenost između grupa te su navedeni temeljni principi formiranja grupa, s posebnim naglaskom na hijerearhijsko i particijsko grupiranje. Objašnjena je i problematika određivanja optimalnog broja grupa. U završnom dijelu rada napravljena je detaljna klasterska analiza na dva realna primjera. Jedan se primjer odnosio na klastersku analizu obrazovnih podataka, a drugi na primjenu klasterske analize u inženjerstvu.

Ključne riječi: Klasterska analiza, hijerarhijsko grupiranje, particijsko grupiranje, Wardova metoda, metrike udaljenosti

Summary and key words

This paper explains cluster analysis and the mathematical apparatus required for its understanding and implementation. Different metrics that measure the distance between groups are defined and the basic principles of group formation are stated, with special emphasis on hierarchical and partition clustering. The issue of determining the optimal number of groups is also explained. In the final part of the paper, a detailed cluster analysis was made on two real examples. One example related to cluster analysis of educational data, and the other to the application of cluster analysis in engineering.

Keywords: Cluster analysis, hierarchical clustering, centroid-based clustering, Ward's method, distance metrics.