Grupiranje srednjoškolaca i studenata prema njihovim interesima, fobijama i navikama klasterskom analizom

Vlaić, Petra

Grupiranje srednjoškolaca i studenata prema njihovim interesima, fobijama i navikama klasterskom analizom

Vlaić, Petra

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Petra Vlaić

GRUPIRANJE SREDNJOŠKOLACA I STUDENATA PREMA NJIHOVIM INTERESIMA, FOBIJAMA I NAVIKAMA KLASTERSKOM ANALIZOM

Diplomski rad

Voditelj rada:
prof. dr. sc. Anamarija Jazbec

Zagreb, 2020.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Za sve trenutke podrške i utjehe, a bilo ih je, hvala mojim roditeljima.
Za sve trenutke bodrenja i razveseljavanja, a bilo ih je, hvala mojim prijateljima.
Za sve savjete i pomoć, a bilo ih je, hvala prof. dr. sc. Anamariji Jazbec.
Bez vas sve bi ovo bilo puno teže.

Sadržaj

Uvod

Klasterska analiza metoda je obrade podataka pri čemu se podatci grupiraju u klastere. Svaki klaster predstavlja grupu objekata čiji elementi imaju najveću moguću sličnost. Grupiranje objekata je učestalo u prirodi, tako o klasterima možemo razmišljati kao o skupinama djece koja treniraju razne sportove ili o biljnim vrstama kojima odgovaraju slični klimatski uvjeti.

Podatci koji se mogu klasterirati su različiti i sukladno tome bira se odgovarajući algoritam i mjera sličnosti. Primjerice, podatci mogu biti numeričke vrijednosti kao što su koncentracije elemenata u tlu, plaće radnika u tvrtkama ili prosjek ocjena studenata na fakultetima, ali i kategorijske, npr. krvne grupe, pripadnost religiji ili bračno stanje ispitanika. Očito je da će se način klasteriranja razlikovati ovisno o podatcima koji su korišteni, odnosno o potrebama i konačnim ciljevima istraživanja.

Metoda pronalazi svoju primjenu u brojnim sferama ljudske djelatnosti. U ovom radu klasterska analiza primijenit će se na skupu podataka o interesima, fobijama i zdravstvenim navikama studenata i srednjoškolaca. Kao mjera sličnosti koristit će se udaljenost između elemenata, a od algoritama provest će se hijerarhijsko te k-means klasteriranje.

Poglavlje 1

Klasterska analiza

1.1 Opis metode

Klaterska analiza je metoda kojom se za određeni skup podataka S određuje grupiranje u podskupove $C_i, i \in \mathbb{N}$ prema njihovoj sličnosti. Podskupovi C_i dobiveni klasterskom analizom nazivaju se **klasteri**, a ovisno o algoritmu i ciljevima istraživanja, konačan broj klastera može i ne mora biti unaprijed definiran. Što su klasteri bolje separirani, to je metoda učinkovitija. Skup svih klastera naziva se **klastering**, u oznaci **C**. Prema [8], razlikujemo hijerarhijski i particijski klastering.

Definicija 1.1.1. *Neka je* $k \in \mathbb{N}$. *Klastering* $\boldsymbol{C} = \{C_1, C_2, \ldots, C_k\}$ *je* ***hijerarhija*** *ako* $\forall\, C_i, C_j \in \boldsymbol{C}, 1 \leq i, j \leq k$ *vrijedi*

$$C_i \cap C_j \in \{\emptyset, C_i, C_j\}. \tag{1.1}$$

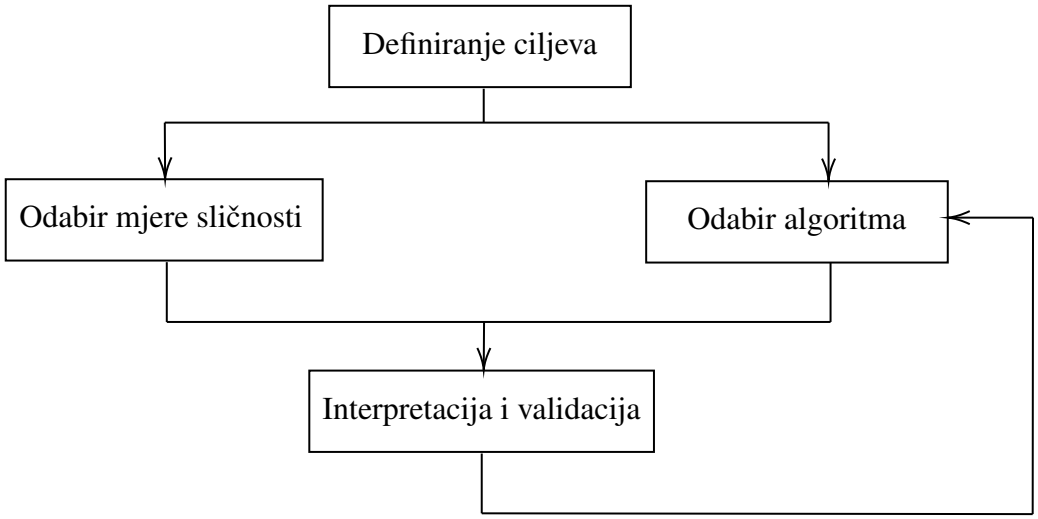
Definicija 1.1.2. *Neka je* $k \in \mathbb{N}$. *Klastering* $\boldsymbol{C} = \{C_1, C_2, \ldots, C_k\}$ *je* ***particija*** *ako vrijedi*

$$\bigcup_i C_i = \boldsymbol{C}, \tag{1.2}$$

$$i \neq j \Rightarrow C_i \cap C_j = \emptyset. \tag{1.3}$$

1.1.1 Koraci metode

Prema [3], dan je shematski prikaz koraka metode. Ne postoji jedinstveni postupak koji bi definirao klastersku analizu, ali generalni koraci su sljedeći:



Slika 1.1: Shematski prikaz

- **Definiranje ciljeva**
  Definiraju se ciljevi analize i formuliraju početne pretpostavke.
- **Prikupljanje podataka**
  Odabire se skup podataka S i mjere se relevantni podatci prema svojim atributima.
- **Početna analiza**
  Nakon prikupljanja podatci se pregledavaju, eliminiraju se neispravne vrijednosti te ukoliko je potrebno, vrši se standardizacija varijabli.
- **Odabir mjere sličnosti**
  Prema karakteristikama podataka odabire se odgovarajuća mjera sličnosti za dani problem.
- **Odabir algoritma**
  Odabire se odgovarajući algoritam klasteriranja i definiraju se njegovi parametri.
- **Validacija**
  Rezultati analize se ocjenjuju i određuje se klastering koji optimizira početni problem.
- **Interpretacija**
  Rezultati se uspoređuju s drugim studijama kako bi se izvukli konačni zaključci i predložila eventualna daljnja analiza.

Klaterska analiza je jedna od novijih metoda obrade podataka i produkt je potreba i suradnje različitih područja znanosti kao što su statistika, računarstvo, operacijska istraživanja itd. U biologiji, klasterska analiza se još naziva i taksonomija, a u domeni strojnog učenja spada u nenadzirane metode učenja. Primjenom klasterske analize nije unaprijed poznato na koji način će se objekti grupirati u klastere.

1.2 Mjere sličnosti

Podatci se grupiraju u klastere ovisno o vrijednostima mjere sličnosti. Dakle, za provedbu klasterske analize nužno je odabrati adekvatnu mjeru. Neka su $x, y \in S$. Prema [8], sličnost d je funkcija koja svakom paru iz S pridružuje realnu vrijednost, tj. $d : (x, y) \to \mathbb{R}$. Osnovna svojstva koja mjera sličnosti mora zadovoljavati su sljedeća:

- $d(x, y) \geq 0$ (nenegativnost),
- $d(x, x) = 0$,
- $d(x, y) = d(y, x)$ (simetričnost).

Ako mjera sličnosti dodatno zadovoljava i:

- $d(x, y) = 0 \Leftrightarrow x = y$,
- $(\forall z \in S)\ d(x, y) \leq d(x, z) + d(z, y)$ (nejednakost trokuta),

kažemo da je *metrika*.

Ovisno o vrsti podataka koji su prikupljeni, razlikovat će se odabir mjere sličnosti. Neka su $x, y \in S$ oblika $x = (x_1, x_2, \ldots, x_n)$ i $y = (y_1, y_2, \ldots, y_n)$, $n \in \mathbb{N}$. $\forall i \in \{1, 2, \ldots, n\}$ varijable x_i, y_i su atributi objekata x, y. Određivanje udaljenosti između objekata skupa S svodi se na određivanje udaljenosti između njihovih atributa (varijabli). Dakle, potrebno je definirati koje vrijednosti atributi mogu poprimiti kako bi se izabrala odgovarajuća mjera sličnosti.

1.2.1 Numeričke varijable

Numeričke varijable poprimaju vrijednosti iz skupa realnih brojeva $\mathbb{R}$. Neka je x numerička varijabla i neka je D skup vrijednosti koje može poprimiti. Ako je D konačan ili beskonačno prebrojiv, kažemo da je varijabla *x diskretna*. Primjerice, diskretne varijable su broj stanova u zgradi, broj sportaša na natjecanju, broj životinjskih vrsta na nekom području i sl.

Ako je D beskonačno neprebrojiv, kažemo da je varijabla x *neprekidna* ili *kontinuirana*. Tada varijabla poprima vrijednost iz određenog intervala realnih brojeva ili iz cijelog $\mathbb{R}$. Primjeri kontinuiranih varijabli su temperatura mora, koncentracija tvari u tlu, visina učenika itd.

Neka su $x, y \in S$, $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n), n \in \mathbb{N}$. Pretpostavimo da su $\forall i \in \{1, 2, \ldots, n\}$ x_i, y_i numeričke varijable. Prema [2] i [9], udaljenost $d : (x, y) \to \mathbb{R}$ moguće je izračunati koristeći sljedeće mjere:

- **Euklidska udaljenost**

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}, \tag{1.4}$$

- **Minkowski udaljenost**

$$d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^q \right)^{\frac{1}{q}}, \; q > 0, \tag{1.5}$$

- **Manhattan udaljenost**

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|, \tag{1.6}$$

- **Čebiševljeva udaljenost**

$$d(x, y) = max_{i=1}^{n} |x_i - y_i|, \tag{1.7}$$

- **Kosinus sličnost**

$$d(x, y) = \cos \alpha = \frac{x^\tau y}{||x||\,||y||}, \tag{1.8}$$

  pri čemu je α kut kojeg razapinju vektori.

- **Mahalanobis udaljenost**

$$d(x, y) = (x - y)^{'} \Sigma^{-1} (x - y), \tag{1.9}$$

  pri čemu je $\Sigma = \mathbb{E}[(X - \mu)(X - \mu)^{'}]$ kovarijacijska matrica, X vektor podataka te μ vektor očekivanja.

Uočimo, za $q = 2$ Minkowski i Euklidska udaljenost su jednake, za $q = 1$ Minkowski odgovara Manhattan udaljenosti, a za $q \to +\infty$ Čebiševljeva udaljenost jednaka je Minkowski udaljenosti.

Napomena 1.2.1. *Svaka metrika je ujedno i mjera. Obrat ne mora vrijediti.*

Vrijednosti koje se dobiju računanjem udaljenosti mogu varirati ovisno o mjernoj jedinici podataka, stoga je u nekim slučajevima za numeričke varijable potrebno prethodno napraviti standardizaciju, odnosno normalizaciju skupa podataka.

Neka su $x^{(1)}, x^{(2)}, \ldots, x^{(n)} \in S$ i neka je $1 \leq i \leq k$ proizvoljan. Prema [5], standardizacija se vrši u dva koraka:

1. Izračunati srednje odstupanje s_i od aritmetičke sredine m_i, $\forall i$:

$$s_i = \frac{1}{n} \sum_{j=1}^{n} |x_i^{(j)} - m_i|, \tag{1.10}$$

   pri čemu su $x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(n)}$ vrijednosti i-tog atributa.

2. Izračunati *z-score*:

$$z_i^{(j)} = \frac{x_i^{(j)} - m_i}{s_i}, 1 \leq j \leq n. \tag{1.11}$$

Metoda se dalje primjenjuje na standardiziranom skupu podataka $\{z^{(1)}, z^{(2)}, \ldots, z^{(n)}\}$.

1.2.2 Binarne varijable

Binarne (dihotomne) varijable mogu poprimiti točno dva stanja koja su međusobno isključiva. Primjerice, svjetlo može biti ili upaljeno ili ugašeno, rezultat je ili ispravan ili neispravan, bacanjem kovanice dobiva se ili pismo ili glava itd. Svako od ovih svojstava možemo označiti s 1 ili s 0, zato se binarne varijable još nazivaju i *indikatorske*.

Neka su (x, y) binarne. Označimo s α broj parova $(0, 0)$, β broj parova $(0, 1)$, γ broj parova $(1, 0)$ te neka je δ broj parova $(1, 1)$. Tada distribuciju para (x, y) možemo prikazati koristeći tablicu frekvencija:

Tablica 1.1: Tablica frekvencija

(x, y)	x:0	x:1	$\sum$
y:0	α	β	$\alpha + \beta$
y:1	γ	δ	$\gamma + \delta$
$\sum$	$\alpha + \gamma$	$\beta + \delta$	n

gdje je $n = \alpha + \beta + \gamma + \delta$ ukupna frekvencija.

Prema [6], sličnost je sada moguće izračunati koristeći neku od sljedećih mjera:

- **Simple Matching koeficijent**

$$d(x,y) = \frac{\alpha + \delta}{n}, \tag{1.12}$$

koji računa omjer simetričnih i ukupnih slučajeva.

- **Jaccard koeficijent**

$$d(x,y) = \frac{\delta}{\beta + \gamma + \delta}, \tag{1.13}$$

koji ne uzima u obzir slučajeve $(0,0)$.

- **Russel-Rao koeficijent**

$$d(x,y) = \frac{\delta}{n}, \tag{1.14}$$

koji računa omjer $(1,1)$ i ukupnih slučajeva.

- **Sorensen koeficijent**

$$d(x,y) = \frac{2\delta}{2\delta + \beta + \gamma}, \tag{1.15}$$

koji veću težinu stavlja na slučaj $(1,1)$.

- **1. Kulczynski koeficijent**

$$d(x,y) = \frac{\delta}{\beta + \gamma}, \tag{1.16}$$

koji ne uzima u obzir slučajeve $(0,0)$.

- **2. Kulczynski koeficijent**

$$d(x,y) = \frac{1}{2}\left(\frac{\delta}{\delta + \beta} + \frac{\delta}{\delta + \gamma}\right), \tag{1.17}$$

koji uzima aritmetičku sredinu rasparenih slučajeva.

Pretpostavka mjere je sljedeća: ako je karakteristika prisutna u jednoj varijabli, tada je prisutna i u drugoj.

- **Braun-Blanquet koeficijent**

$$d(x,y) = \begin{cases} \delta/(\delta + \gamma), & (\delta + \gamma) > (\delta + \beta) \\ \delta/(\delta + \beta), & \text{inače} \end{cases}. \tag{1.18}$$

- **Kocher i Wong koeficijent**

$$d(x,y) = \frac{\delta n}{(\delta + \gamma)(\alpha + \beta)}. \tag{1.19}$$

Napomena 1.2.2. *Jaccard i Sorensen koeficijenti su ekvivalentni.*

1.2.3 Kategorijske varijable

Vrijednost kategorijskih (opisnih) varijabli određena je pripadanjem u kategorije. Razlikujemo *ordinalne* i *nominalne* varijable.

Karakteristika ordinalnih varijabli jest da se među kategorijama može uspostaviti prirodni poredak, odnosno između kategorija postoji neki uređaj ($<, >$). Kao primjer, moguće je promatrati kategorije ocjena u školi (nedovoljan, dovoljan, dobar, vrlo dobar, izvrstan) ili stupanj stručne spreme (NKV, KV, VKV, SSS, VŠS, VSS). Vrijednost nominalnih varijabli također se može svrstati u kategorije, međutim na tim kategorijama ne postoji uređaj. Primjeri nominalnih varijabli su boja kose (crna, smeđa, plava, ...), bračni status (slobodan, u braku, razveden, ...), regionalna pripadnost i sl.

Neka su x, y objekti čiju sličnost želimo odrediti. Prema [5], ako su atributi objekata nominalni, sličnost između objekata jednaka je:

$$d(x, y) = \frac{p - m}{p}, \tag{1.20}$$

pri čemu je p ukupan broj atributa, a m broj poklapanja.

Neka su atributi objekata ordinalni. Neka je i proizvoljan ordinalni atribut, a S_i broj stanja koja može poprimiti. Budući da na ordinalnim varijablama postoji uređaj, stanja S_i su uređena i moguće je supstituirati vrijednosti rangom $r_i \in \{1, 2, \ldots, S_i\}$. Atributi ne moraju imati istu veličinu domene pa je potrebno skaliranje na vrijednosti iz intervala $[0, 1]$. Definiramo:

$$z_i^{(j)} = \frac{r_i^{(j)} - 1}{S_i - 1}. \tag{1.21}$$

Sada se sličnost računa koristeći neku od mjera za numeričke varijable primijenjenu na $z_i^{(j)}$.

1.2.4 Univerzalne mjere

U nekim slučajevima, klasterska analiza primjenjuje se na različitim vrstama podataka. Neka su $x_1, x_2, \ldots, x_n$ prikupljeni podatci. Tada je prikladna mjera sličnosti *Gower metoda*:

$$d(x_i, x_j) = \frac{1}{n} \sum_{k=1}^{n} s_{ij}^k, \tag{1.22}$$

pri čemu se udaljenost računa kao aritmetička sredina sličnosti varijabli podataka x_i i x_j. Prema [8], Gower metoda ne računa sličnost na isti način za sve varijable.

- Ako su varijable numeričke ili ordinalne, tada je

$$s_{ij}^k = 1 - \frac{|x_{ik} - x_{jk}|}{r_k}, \tag{1.23}$$

  pri čemu je $r_k = \max(x_{.k}) - \min(x_{.k})$.

- Ako su varijable nominalne ili binarne, tada je

$$s_{ij}^k = \begin{cases} 0, & x_{ik} \neq x_{jk} \\ 1, & x_{ik} = x_{jk} \end{cases}. \tag{1.24}$$

1.3 Odabir algoritma

Kada je odabrana odgovarajuća mjera sličnosti, potrebno je odabrati i algoritam koji će se primijeniti. Prema [5], dobar algoritam trebao bi imati sljedeća svojstva:

- skalabilnost,
- sposobnost analize različitih tipova varijabli,
- nepristranost pri određivanju oblika klastera,
- dobro podnošenje šumova,
- minimalni zahtjevi na početnim parametrima,
- neovisnost o redoslijedu unosa podataka,
- lakoća interpretabilnosti i iskoristivost rezultata.

Generalno, nije moguće definirati optimalan algoritam budući da izbor ovisi o skupu podataka, ciljevima istraživanja i karakteristikama algoritma. Algoritmi se razlikuju ujedno i prema pretpostavkama koje moraju biti zadovoljene pa svaki od algoritama klasteriranja ima svoje mane i prednosti. Osnovna podjela algoritama klasteriranja je na hijerarhijske i nehijerarhijske (partitivne).

1.3.1 Hijerarhijski algoritmi

Karakteristika hijerarhijskih algoritama klasteriranja je kreiranje hijerarhijske dekompozicije skupa objekata na temelju odabrane mjere sličnosti. Prema [4], razlikujemo *aglomerativne* i *razdjeljujuće* hijerarhijske algoritme. Za aglomerativne algoritme svaki objekt je zaseban klaster. Novi klasteri nastaju spajanjem postojećih u parove sve dok svi klasteri nisu spojeni u jedan. S druge strane, razdjeljujući algoritam gleda na cjelokupni skup podataka kao na jedan klaster koji se separira na manje (engl. *divide and conquer* algoritam). Upravo zbog ovako definiranog izbora klastera, konačan broj klastera ne mora biti unaprijed definiran.

Neka je $S = \{x_1, x_2, \ldots, x_n\}, n \in \mathbb{N}$ skup podataka. Koristeći mjeru sličnosti $d : S \to \mathbb{R}$ računamo matricu udaljenosti D:

$$D = \begin{bmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{bmatrix} \tag{1.25}$$

pri čemu je $d_{ij} = d(x_i, x_j), 1 \leq i, j \leq n$. Uočimo, $d_{ij} = d_{ji}, \forall i, j$.

- Koraci aglomerativnog algoritma su sljedeći:

  1. Svaki podatak zaseban je klaster: $C = \{X_1, X_2, \ldots, X_n\}, X_i = \{x_i\}, \forall i$.
  2. Koristeći matricu udaljenosti, odrediti najbliži par klastera (X, Y):

  $$d(X, Y) = \min_{i,j} d(X_i, X_j) = \min_{i,j} d_{ij}.$$

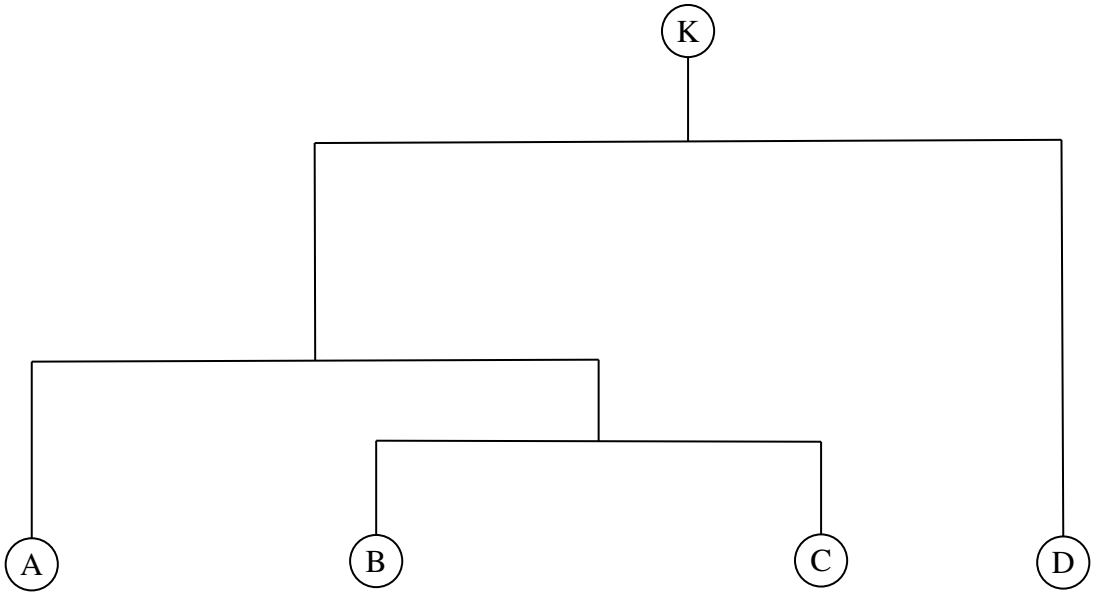
  3. Definirati novi klaster $Z = X \cup Y$.
  4. $X := Z, Y \in C \setminus S$
  5. Ponavljati postupak sve dok postoje barem dva klastera.

- Koraci razdjeljujućeg algoritma su sljedeći:

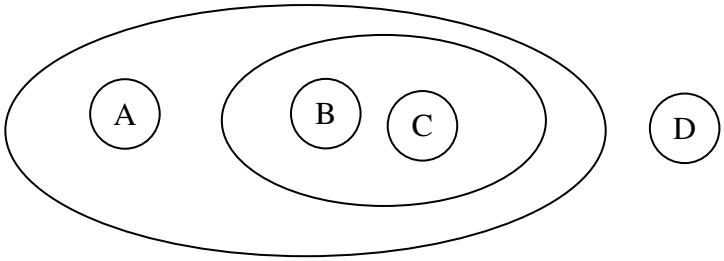
  1. Svi podatci čine jedan klaster: $C = \{x_1, x_2, \ldots, x_n\}$.
  2. Koristeći partitivni algoritam, separirati C na dva klastera: $C = X \cup Y$.
  3. Koristeći partitivni algoritam, separirati novodobivene X i Y na po dva klastera.
  4. Ponavljati postupak sve dok svaki podatak nije u zasebnom klasteru.

Hijerarhija klastera dobivena ovim algoritmima prikazuje se stablom koje se naziva *dendrogram*. Elementi dendrograma su: korijen, čvorovi, grane (unutarnja, vanjska) i listovi. Korijen predstavlja skup svih podataka, a svaki podatak je list. Listovi su na najnižoj razini dendrograma, dok je korijen na najvišoj. Grane dendrograma povezane su čvorovima. Dakle, sličnost (udaljenost) objekata je zapravo visina najnižeg unutarnjeg čvora kojeg dijele. Na vanjskim granama su outlieri, odnosno ekstremne vrijednosti.

Na slici 1.2 dan je primjer jednostavnog dendrograma skupa $\{A, B, C, D\}$, pri čemu K označava korijen dendrograma. Objekt D vanjskom je granom povezan s korijenom - to je outlier i pripada zasebnom klasteru, dok su B i C elementi istog klastera. Objekt A unutarnjom je granom povezan s klasterom kojeg sačinjavaju B i C. Dani dendrogram odgovara skupu koji je prikazan na slici 1.3. Ovisno o razini čvorova koju promatramo, moguće je dobiti finiju podjelu na klastere. Primjerice, kada bi se dendrogram „odrezao" na razini najnižeg čvora, dobiju se klasteri $\{A\}, \{B, C\}$ i $\{D\}$.



Slika 1.2: Jednostavan dendrogram



Slika 1.3: Pripadni skup

Obje vrste hijerarhijskog algoritma kao korak imaju računanje udaljenosti između klastera. Neka su C_1, C_2 proizvoljni klasteri. Prema [8], njihovu udaljenost moguće je odrediti upotrebom neke od sljedećih metoda:

- **Metoda minimuma**

$$d(C_1, C_2) = \min_{x,y}\{d(x, y) \mid x \in C_1, y \in C_2\} \tag{1.26}$$

  Udaljenost između klastera je najmanja udaljenost između elemenata klastera.

- **Metoda maksimuma**

$$d(C_1, C_2) = \max_{x,y}\{d(x, y) \mid x \in C_1, y \in C_2\} \tag{1.27}$$

  Udaljenost između klastera je najveća udaljenost između elemenata klastera.

- **Metoda prosjeka**

$$d(C_1, C_2) = \frac{1}{n_{C_1} n_{C_2}} \sum_{x \in C_1} \sum_{y \in C_2} d(x, y), \tag{1.28}$$

  pri čemu je n_{C_1} broj elemenata klastera C_1, a n_{C_2} broj elemenata klastera C_2.
  Udaljenost između klastera je aritmetička sredina udaljenosti između svakog para elemenata prvog i drugog klastera.

- **Ward metoda**

$$d(C_1, C_2) = \frac{n_1 n_2}{n_1 + n_2} d^2(t_1, t_2), \tag{1.29}$$

  gdje su t_1, t_2 centri klastera C_1, C_2.

Metodom minimuma dobiva se udaljenost između najsličnijih elemenata, a metodom maksimuma udaljenost između najmanje sličnih elemenata klastera. Wardovom metodom klasteri su često eliptični, a metoda prosjeka je najčešće korištena metoda.

1.3.2 Nehijerarhijski algoritmi

Nehijerarhijski algoritmi klasteriranja particioniraju skup podataka S na međusobno disjunktne klastere pri čemu je konačan broj klastera unaprijed definiran. Neka je $card(S) = n \in \mathbb{N}$. Particijsko klasteriranje dijeli skup podataka na $k \leq n$ klastera sa svojstvima da svaki klaster sadrži barem jedan element i svaki element pripada točno jednom klasteru. Iznimka je tkz. „fuzzy" klasteriranje za koje vrijedi da jedan element može biti dio više klastera. Najpoznatiji primjeri partitivnih algoritama su: k-means, PAM, CLARA i CLARANS.

Cilj k-means algoritma je odrediti particiju n-članog skupa na k-člani tako da je udaljenost između članova grupe (klastera) i njenog predstavnika minimalna. Predstavnik tj. centar klastera naziva se *centroid*. Postupak je opisan pomoću pseudoalgoritma 1:

Algorithm 1 K-Means

procedure K-MEANS(n)
1. Odrediti broj klastera k.
2. Slučajnim odabirom definirati centroide $\{c_1, c_2, \ldots, c_k\}$.

repeat
3. Centroidima pridružiti najbliže objekte.
4. Izračunati aritmetičke sredine objekata istog klastera.
5. Dobivene aritmetičke sredine novi su centroidi.

until Centroidi su nepromijenjeni.
return $\{c_1, c_2, \ldots, c_k\}$ ▷ $\{c_1, c_2, \ldots, c_k\}$ su optimalni.
end procedure

K-means algoritam optimizira prosječnu udaljenost objekata unutar istog klastera, odnosno kreira k klastera tako da je suma kvadrata pogrešaka (SSE) minimalna.

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \|x_{ij} - \mu_i\|^2 \tag{1.30}$$

Prednosti k-means algoritma su jednostavnost implementacije, skalabilnost te nepristranost pri određivanju oblika klastera. Unatoč tome što je konvergencija zagarantirana, mana algoritma je unaprijed definiran broj klastera te osjetljivost na šumove. Dodatno, zbog inicijalnog slučajnog odabira centroida, moguće je dobiti rezultat koji je samo lokalno optimalan.

U članku [5], navedena su neka unaprijeđenja u odnosu na k-means algoritam. PAM algoritam je nadogradnja na k-means s ciljem da se postigne manja osjetljivost na outliere. Umjesto da se definira centar klastera, svaki klaster predstavljen je svojim *medoidom* - objektom koji je „najcentralnije" postavljen unutar klastera. Mana ovakvog algoritma je

veća složenost za velike n i k. CLARA algoritam je zapravo PAM algoritam primijenjen na manjim uzorcima. Kao konačan rezultat uzima se onaj najbolji (najoptimalniji) među uzorcima. Mana ovakvog pristupa je mogućnost da je dobiveno rješenje samo lokalno optimalno. CLARANS algoritam vrši i provjeru optimalnosti izbora medoida, međutim upravo zbog toga nije prikladan izbor za velike količine podataka. Učinkovitost svakog od algoritama dodatno ovisi o postojanju pristranosti (engl. *bias*).

1.4 Primjena metode

Kao što je objašnjeno u [1], klastersku analizu moguće je primijeniti u raznim granama znanosti kao što su medicina, ekonomija, računarstvo, sociologija i sl.

1. Biologija, botanika, mikrobiologija, zoologija i slične grane znanosti koriste klastersku analizu kao sredstvo razvitka taksonomije, odnosno za definiranje raznih biljnih i životinjskih vrsta i podvrsta.

2. U medicini, objekti klasteriranja mogu biti npr. bolesti, simptomi ili pacijenti. Cilj analize je dobiti učinkovitije metode dijagnostike pacijenata kako bi se poboljšao njihov tretman te ubrzao oporavak.

3. Društvene znanosti poput psihologije, socioligije, antropologije pa čak i kriminalistike daju brojna polja pogodna za klastersku analizu. Metode učenja, faktori ljudske djelatnosti, obitelji, susjedstva, društvene organizacije, prekršaji i zločini, kulture, jezici pa čak i nalazišta su objekti na koje bi se mogla primijeniti klasterska analiza u svrhu dobivanja više informacija o obrascima ljudskog ponašanja te razvitku društva.

4. Znanosti poput geologije i geografije koriste klastersku analizu prilikom klasifikacije stijena, vrsta tla, riječnih sustava, gradova te regija kako bi se pronašla optimalna područja kako za potrebe zajednice, tako i za planiranje daljnjeg razvitka.

5. U granama računalne znanosti poput umjetne inteligencije, računalnog vida te kibernetike klasteriranje je moguće na uzorcima govora, pisma, otisaka prstiju, fotografijama, radiovalovima i raznim drugim objektima. Bolja klasifikacija omogućava unaprjeđenje analize računalnih sustava, odnosno veću preciznost što je osobito korisno kod novijih metoda analize prometa.

6. Što se tiče ekonomskih i političkih znanosti, metoda je primjenjiva pri analizi tržišta, ciljnih skupina te marketinških strategija. Ovakva primjena bilježi osobit rast posljednjih nekoliko godina.

Poglavlje 2

Primjer

2.1 Prikupljanje podataka

U svrhu prikupljanja podataka za izradu diplomskog rada, pripremljena je anketa. S obzirom na to da velike količine podataka najčešće nisu javno dostupne, kao inspiracija za pitanja poslužilo je istraživanje [7]. U anketi se ispituju interesi, hobiji, zdravstvene navike te fobije studenata i srednjoškolaca. Ideja je prikupiti što veći broj ispitanika iz raznih dijelova Hrvatske kako bi se uočile neke možda neočekivane grupacije. Iz tog razloga, anketa je pripremljena u online formatu i distribuirana putem društvenih mreža.

Anketa je bila javno dostupna četiri mjeseca i u tom vremenskom periodu prikupljeno je 700 različitih odgovora od ispitanika diljem zemlje. Nakon čišćenja podataka, 699 ih je ušlo u daljnju analizu. Cilj je bio postići ravnotežu između broja studenata i srednjoškolaca kako bi se u konačnici dobili statistički značajni rezultati.

Cjelokupna anketa dana je u poglavlju 4. U prvih 12 pitanja tražili su se osobni podatci o ispitaniku kao što su spol, obrazovanje, zaposlenost, mjesto rođenja i slično. Ostala pitanja podijeljena su u pet odlomaka: *Glazba, Filmovi, Fobije, Hobiji* i *Interesi*. U tim odlomcima odgovori su rangirani brojevima od 1 do 7, pri čemu 1 označava najslabiji interes, a 7 jako izraženi interes. Posljednji odlomak je *Zdravstvene navike*, gdje se ispitivala učestalost konzumacije alkohola, duhana i droga među mladim ljudima. S obzirom na to da je anketa online i anonimna, nema razloga sumnjati u istinitost odgovora.

2.2 Analiza

Analiza ankete provedena je u SAS softveru i na podatcima će biti primijenjeno hijerarhijsko i k-means klasteriranje. Zbog jednakog raspona vrijednosti, standardizacija nije bila potrebna.

Tablica 2.1: Deskriptivna statistika (SAS ispis)

The MEANS Procedure

Variable	Label	N	Minimum	Mean	Maximum	Std Dev
VoliG	Voli glazbu	699	1.0000000	6.4978541	7.0000000	0.9581103
Pop	Pop	699	1.0000000	4.9742489	7.0000000	1.6703793
Rock	Rock	699	1.0000000	4.3762518	7.0000000	1.9762362
Metal	Metal	699	1.0000000	2.0200286	7.0000000	1.6155227
HHRap	Hip hop\Rap	699	1.0000000	3.8741059	7.0000000	1.9147001
Jazz	Jazz	699	1.0000000	2.9170243	7.0000000	1.7308872
Narodno	Narodna glazba	699	1.0000000	3.5321888	7.0000000	2.2038082
Techno	Techno	699	1.0000000	3.4792561	7.0000000	1.9836182
Klasika	Klasika	699	1.0000000	3.1630901	7.0000000	1.8015465
Trash	Trash	699	1.0000000	4.0100143	7.0000000	2.1287128
House	House	699	1.0000000	3.0686695	7.0000000	1.9056182
VoliF	Voli filmove	699	1.0000000	6.2761087	7.0000000	1.1378554
Horor	Horor	699	1.0000000	3.2303290	7.0000000	2.1848387
Triler	Triler	699	1.0000000	4.9384835	7.0000000	1.8122295
Komedija	Komedija	699	1.0000000	5.7052933	7.0000000	1.4138541
Romanticni	Romantični film	699	1.0000000	4.6022890	7.0000000	1.9897432
SCI_FI	SCI-FI	699	1.0000000	4.3276109	7.0000000	2.0944657
Dokumentarni	Dokumentarni film	699	1.0000000	4.4878398	7.0000000	1.7972626
Akcijski	Akcijski film	699	1.0000000	4.9227468	7.0000000	1.7253496
Animirani	Animirani film	699	1.0000000	4.7596567	7.0000000	1.8556899
Western	Western	699	1.0000000	2.6065808	7.0000000	1.6837372
Let	Letenje	699	1.0000000	2.2589413	7.0000000	1.6817458
Visina	Visina	699	1.0000000	3.2074392	7.0000000	1.9258910
Nevrijeme	Nevrijeme	699	1.0000000	2.0329041	7.0000000	1.5065488
Mrak	Mrak	699	1.0000000	2.4391989	7.0000000	1.7263199
Pauci	Pauci	699	1.0000000	3.2188841	7.0000000	2.1470012
Dizalo	Dizalo	699	1.0000000	1.8927039	7.0000000	1.5747619
Zmije	Zmije	699	1.0000000	3.9055794	7.0000000	2.1841341
MaliPr	Mali prostori	699	1.0000000	2.4177396	7.0000000	1.8604695
Psi	Psi	699	1.0000000	1.6480687	7.0000000	1.2238233
JavniGovor	Javni govor	699	1.0000000	3.6709585	7.0000000	1.9993984
Glodavci	Glodavci	699	1.0000000	3.0214592	7.0000000	2.0420641
Zubar	Zubar	699	1.0000000	2.3218884	7.0000000	1.8637605
Igle	Igle	699	1.0000000	2.4377682	7.0000000	1.8948138
Krv	Krv	699	1.0000000	1.8597997	7.0000000	1.5450542
Bacili	Bacili	699	1.0000000	2.1959943	7.0000000	1.6394793
Citanje	Čitanje	699	1.0000000	3.9885551	7.0000000	1.8324942
Jezici	Jezici	699	1.0000000	4.0329041	7.0000000	1.8042624
LikovnaU	Likovna umjetnost	699	1.0000000	2.5350501	7.0000000	1.8104644
Gluma	Gluma	699	1.0000000	2.0114449	7.0000000	1.6437380
Ples	Ples	699	1.0000000	2.7839771	7.0000000	2.0151221
Sviranje	Sviranje	699	1.0000000	2.3190272	7.0000000	2.0035273
Pjevanje	Pjevanje	699	1.0000000	3.2389127	7.0000000	2.1694352
Koncerti	Koncerti	699	1.0000000	3.7067239	7.0000000	1.9765380
Kazaliste	Kazalište	699	1.0000000	2.9284692	7.0000000	1.7955786
VIgrice	Video igrice	699	1.0000000	3.2203147	7.0000000	2.2322434
Automobili	Automobili	699	1.0000000	2.7567954	7.0000000	2.0775286
Shopping	Shopping	699	1.0000000	3.9055794	7.0000000	1.9209926
SportP	Sport - profesionalno	699	1.0000000	2.1802575	7.0000000	2.0324331
SportA	Sport - amaterski	699	1.0000000	4.0972818	7.0000000	2.2645217
Pov	Povijest	699	1.0000000	3.6523605	7.0000000	2.0629504
Zem	Geografija	699	1.0000000	3.6824034	7.0000000	1.9904836
Psih	Psihologija	699	1.0000000	4.5278970	7.0000000	1.9964890
Pol	Politika	699	1.0000000	3.1230329	7.0000000	2.0500259
Mat	Matematika	699	1.0000000	4.1430615	7.0000000	2.2772369
Fiz	Fizika	699	1.0000000	3.4334764	7.0000000	2.1323194
Inf	Informatika	699	1.0000000	3.8855508	7.0000000	2.1633983
Bio	Biologija	699	1.0000000	3.8826896	7.0000000	2.1426159
Kem	Kemija	699	1.0000000	3.1516452	7.0000000	2.1012773
Med	Medicina	699	1.0000000	3.6065808	7.0000000	2.2719196
Pravo	Pravo	699	1.0000000	2.4992847	7.0000000	1.9146358
Ekon	Ekonomija	699	1.0000000	2.9213162	7.0000000	2.0560458
Elekt	Elektrotehnika	699	1.0000000	2.5293276	7.0000000	1.8849991
Robot	Robotika	699	1.0000000	2.5994278	7.0000000	1.9506225
Religija	Religija	699	1.0000000	3.1659514	7.0000000	2.2059926

U ovoj tablici tablici prikazane su vrijednosti minimuma, aritmetičke sredine, maksimuma i standardne devijacije za svaku od varijabli iz odlomaka *Glazba, Filmovi, Fobije* te *Hobiji i interesi*.

Tablica 2.2: Tablica frekvencija ispitanika (SAS ispis)

The FREQ Procedure

Skola				
Skola	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Srednja škola-gimnazija	258	36.91	258	36.91
Srednja škola-strukovna	74	10.59	332	47.50
Sveučilište	343	49.07	675	96.57
Veleučilište	24	3.43	699	100.00

U tablici 2.2 prikazana je frekvencija ispitanika prema vrsti obrazovne institucije. Može se primijetiti da je broj studenata i srednjoškolaca ravnomjeran ($52,5\% : 47,5\%$).

2.2.1 Glazba

U sljedećoj tablici prikazane su vrijednosti aritmetičkih sredina odgovora ispitanika na pitanja prema vrsti obrazovne institucije. Može se primijetiti da svi ispitanici pokazuju znatan interes za slušanje glazbe.

Tablica 2.3: Aritmetičke sredine podataka o glazbi (SAS ispis)

Obs	Skola	Voli_glazbu	Pop	Rock	Metal	Hip_hop_Rap	Jazz	Narodno	Techno	Klasika	Trash	House
1	Srednja škola-gimnazija	6.66	5.14	4.08	1.84	4.25	2.7	3.83	3.54	2.81	3.47	2.84
2	Srednja škola-strukovna	6.45	4.53	3.09	1.66	4.34	2.49	4.05	4.24	2.64	2.92	3.15
3	Sveučilište	6.4	4.97	4.87	2.24	3.5	3.13	3.2	3.28	3.53	4.67	3.2
4	Veleučilište	6.38	4.63	4.46	1.92	3.71	3.54	3.42	3.25	3.33	3.67	3.38

Vrijednosti kategorija ovisno o obrazovanju su relativno podjednake, primjerice slab interes za metal te snažan interes za pop glazbu.

Na podatcima je primijenjeno hijerarhijsko klasteriranje koristeći Ward metodu, metodu minimuma, metodu maksimuma te metodu prosjeka. Mjera sličnosti koja se upotrebljava je euklidska.

Ward metoda

Wardovom metodom dobivena je sljedeća podjela na klastere:

Tablica 2.4: Podjela ispitanika na klastere prema glazbi Ward metodom (SAS ispis)

Cluster History						
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Tie
3	Sveučilište	Veleučilište	2	0.1278	.872	
2	Srednja škola-gimnazija	Srednja škola-strukovna	2	0.1817	.691	
1	CL2	CL3	4	0.6906	.000	

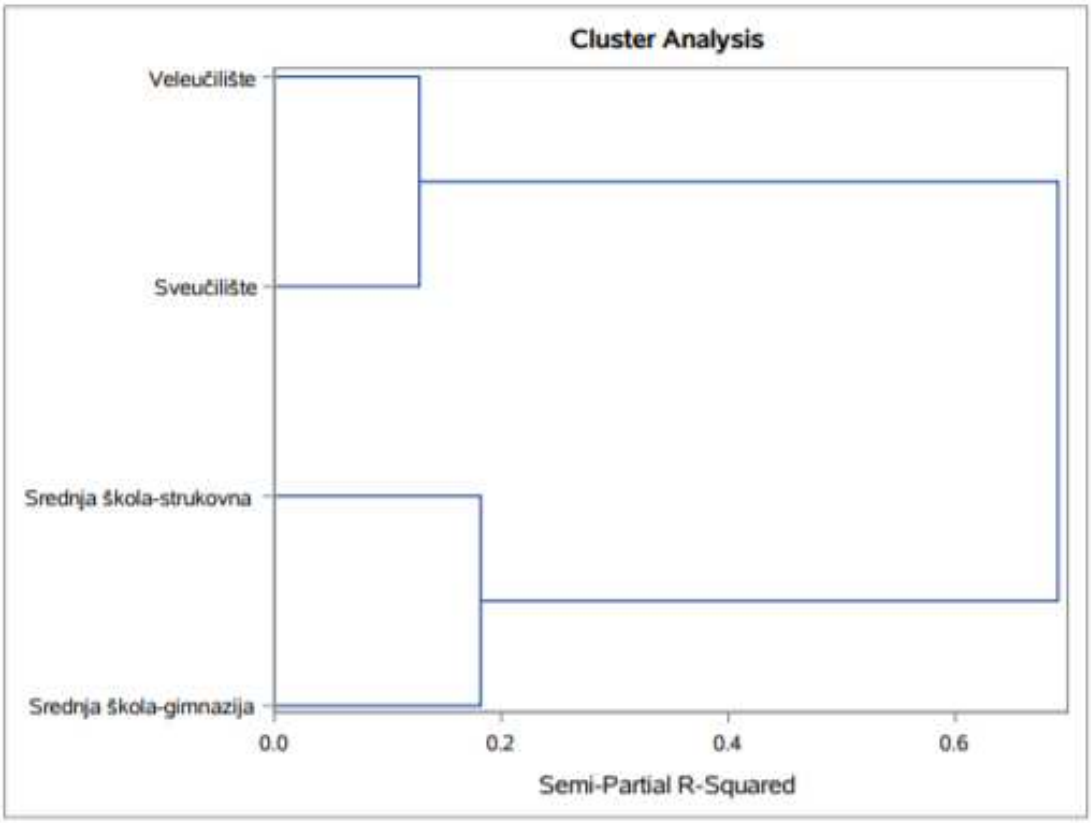
Gimnazija i strukovna škola pripadaju jednom klasteru, a sveučilište i veleučilište drugom klasteru. SAS kod kojim je dobivena ova podjela je sljedeći:

```
proc distance data = podatci out = DIST method = Euclid;
    var interval(Voli_glazbu Pop Rock Metal Hip_hop_Rap Jazz Narodno
    Techno Klasika Trash House);
    id Skola;
run;

proc cluster data = DIST method = Ward outtree = Tree;
    id Skola;
run;
```

Procedura *distance* računa matricu udaljenosti podataka prema mjeri koja je definirana naredbom *method*. Procedura *cluster* provodi hijerarhijsku klaster analizu prema vrsti koja je određena naredbom *method*.

Na slici 2.1 prikazan je pripadni dendrogram podataka o glazbenim interesima. Klasteri srednjih škola i sveučilišta jasno su separirani.



Slika 2.1: Dendrogram ispitanika prema podatcima o glazbi - Ward metoda (SAS ispis)

Metoda minimuma

Metodom minimuma dobivena je sljedeća podjela na klastere:

Tablica 2.5: Podjela ispitanika na klastere prema glazbi metodom minimuma (SAS ispis)

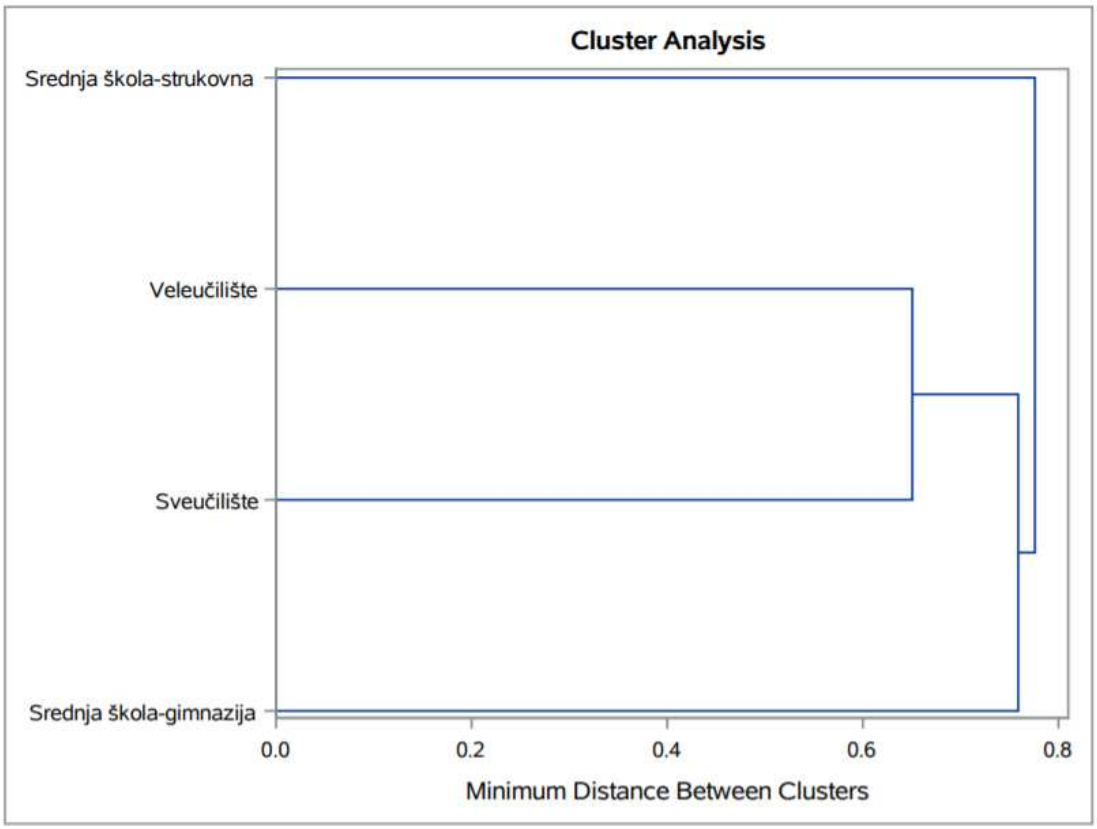
Cluster History					
Number of Clusters	Clusters Joined		Freq	Norm Minimum Distance	Tie
3	Sveučilište	Veleučilište	2	0.6506	
2	Srednja škola-gimnazija	CL3	3	0.7588	
1	CL2	Srednja škola-strukovna	4	0.7758	

Za razliku od Ward metode, dobivena je podjela na tri klastera. Sveučilište i veleučilište pripadaju jednom klasteru, a srednje škole su u zasebnim klasterima. SAS kod kojim je dobivena ova podjela je sljedeći:

```
proc distance data = podatci out = DIST method = Euclid;
    var interval(Voli_glazbu Pop Rock Metal Hip_hop_Rap Jazz Narodno
    Techno Klasika Trash House);
    id Skola;
run;

proc cluster data = DIST method = Sin outtree = Tree;
    id Skola;
run;
```

Na slici 2.2 prikazan je pripadni dendrogram podataka o glazbenim interesima. Sveučilište i veleučilište nalaze se u jednom klasteru, a gimnazija je unutarnjom granom povezana s njima. Strukovna škola spada u zaseban klaster.



Slika 2.2: Dendrogram ispitanika prema podatcima o glazbi - metoda minimuma (SAS ispis)

Metoda maksimuma

Metodom maksimuma dobivena je sljedeća podjela na klastere:

Tablica 2.6: Podjela ispitanika na klastere prema glazbi metodom maksimuma (SAS ispis)

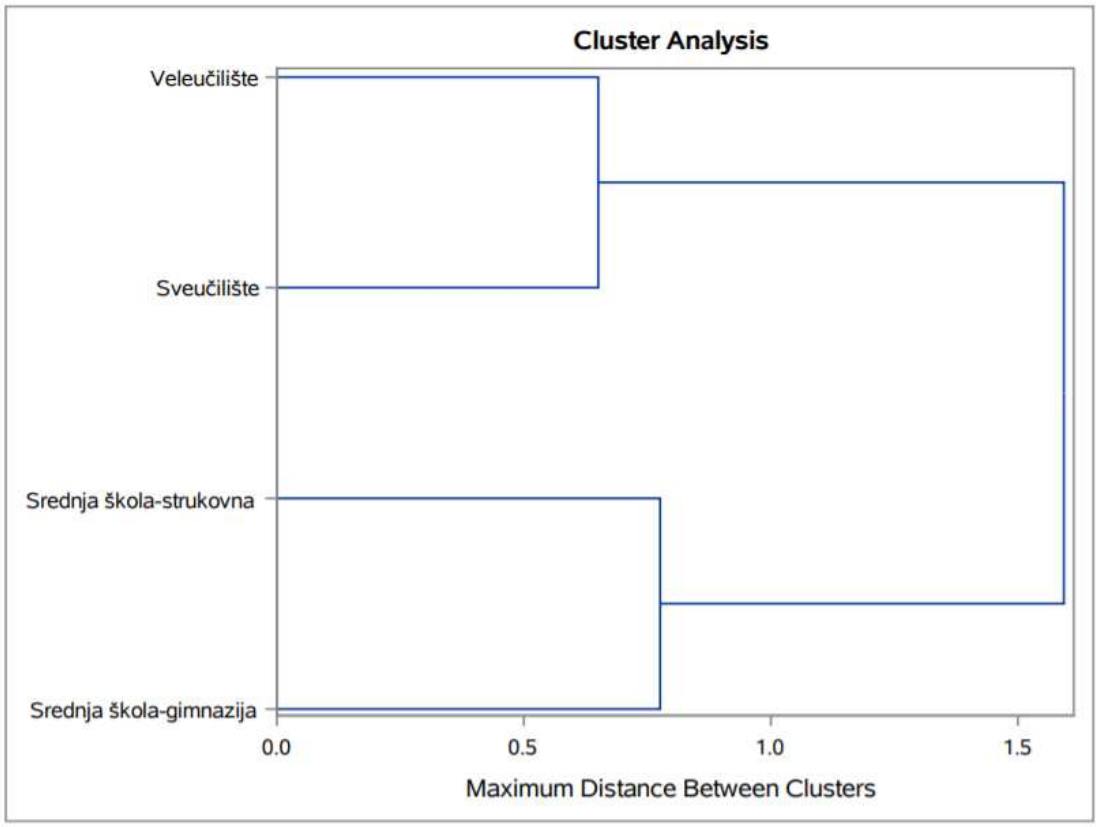
Cluster History					
Number of Clusters	Clusters Joined		Freq	Norm Maximum Distance	Tie
3	Sveučilište	Veleučilište	2	0.6506	
2	Srednja škola-gimnazija	Srednja škola-strukovna	2	0.7758	
1	CL2	CL3	4	1.5931	

Kao i kod Ward metode, dobivena je podjela na dva klastera. Sveučilište i veleučilište pripadaju jednom klasteru, a strukovna škola i gimnazija drugom klasteru. SAS kod kojim je dobivena ova podjela je sljedeći:

```
proc distance data = podatci out = DIST method = Euclid;
    var interval(Voli_glazbu Pop Rock Metal Hip_hop_Rap Jazz Narodno
    Techno Klasika Trash House);
    id Skola;
run;

proc cluster data = DIST method = Com outtree = Tree;
    id Skola;
run;
```

Na slici 2.3 prikazan je pripadni dendrogram podataka o glazbenim interesima. Obrazovne institucije jasno su separirane u klastere prema nivou obrazovanja.



Slika 2.3: Dendrogram ispitanika prema podatcima o glazbi - metoda maksimuma (SAS ispis)

Metoda prosjeka

Metodom prosjeka dobivena je sljedeća podjela na klastere:

Tablica 2.7: Podjela ispitanika na klastere prema glazbi metodom prosjeka (SAS ispis)

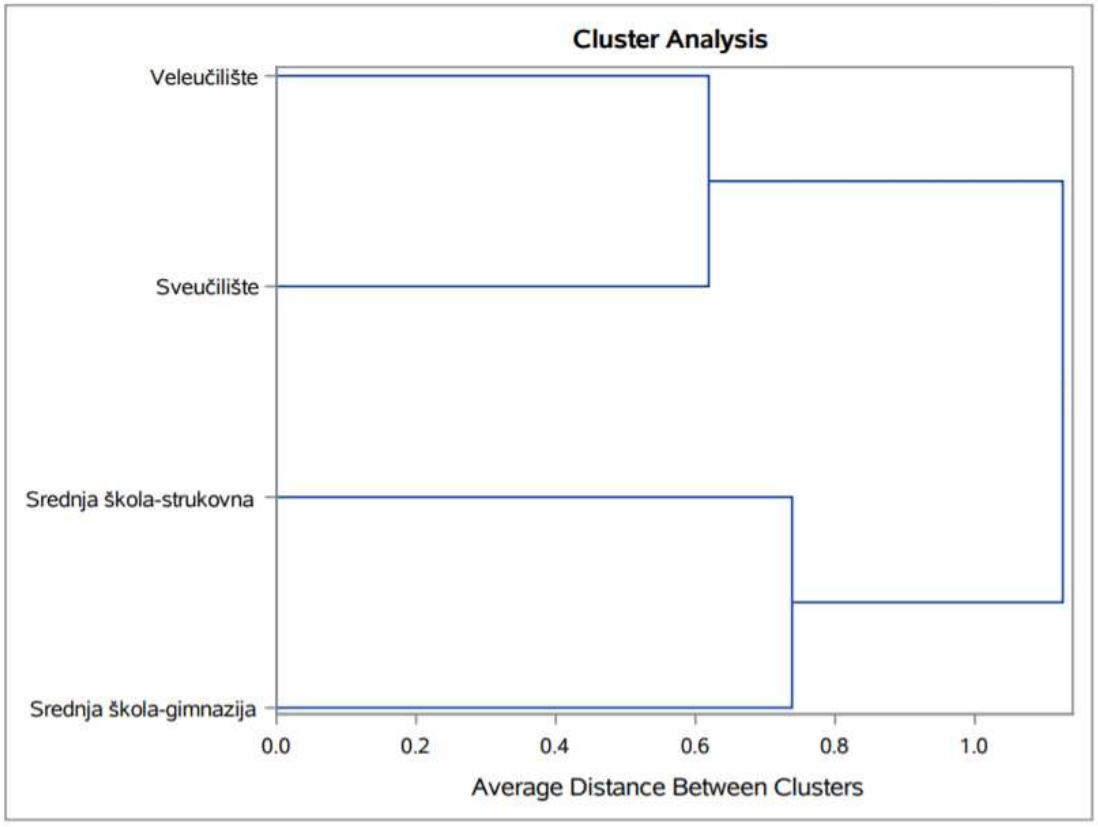
Cluster History					
Number of Clusters	Clusters Joined		Freq	Norm RMS Distance	Tie
3	Sveučilište	Veleučilište	2	0.6191	
2	Srednja škola-gimnazija	Srednja škola-strukovna	2	0.7383	
1	CL2	CL3	4	1.126	

Kao i kod Ward metode i metode maksimuma, dobivena je podjela na dva klastera. Sveučilište i veleučilište ponovno pripadaju jednom klasteru, a strukovna škola i gimnazija drugom klasteru. SAS kod kojim je dobivena ova podjela je sljedeći:

```
proc distance data = podatci out = DIST method = Euclid;
    var interval(Voli_glazbu Pop Rock Metal Hip_hop_Rap Jazz Narodno
    Techno Klasika Trash House);
    id Skola;
run;

proc cluster data = DIST method = Ave outtree = Tree;
    id Skola;
run;
```

Na slici 2.4 prikazan je pripadni dendrogram podataka o glazbenim interesima. Obrazovne institucije ponovno su jasno separirane u klastere prema nivou obrazovanja.



Slika 2.4: Dendrogram ispitanika prema podatcima o glazbi - metoda prosjeka (SAS ispis)

Za ostale skupine pitanja bit će primijenjeno hijerarhijsko klasteriranje uz Ward metodu.

K-means

Konačno, na podatcima je primijenjen nehijerarhijski k-means algoritam. U prvom slučaju konačan broj klastera je ograničen na dva naredbom *maxclusters*. Kod iz SAS-a je sljedeći:

```
proc fastclus data = podatci out = klaster2 maxclusters = 2
    nomiss maxiter = 300;
    var Voli_glazbu Pop Rock Metal Hip_hop_Rap Jazz Narodno Techno
    Klasika Trash House;
run;

proc print data = klaster2;
    var Skola cluster;
run;
```

Procedura *fastclus* provodi k-means algoritam na podatcima. Naredbom *nomiss* u analizu ne ulaze vrijednosti koje nedostaju, a naredba *print* služi za konačan ispis rezultata.

Tablica 2.8: Podjela ispitanika na dva klastera prema podatcima o glazbi (SAS ispis)

Obs	Skola	CLUSTER
1	Srednja škola-gimnazija	2
2	Srednja škola-strukovna	2
3	Sveučilište	1
4	Veleučilište	1

Postavljanjem ograničenja konačnog broja klastera na dva dobiven je klaster koji sadrži gimnaziju i srednju školu te klaster koji sadrži sveučilište i veleučilište.

Tablica 2.9: Aritmetičke sredine varijabli klastera glazbe - dva klastera (SAS ispis)

Cluster Means							
Cluster	Voli_glazbu	Pop	Rock	Metal	Hip_hop_Rap	Jazz	Narodno
1	6.390000000	4.800000000	4.665000000	2.080000000	3.605000000	3.335000000	3.310000000
2	6.555000000	4.835000000	3.585000000	1.750000000	4.295000000	2.595000000	3.940000000

Cluster Means				
Cluster	Techno	Klasika	Trash	House
1	3.265000000	3.430000000	4.170000000	3.290000000
2	3.890000000	2.725000000	3.195000000	2.995000000

Kada se promotre aritmetičke sredine odgovora ispitanika, uočavaju se razlike između klastera. Za svaku kategoriju odgovori se razlikuju, osim za kategoriju pop gdje su aritmetičke sredine približno jednake.

U drugom slučaju, konačan broj klastera ograničen je na tri. Pripadni SAS kod je sljedeći:

```
proc fastclus data = podatci out = klaster2 maxclusters = 3
    nomiss maxiter = 300;
    var Voli_glazbu Pop Rock Metal Hip_hop_Rap Jazz Narodno Techno
    Klasika Trash House;
run;

proc print data = klaster2;
    var Skola cluster;
run;
```

Dobivena je sljedeća podjela na klastere:

Tablica 2.10: Podjela ispitanika na tri klastera prema podatcima o glazbi (SAS ispis)

Obs	Skola	CLUSTER
1	Srednja škola-gimnazija	1
2	Srednja škola-strukovna	2
3	Sveučilište	3
4	Veleučilište	3

Gimnazija i strukovna škola pripadaju u zasebne klastere, dok su sveučilište i veleučilište dio istog klastera. Što se tiče odgovora ispitanika, vrijednosti aritmetičkih sredina dane su u tablici 2.10. Za svaku od kategorija aritmetičke sredine odgovora razlikuju se između klastera. Najusklađeniji su odgovori što se tiče interesa prema glazbi, a najveće odstupanje uočava se u kategoriji trash glazbe.

Tablica 2.11: Aritmetičke sredine varijabli klastera glazbe - tri klastera (SAS ispis)

Cluster Means							
Cluster	Voli_glazbu	Pop	Rock	Metal	Hip_hop_Rap	Jazz	Narodno
1	6.660000000	5.140000000	4.080000000	1.840000000	4.250000000	2.700000000	3.830000000
2	6.450000000	4.530000000	3.090000000	1.660000000	4.340000000	2.490000000	4.050000000
3	6.390000000	4.800000000	4.665000000	2.080000000	3.605000000	3.335000000	3.310000000

Cluster Means				
Cluster	Techno	Klasika	Trash	House
1	3.540000000	2.810000000	3.470000000	2.840000000
2	4.240000000	2.640000000	2.920000000	3.150000000
3	3.265000000	3.430000000	4.170000000	3.290000000

2.2.2 Filmovi

U tablici 2.12 prikazane su aritmetičke sredine interesa ispitanika ovisno kojoj instituciji pripadaju. Kao i za glazbu, očit je velik interes svih ispitanika. Za većinu kategorija interes je ujednačen, ali uočavaju se neke razlike. Dokumentarni filmovi najmanje zanimaju ispitanike strukovnih škola, dok oni prednjače po interesu za horore. Također, trileri su najzastupljeniji kod ispitanika koji pohađaju sveučilište.

Tablica 2.12: Aritmetičke sredine podataka o filmovima (SAS ispis)

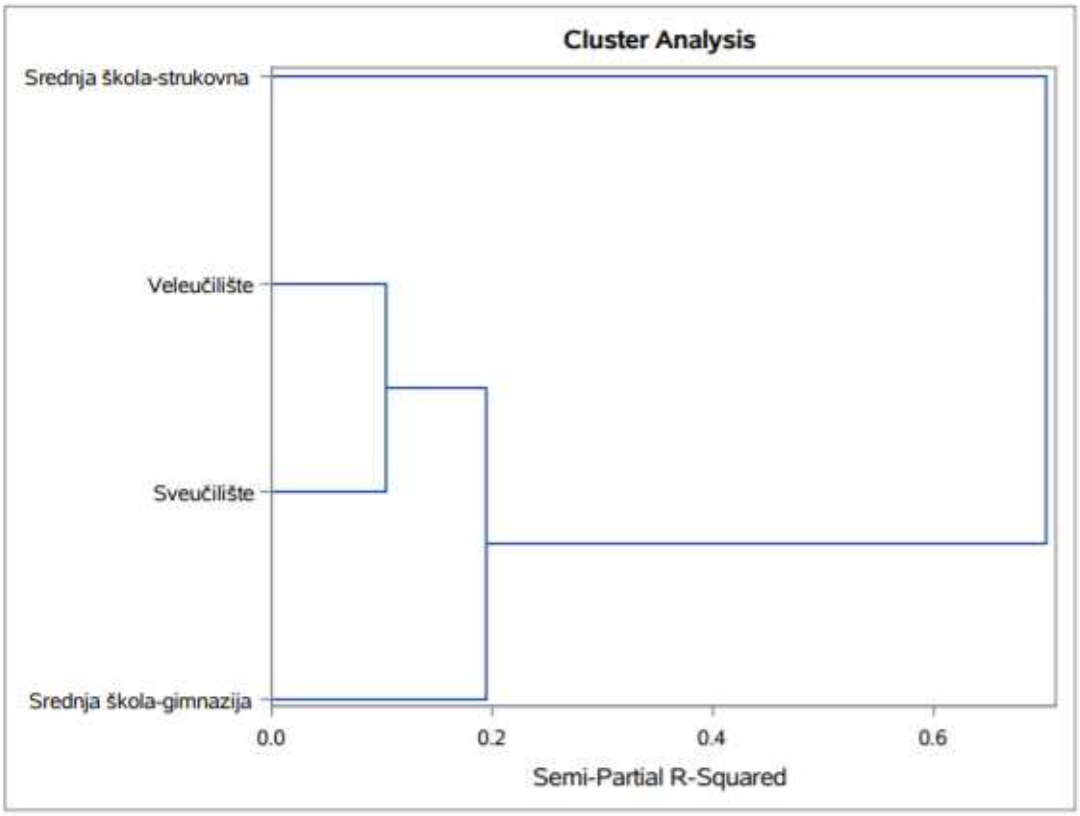
Obs	Skola	Voli_film	Horor	Triler	Komedija	Romanticni_film	SCI_FI	Dokumentarni_film	Akcijski_film	Animirani_film	Western
1	Srednja škola-gimnazija	6.41	3.19	4.7	5.8	4.82	4.48	4.19	5.17	4.98	2.48
2	Srednja škola-strukovna	6.26	4.34	4.43	5.97	4.59	3.8	3.92	5.34	4.2	2.86
3	Sveučilište	6.19	3.04	5.25	5.56	4.45	4.31	4.79	4.64	4.72	2.65
4	Veleučilište	6.17	2.96	4.67	5.92	4.46	4.5	5.13	5	4.71	2.5

Primjenom hijerarhijskog klasteriranja Wardovom metodom, dobiveni su sljedeći rezultati:

Tablica 2.13: Podjela na klastere prema filmovima (SAS ispis)

Cluster History						
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Tie
3	Sveučilište	Veleučilište	2	0.1036	.896	
2	Srednja škola-gimnazija	CL3	3	0.1945	.702	
1	CL2	Srednja škola-strukovna	4	0.7019	.000	

U ovom slučaju dobivena je podjela na tri klastera, pri čemu sveučilište i veleučilište pripadaju istom klasteru. Strukovna škola povezana je s korijenom preko vanjske grane i možemo je smatrati outlierom. Na slici 2.5 prikazan je pripadni dendrogram. Odabrana mjera sličnosti je ponovno euklidska.



Slika 2.5: Dendrogram ispitanika prema podatcima o filmovima (SAS ispis)

Pripadni SAS kod je sljedeći:

```
proc distance data = podatci out = DIST method = Euclid;
    var interval(Voli_film Horor Triler Komedija Romanticni_film SCI_FI
    Dokumentarni_film Akcijski_film Animirani_film Western);
    id Skola;
run;

proc cluster data = DIST method = Ward outtree = Tree;
    id Skola;
run;
```

Nakon hijerarhijskog, na podatcima je primijenjen i nehijerarhijski k-means algoritam uz ograničenje na maksimalno dva klastera. Dobivena je sljedeća podjela:

Obs	Skola	CLUSTER
1	Srednja škola-gimnazija	1
2	Srednja škola-strukovna	2
3	Sveučilište	1
4	Veleučilište	1

Tablica 2.14: Podjela ispitanika na dva klastera prema podatcima o filmovima (SAS ispis)

Kada se promotre aritmetičke sredine odgovora ispitanika pojedinih klastera, može se uočiti gotovo jednak interes za neke kategorije, npr. romantični filmovi i komedije, dok su kod ostalih razlike puno veće, primjerice zanimanje za horore i dokumentarne filmove se drastično razlikuje.

Tablica 2.15: Aritmetičke sredine varijabli klastera filma - dva klastera (SAS ispis)

Cluster Means						
Cluster	Voli_film	Horor	Triler	Komedija	Romanticni_film	SCI_FI
1	6.256666667	3.063333333	4.873333333	5.760000000	4.576666667	4.430000000
2	6.260000000	4.340000000	4.430000000	5.970000000	4.590000000	3.800000000

Cluster Means				
Cluster	Dokumentarni_film	Akcijski_film	Animirani_film	Western
1	4.703333333	4.936666667	4.803333333	2.543333333
2	3.920000000	5.340000000	4.200000000	2.860000000

U drugom slučaju postavljeno je ograničenje na tri klastera. Sveučilište i veleučilište pripadaju istom klasteru, a gimnazija i strukovna škola su u zasebnim klasterima, kao što je vidljivo u tablici 2.16.

Tablica 2.16: Podjela ispitanika na tri klastera prema podatcima o filmovima (SAS ispis)

Obs	Skola	CLUSTER
1	Srednja škola-gimnazija	1
2	Srednja škola-strukovna	2
3	Sveučilište	3
4	Veleučilište	3

Aritmetičke sredine po klasterima dane su u tablici 2.17. Moguće je uočiti odstupanja u odgovorima po klasterima, osim za neke kategorije kao što su komedija i western.

Tablica 2.17: Aritmetičke sredine varijabli klastera filma - tri klastera (SAS ispis)

Cluster Means						
Cluster	Voli_film	Horor	Triler	Komedija	Romanticni_film	SCI_FI
1	6.410000000	3.190000000	4.700000000	5.800000000	4.820000000	4.480000000
2	6.260000000	4.340000000	4.430000000	5.970000000	4.590000000	3.800000000
3	6.180000000	3.000000000	4.960000000	5.740000000	4.455000000	4.405000000

Cluster Means				
Cluster	Dokumentarni_film	Akcijski_film	Animirani_film	Western
1	4.190000000	5.170000000	4.980000000	2.480000000
2	3.920000000	5.340000000	4.200000000	2.860000000
3	4.960000000	4.820000000	4.715000000	2.575000000

Pripadni SAS kod je sljedeći:

```
proc fastclus data = podatci out = klaster2 maxclusters = 3
    nomiss maxiter = 300;
    var Voli_film Horor Triler Komedija Romanticni_film SCI_FI
    Dokumentarni_film Akcijski_film Animirani_film Western;
run;

proc print data = klaster2;
    var Skola cluster;
run;
```

2.2.3 Fobije

Među studentima i srednjoškolcima aritmetičke sredine odgovora za fobije podjednake su za svaki od tipova, što je vidljivo u sljedećoj tablici:

Tablica 2.18: Aritmetičke sredine podataka o fobijama (SAS ispis)

Obs	Skola	Letenje	Visina	Nevrijeme	Mrak	Pauci	Dizalo	Zmije
1	Srednja škola-gimnazija	2.2	3.05	1.98	2.66	3.49	1.83	3.83
2	Srednja škola-strukovna	2.28	3.16	1.93	2.27	3.28	1.99	3.62
3	Sveučilište	2.33	3.36	2.13	2.36	3.06	1.97	4.02
4	Veleučilište	1.79	2.83	1.42	1.75	2.38	1.25	3.96

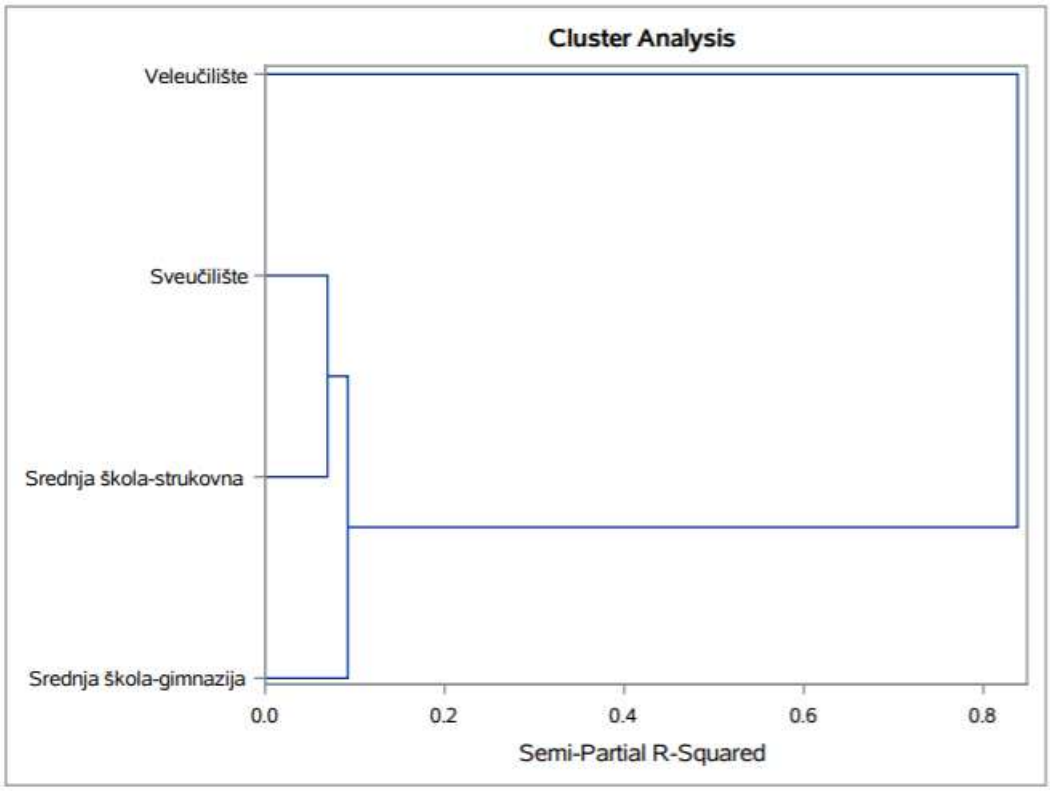
Obs	Mali_prostori	Psi	Javni_govor	Glodavci	Zubar	Igle	Krv	Bacili
1	2.4	1.71	3.57	3.17	2.03	2.51	1.87	2.2
2	2.61	1.55	3.5	2.82	2.23	2.41	1.81	2.27
3	2.44	1.65	3.77	3	2.53	2.39	1.88	2.22
4	1.75	1.29	3.92	2.33	2.75	2.42	1.54	1.63

Hijerarhijskim klasteriranjem Wardovom metodom dobiveni su sljedeći rezultati:

Tablica 2.19: Podjela na klastere prema fobijama (SAS ispis)

Cluster History						
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Tie
3	Srednja škola-strukovna	Sveučilište	2	0.0695	.930	
2	Srednja škola-gimnazija	CL3	3	0.0922	.838	
1	CL2	Veleučilište	4	0.8383	.000	

Dakle, ispitanici su podijeljeni u tri klastera pri čemu je veleučilište vanjskom granom, a gimnazija unutarnjom granom povezana s korijenom dendrograma (Slika 2.6).



Slika 2.6: Dendrogram ispitanika prema podatcima o fobijama (SAS ispis)

Pripadni SAS kod je sljedeći:

```
proc distance data = podatci out = DIST method = Euclid;
    var interval(Letenje Visina Nevrijeme Mrak Pauci Dizalo Zmije
    Mali_prostori Psi Javni_govor Glodavci Zubar Igle Krv Bacili);
    id Skola;
run;

proc cluster data = DIST method = Ward outtree = Tree;
    id Skola;
run;
```

Na podatcima je primijenjeno i nehijerarhijsko klasteriranje k-means algoritmom uz ograničenje na dva klastera. Tada gimnazija, strukovna škola te sveučilište pripadaju istom klasteru, a veleučilište je u zasebnom.

Tablica 2.20: Podjela ispitanika na dva klastera prema podatcima o fobijama (SAS ispis)

Obs	Skola	CLUSTER
1	Srednja škola-gimnazija	1
2	Srednja škola-strukovna	1
3	Sveučilište	1
4	Veleučilište	2

SAS kod je sljedeći:

```
proc fastclus data = podatci out = klaster2 maxclusters = 2
    nomiss maxiter = 300;
    var Letenje Visina Nevrijeme Mrak Pauci Dizalo Zmije Mali_prostori
    Psi Javni_govor Glodavci Zubar Igle Krv Bacili;
run;

proc print data = klaster2;
    var Skola cluster;
run;
```

U tablici 2.21 dane su aritmetičke sredine odgovora ispitanika po kalsterima. Dok se za neke kategorije odgovori znatno razlikuju, zanimljivo je primijetiti približno jednake odgovore za igle i zmije.

Tablica 2.21: Aritmetičke sredine varijabli klastera fobija - dva klastera (SAS ispis)

Cluster Means								
Cluster	**Letenje**	**Visina**	**Nevrijeme**	**Mrak**	**Pauci**	**Dizalo**	**Zmije**	**Mali_prostori**
1	2.270000000	3.190000000	2.013333333	2.430000000	3.276666667	1.930000000	3.823333333	2.483333333
2	1.790000000	2.830000000	1.420000000	1.750000000	2.380000000	1.250000000	3.960000000	1.750000000

Cluster Means							
Cluster	**Psi**	**Javni_govor**	**Glodavci**	**Zubar**	**Igle**	**Krv**	**Bacili**
1	1.636666667	3.613333333	2.996666667	2.263333333	2.436666667	1.853333333	2.230000000
2	1.290000000	3.920000000	2.330000000	2.750000000	2.420000000	1.540000000	1.630000000

U drugom slučaju, napravljena je i podjela na klastere k-means algoritmom uz ograničenje na tri klastera. Tada se dobiva sljedeća podjela:

Tablica 2.22: Podjela ispitanika na tri klastera prema podatcima o fobijama (SAS ispis)

Obs	**Skola**	**CLUSTER**
1	Srednja škola-gimnazija	1
2	Srednja škola-strukovna	3
3	Sveučilište	3
4	Veleučilište	2

Kao što se vidi iz tablice 2.22, gimnazija i veleučilište pripadaju zasebnim klasterima dok su strukovna škola i sveučilište dio istog klastera.

Prema aritmetičkim sredinama odgovora ispitanika, uočavaju se sličnosti između prvog i trećeg klastera, dok drugi klaster u kojem je veleučilište generalno odstupa. Slično se moglo vidjeti i na dendrogramu 2.6, gdje je veleučilište outlier. Može se primijetiti da kategorije zmije i igle ponovno postižu približno jednake odgovore.

Tablica 2.23: Aritmetičke sredine varijabli klastera fobija - tri klastera (SAS ispis)

Cluster Means								
Cluster	Letenje	Visina	Nevrijeme	Mrak	Pauci	Dizalo	Zmije	Mali_prostori
1	2.200000000	3.050000000	1.980000000	2.660000000	3.490000000	1.830000000	3.830000000	2.400000000
2	1.790000000	2.830000000	1.420000000	1.750000000	2.380000000	1.250000000	3.960000000	1.750000000
3	2.305000000	3.260000000	2.030000000	2.315000000	3.170000000	1.980000000	3.820000000	2.525000000

Cluster Means							
Cluster	Psi	Javni_govor	Glodavci	Zubar	Igle	Krv	Bacili
1	1.710000000	3.570000000	3.170000000	2.030000000	2.510000000	1.870000000	2.200000000
2	1.290000000	3.920000000	2.330000000	2.750000000	2.420000000	1.540000000	1.630000000
3	1.600000000	3.635000000	2.910000000	2.380000000	2.400000000	1.845000000	2.245000000

SAS kod je sljedeći:

```
proc fastclus data = podatci out = klaster2 maxclusters = 3
    nomiss maxiter = 300;
   var Letenje Visina Nevrijeme Mrak Pauci Dizalo Zmije Mali_prostori
   Psi Javni_govor Glodavci Zubar Igle Krv Bacili;
run;

proc print data = klaster2;
    var Skola cluster;
run;
```

2.2.4 Hobiji

Aritmetičke sredine odgovora ispitanika dane su u tablici 2.24. Za razliku od prethodnih kategorija, postoje veće razlike ovisno o obrazovnoj instituciji. Čitanje i odlasci u kazalište su najmanje zastupljeni hobiji kod ispitanika iz strukovnih škola dok prednjače po profesionalnom bavljenju sportom i video igricama. S druge strane, gimnazijalci prednjače po jezicima i sviranju.

Tablica 2.24: Aritmetičke sredine podataka o hobijima (SAS ispis)

Obs	Skola	Citanje	Jezici	Likovna_umjetnost	Gluma	Ples	Sviranje	Pjevanje	Koncerti	Kazaliste
1	Srednja škola-gimnazija	3.91	4.24	2.79	2.27	3.04	2.59	3.43	3.6	2.95
2	Srednja škola-strukovna	2.74	4.04	2.86	2.2	2.78	1.96	2.81	3.05	2.14
3	Sveučilište	4.3	3.89	2.3	1.8	2.62	2.24	3.19	3.94	3.12
4	Veleučilište	4.21	3.83	2.17	1.71	2.29	1.54	3.25	3.42	2.46

Obs	Video_igrice	Automobili	Shopping	Sport_profesionalno	Sport_amaterski
1	3.09	2.4	4.05	2.45	4.23
2	4.36	3.43	4.35	2.91	3.09
3	3.01	2.84	3.7	1.85	4.18
4	4.08	3.25	3.96	1.79	4.5

Hijerarhijskim klasteriranjem Wardovom metodom dobivena je podjela na tri klastera. Sveučilište i gimnazija pripadaju istom klasteru, a veleučilište i strukovna škola su u zasebnim klasterima. Pripadni SAS kod je sljedeći:

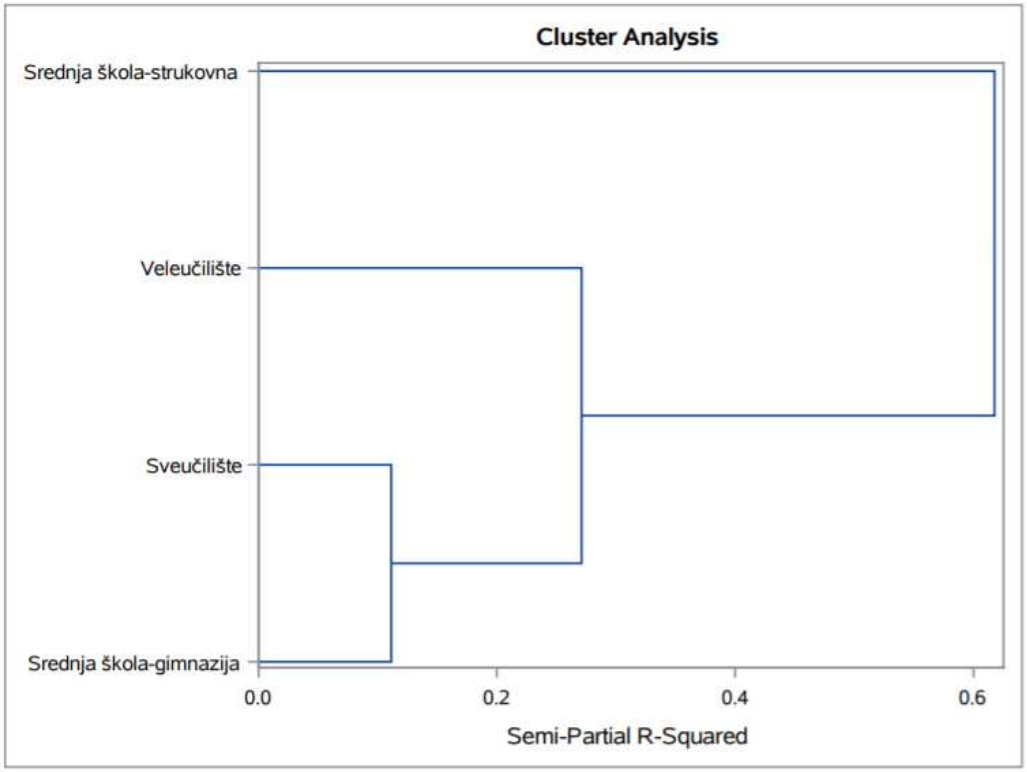
```
proc distance data = podatci out = DIST method = Euclid;
    var interval(Citanje Jezici Likovna_umjetnost Gluma Ples Sviranje
    Pjevanje Koncerti Kazaliste Video_igrice Automobili Shopping
    Sport_profesionalno Sport_amaterski);
    id Skola;
run;

proc cluster data = DIST method = Ward outtree = Tree;
    id Skola;
run;
```

Tablica 2.25: Podjela na klastere prema hobijima (SAS ispis)

Cluster History						
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Tie
3	Srednja škola-gimnazija	Sveučilište	2	0.1114	.889	
2	CL3	Veleučilište	3	0.2711	.618	
1	CL2	Srednja škola-strukovna	4	0.6176	.000	

Strukovna škola se najviše razlikuje po odgovorima i vanjskom granom je povezana s korijenom dendrograma. Sveučilište i gimnazija spadaju u zaseban klaster, a veleučilište je s njima povezano unutarnjom granom i time sličnije po odgovorima upravo tom klasteru. Na slici 2.7 prikazan je odgovarajući dendrogram.



Slika 2.7: Dendrogram ispitanika prema podatcima o hobijima (SAS ispis)

Na skupu podataka primijenjen je k-means algoritam uz ograničenje konačnog broja klastera na dva. U tom slučaju gimnazija, veleučilište i sveučilište pripadaju istom klasteru dok je strukovna škola odvojena.

Tablica 2.26: Podjela ispitanika na dva klastera prema podatcima o hobijima (SAS ispis)

Obs	Skola	CLUSTER
1	Srednja škola-gimnazija	1
2	Srednja škola-strukovna	2
3	Sveučilište	1
4	Veleučilište	1

Tablica 2.27: Aritmetičke sredine varijabli klastera hobija - 2 klastera (SAS ispis)

Cluster Means								
Cluster	Citanje	Jezici	Likovna_umjetnost	Gluma	Ples	Sviranje	Pjevanje	Koncerti
1	4.140000000	3.986666667	2.420000000	1.926666667	2.650000000	2.123333333	3.290000000	3.653333333
2	2.740000000	4.040000000	2.860000000	2.200000000	2.780000000	1.960000000	2.810000000	3.050000000

Cluster Means						
Cluster	Kazaliste	Video_igrice	Automobili	Shopping	Sport_profesionalno	Sport_amaterski
1	2.843333333	3.393333333	2.830000000	3.903333333	2.030000000	4.303333333
2	2.140000000	4.360000000	3.430000000	4.350000000	2.910000000	3.090000000

Može se primijetiti da se za svaku kategoriju aritmetičke sredine odgovora ispitanika razlikuju, osim za ples gdje su vrijednosti približno jednake.

Pripadni SAS kod je sljedeći:

```
proc fastclus data = podatci out = klaster2 maxclusters = 2
    nomiss maxiter = 300;
    var Citanje Jezici Likovna_umjetnost Gluma Ples Sviranje Pjevanje
    Koncerti Kazaliste Video_igrice Automobili Shopping
    Sport_profesionalno Sport_amaterski;
run;

proc cluster data = DIST method = Ward outtree = Tree;
    id Skola;
run;
```

U drugom slučaju primijenjen je k-means algoritam s ograničenjem na tri klastera. Dobivena je sljedeća podjela:

Tablica 2.28: Podjela ispitanika na tri klastera prema podatcima o hobijima (SAS ispis)

Obs	Skola	CLUSTER
1	Srednja škola-gimnazija	1
2	Srednja škola-strukovna	2
3	Sveučilište	1
4	Veleučilište	3

Kao što je prikazano na dendrogramu 2.7, sveučilište i gimnazija pripadaju istom klasteru, a strukovna škola i veleučilište su u odvojenim klasterima. Aritmetičke sredine odgovora ispitanika dane su u tablici 2.29. U svim kategorijama postoje znatne razlike između odgovora, međutim vidljivo je da se odgovori ispitanika klastera 2 najviše razlikuju od ostalih, odnosno da studenti s veleučilišta imaju najrazličitije interese. Slična kategorizacija odgovara i dendrogramu na kojem je veleučilište prikazano kao outlier.

Tablica 2.29: Aritmetičke sredine varijabli klastera hobija - 3 klastera (SAS ispis)

Cluster Means								
Cluster	Citanje	Jezici	Likovna_umjetnost	Gluma	Ples	Sviranje	Pjevanje	Koncerti
1	4.105000000	4.065000000	2.545000000	2.035000000	2.830000000	2.415000000	3.310000000	3.770000000
2	2.740000000	4.040000000	2.860000000	2.200000000	2.780000000	1.960000000	2.810000000	3.050000000
3	4.210000000	3.830000000	2.170000000	1.710000000	2.290000000	1.540000000	3.250000000	3.420000000

Cluster Means						
Cluster	Kazaliste	Video_igrice	Automobili	Shopping	Sport_profesionalno	Sport_amaterski
1	3.035000000	3.050000000	2.620000000	3.875000000	2.150000000	4.205000000
2	2.140000000	4.360000000	3.430000000	4.350000000	2.910000000	3.090000000
3	2.460000000	4.080000000	3.250000000	3.960000000	1.790000000	4.500000000

Pripadni SAS kod je sljedeći:

```
proc fastclus data = podatci out = klaster2 maxclusters = 3
    nomiss maxiter = 300;
    var Citanje Jezici Likovna_umjetnost Gluma Ples Sviranje Pjevanje
    Koncerti Kazaliste Video_igrice Automobili Shopping
    Sport_profesionalno Sport_amaterski;
run;

proc cluster data = DIST method = Ward outtree = Tree;
    id Skola;
run;
```

Dakle, k-means algoritmom s ograničenjem maksimalnog broja klastera na tri uz euklidsku mjeru dobivena je jednaka podjela na klastere koja je sugerirana dendrogramom.

2.2.5 Interesi

Što se tiče interesa, odgovori ispitanika su sljedeći:

Tablica 2.30: Aritmetičke sredine podataka o interesima (SAS ispis)

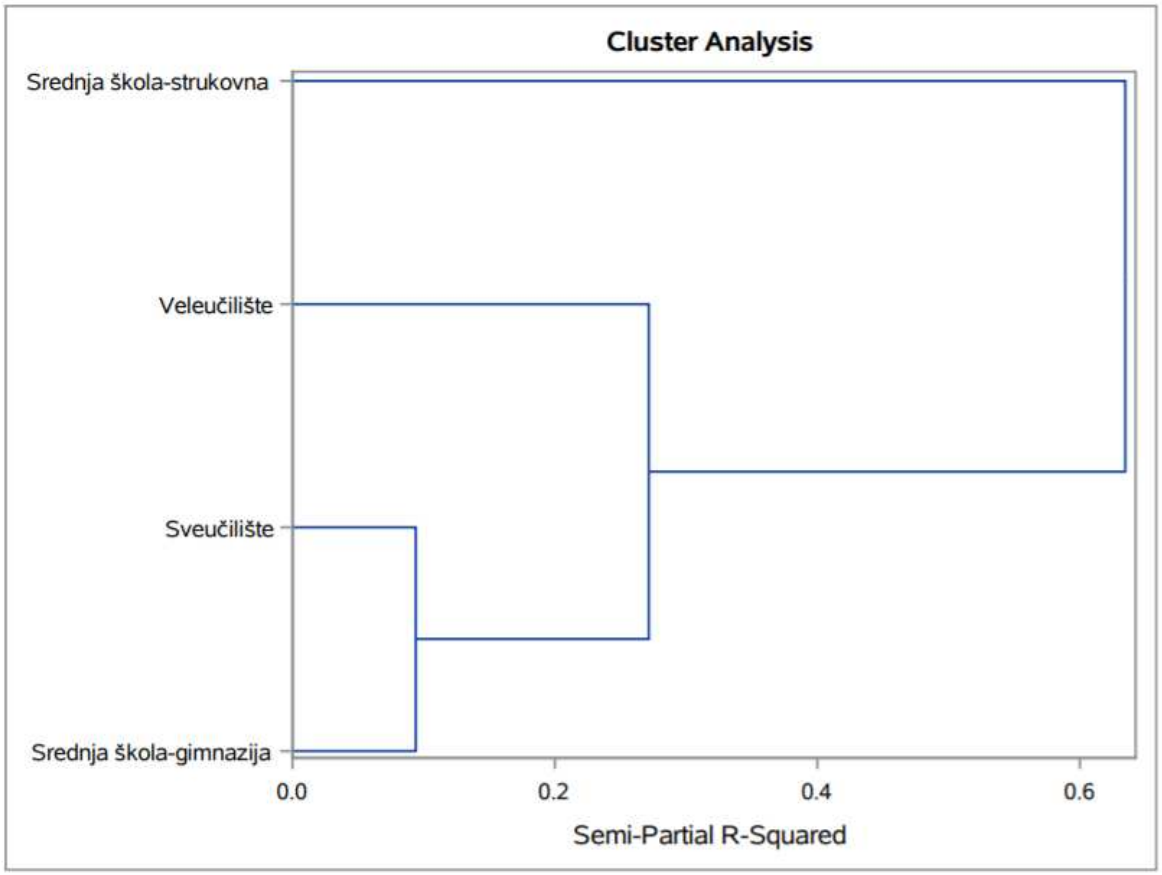
Obs	Skola	Povijest	Geografija	Psihologija	Politika	Matematika	Fizika	Informatika	Biologija	Kemija
1	Srednja škola-gimnazija	3.66	3.55	4.69	3.09	4.12	3.6	3.39	4.05	3.53
2	Srednja škola-strukovna	3.23	3.46	3.32	2.86	2.97	2.68	3.64	3.03	2.22
3	Sveučilište	3.72	3.78	4.61	3.22	4.45	3.53	4.31	3.97	3.09
4	Veleučilište	4	4.33	5.33	2.92	3.54	2.58	3.88	3.42	2.88

Obs	Medicina	Pravo	Ekonomija	Elektrotehnika	Robotika	Religija
1	3.72	2.77	2.69	2.43	2.59	3.13
2	2.86	2.59	3.09	2.8	2.78	2.68
3	3.63	2.29	3.06	2.53	2.57	3.29
4	4.25	2.29	2.92	2.75	2.58	3.25

Hijerarhijskim klasteriranjem uz Ward metodu dobivena je podjela na tri klastera. Kao i kod hobija, interesi ispitanika se razlikuju ovisno o obrazovanju. Primjerice, interes za matematiku i kemiju je znatno veći među gimnazijalcima nego učenicima strukovnih škola. Slično, interes za medicinu najviše iskazuju studenti veleučilišta.

Tablica 2.31: Podjela na klastere prema interesima (SAS ispis)

Cluster History						
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Tie
3	Srednja škola-gimnazija	Sveučilište	2	0.0939	.906	
2	CL3	Veleučilište	3	0.2715	.635	
1	CL2	Srednja škola-strukovna	4	0.6345	.000	



Slika 2.8: Dendrogram ispitanika prema podatcima o interesima (SAS ispis)

Pripadni SAS kod je sljedeći:

```
proc distance data = podatci out = DIST method = Euclid;
    var interval(Povijest Geografija Psihologija Politika Matematika
    Fizika Informatika Biologija Kemija Medicina Pravo Ekonomija
    Elektrotehnika Robotika Religija);
    id Skola;
run;

proc cluster data = DIST method = Ward outtree = Tree;
    id Skola;
run;
```

Primjenom k-means algoritma uz ograničenje konačnog broja klastera na dva uz euklidsku mjeru, dobivena je sljedeća podjela:

Tablica 2.32: Podjela ispitanika na dva klastera prema podatcima o interesima (SAS ispis)

Obs	Skola	CLUSTER
1	Srednja škola-gimnazija	1
2	Srednja škola-strukovna	2
3	Sveučilište	1
4	Veleučilište	1

Gimnazija, sveučilište i veleučilište pripadaju istom klasteru, dok je strukovna škola u zasebnom klasteru. Promatrajući dendrogram 2.8, ista podjela bi se dobila kada bi se dendrogram „odrezao" na prvoj razini. Pripadne aritmetičke sredine klastera dane su u tablici 2.33.

Tablica 2.33: Aritmetičke sredine varijabli klastera interesa - 2 klastera (SAS ispis)

Cluster Means								
Cluster	Povijest	Geografija	Psihologija	Politika	Matematika	Fizika	Informatika	Biologija
1	3.793333333	3.886666667	4.876666667	3.076666667	4.036666667	3.236666667	3.860000000	3.813333333
2	3.230000000	3.460000000	3.320000000	2.860000000	2.970000000	2.680000000	3.640000000	3.030000000

Cluster Means							
Cluster	Kemija	Medicina	Pravo	Ekonomija	Elektrotehnika	Robotika	Religija
1	3.166666667	3.866666667	2.450000000	2.890000000	2.570000000	2.580000000	3.223333333
2	2.220000000	2.860000000	2.590000000	3.090000000	2.800000000	2.780000000	2.680000000

Očigledna su određena odstupanja prema kategorijama interesa, primjerice ispitanici klastera 2 (Srednja škola - strukovna) znatno su manje zainteresirani za prirodoslovne znanosti, dok pokazuju veći interes za ekonomiju i elektrotehniku.

Pripadni SAS kod je sljedeći:

```
proc fastclus data = podaci out = klaster2 maxclusters = 2
    nomiss maxiter = 300;
    var Povijest Geografija Psihologija Politika Matematika Fizika
    Informatika Biologija Kemija Medicina Pravo Ekonomija
    Elektrotehnika Robotika Religija;
run;

proc print data = klaster2;
    var Skola cluster;
run;
```

Nadalje, na podatcima je primijenjeno i k-means klasteriranje s ograničenjem na tri klastera. Dobivena je jednaka podjela koja je i sugerirana dendrogramom: gimnazija i sveučilište pripadaju istom klasteru, a veleučilište i strukovna škola su u zasebnim klasterima. Za razliku od hobija, veće su razlike između klastera.

Tablica 2.34: Podjela ispitanika na tri klastera prema podatcima o interesima (SAS ispis)

Obs	Skola	CLUSTER
1	Srednja škola-gimnazija	1
2	Srednja škola-strukovna	2
3	Sveučilište	1
4	Veleučilište	3

Slično kao i kod podjele na dva klastera, gimnazijalci i studenti sveučilišta iskazuju najveći interes prema prirodoslovnim znanostima, kao što su matematika, fizika i biologija. Veleučilište bilježi skok u interesu za povijest, geografiju te psihologiju, a kategorije u kojima ispitanici strukovnih škola prednjače nemaju znatno veći prosjek odgovora od ispitanika ostalih klastera.

Tablica 2.35: Aritmetičke sredine varijabli klastera interesa - 3 klastera (SAS ispis)

Cluster Means								
Cluster	Povijest	Geografija	Psihologija	Politika	Matematika	Fizika	Informatika	Biologija
1	3.690000000	3.665000000	4.650000000	3.155000000	4.285000000	3.565000000	3.850000000	4.010000000
2	3.230000000	3.460000000	3.320000000	2.860000000	2.970000000	2.680000000	3.640000000	3.030000000
3	4.000000000	4.330000000	5.330000000	2.920000000	3.540000000	2.580000000	3.880000000	3.420000000

Cluster Means							
Cluster	Kemija	Medicina	Pravo	Ekonomija	Elektrotehnika	Robotika	Religija
1	3.310000000	3.675000000	2.530000000	2.875000000	2.480000000	2.580000000	3.210000000
2	2.220000000	2.860000000	2.590000000	3.090000000	2.800000000	2.780000000	2.680000000
3	2.880000000	4.250000000	2.290000000	2.920000000	2.750000000	2.580000000	3.250000000

Pripadni SAS kod je sljedeći:

```
proc fastclus data = podaci out = klaster2 maxclusters = 3
    nomiss maxiter = 300;
    var Povijest Geografija Psihologija Politika Matematika Fizika
    Informatika Biologija Kemija Medicina Pravo Ekonomija
    Elektrotehnika Robotika Religija;
run;

proc print data = klaster2;
    var Skola cluster;
run;
```

2.2.6 Zdravstvene navike

U tablici 2.36 dane su frekvencije korištenja cigareta ovisno o pripadnosti obrazovnoj instituciji. Ono što je zajedničko svakoj od institucija je znatno manji broj pušača u odnosu na nepušače, a najmanje pušača je među gimnazijalcima.

Tablica 2.36: Tablica frekvencija ispitanika ovisno o korištenju cigareta (SAS ispis)

The FREQ Procedure

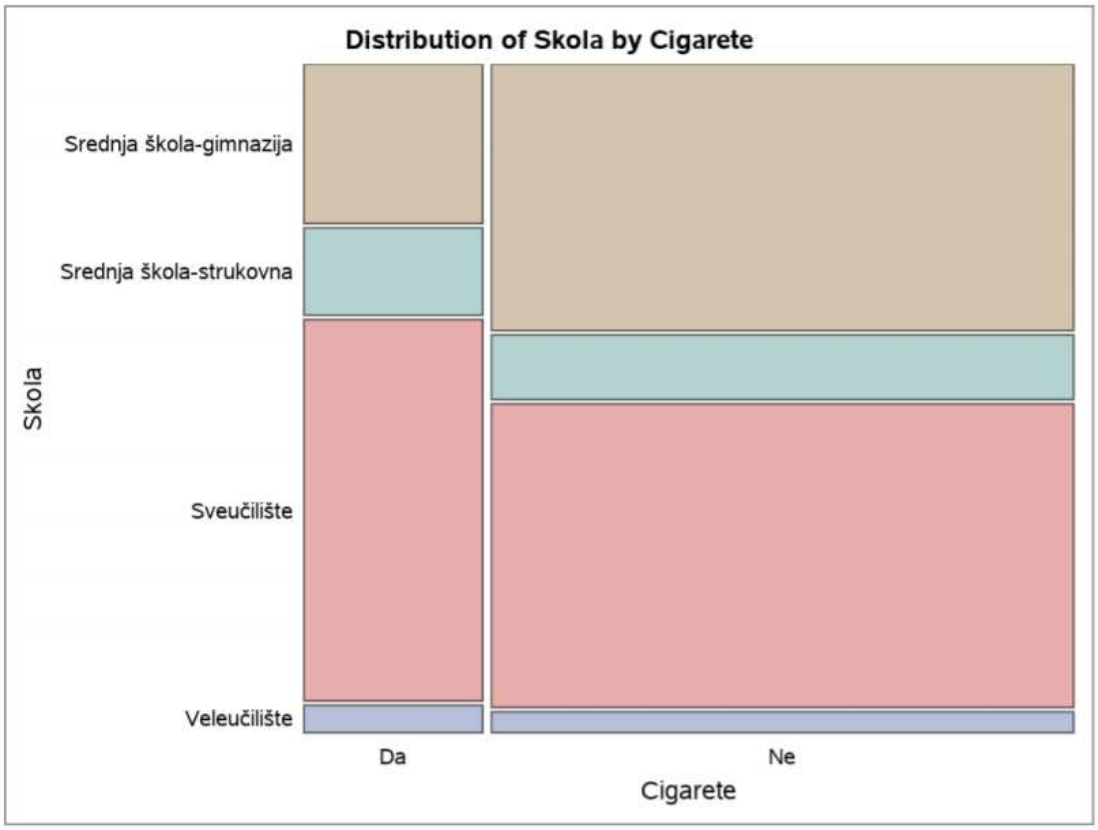
Frequency
Percent
Row Pct
Col Pct

Table of Skola by Cigarete			
Skola(Škola)	Cigarete(Cigarete)		
	Da	Ne	Total
Srednja škola-gimnazija	40 5.72 15.50 24.39	218 31.19 84.50 40.75	258 36.91
Srednja škola-strukovna	22 3.15 29.73 13.41	52 7.44 70.27 9.72	74 10.59
Sveučilište	95 13.59 27.70 57.93	248 35.48 72.30 46.36	343 49.07
Veleučilište	7 1.00 29.17 4.27	17 2.43 70.83 3.18	24 3.43
Total	164 23.46	535 76.54	699 100.00

Mozaik distribucije cigareta prema školama dan je na slici 2.9.

SAS kod kojim su dobiveni rezultati je sljedeći:

```
proc freq data = podatci;
    table skola*cigarete skola*alkohol skola*droga /all plots = mosaic;
    label skola = "Škola";
run;
```



Slika 2.9: Mozaik distribucije cigareta (SAS ispis)

U tablici 2.37 prikazane su frekvencije konzumacije alkohola ovisno o obrazovnoj instituciji. Za razliku od cigareta, u svakoj instituciji znatan broj ispitanika odgovorio je potvrdno. Daleko najveći postotak onih koji konzumiraju alkohol je sa sveučilišta (87, 17%), a najmanji postotak pripada strukovnim školama (66, 22%).

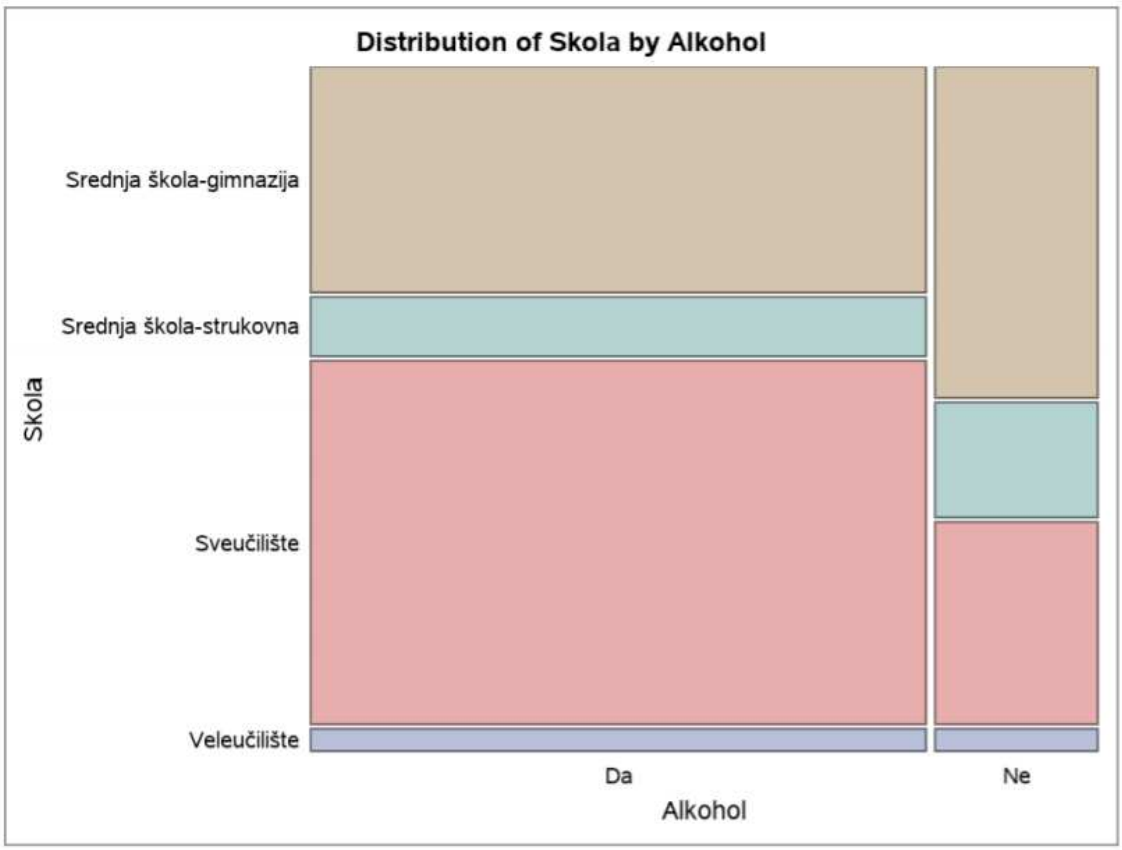
Tablica 2.37: Tablica frekvencija ispitanika ovisno o konzumaciji alkohola (SAS ispis)

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of Skola by Alkohol		
	Alkohol(Alkohol)		
Skola(Škola)	Da	Ne	Total
Srednja škola-gimnazija	186 26.61 72.09 33.63	72 10.30 27.91 49.32	258 36.91
Srednja škola-strukovna	49 7.01 66.22 8.86	25 3.58 33.78 17.12	74 10.59
Sveučilište	299 42.78 87.17 54.07	44 6.29 12.83 30.14	343 49.07
Veleučilište	19 2.72 79.17 3.44	5 0.72 20.83 3.42	24 3.43
Total	553 79.11	146 20.89	699 100.00

Na slici 2.10 dan je mozaik pripadnih frekvencija. Jasno se vidi da je znatno veći broj ispitanika koji konzumiraju alkohol u odnosu na one koji ne konzumiraju. Unatoč tome što je ispitanika s veleučilišta manje (samo 24), čak 79,19% konzumira alkohol barem na društvenim događanjima.

Na dendrogramu 2.12 jasno se vidi podjela na tri klastera. Srednje škole pripadaju istom klasteru dok su sveučilište i veleučilište u zasebnim klasterima. Budući da je veleučilište vanjskom granom povezano s korijenom dendrograma, a sveučilište unutarnjom, može se zaključiti da postoji veća sličnost između srednjoškolaca i studenata sa sveučilišta.



Slika 2.10: Mozaik distribucije alkohola (SAS ispis)

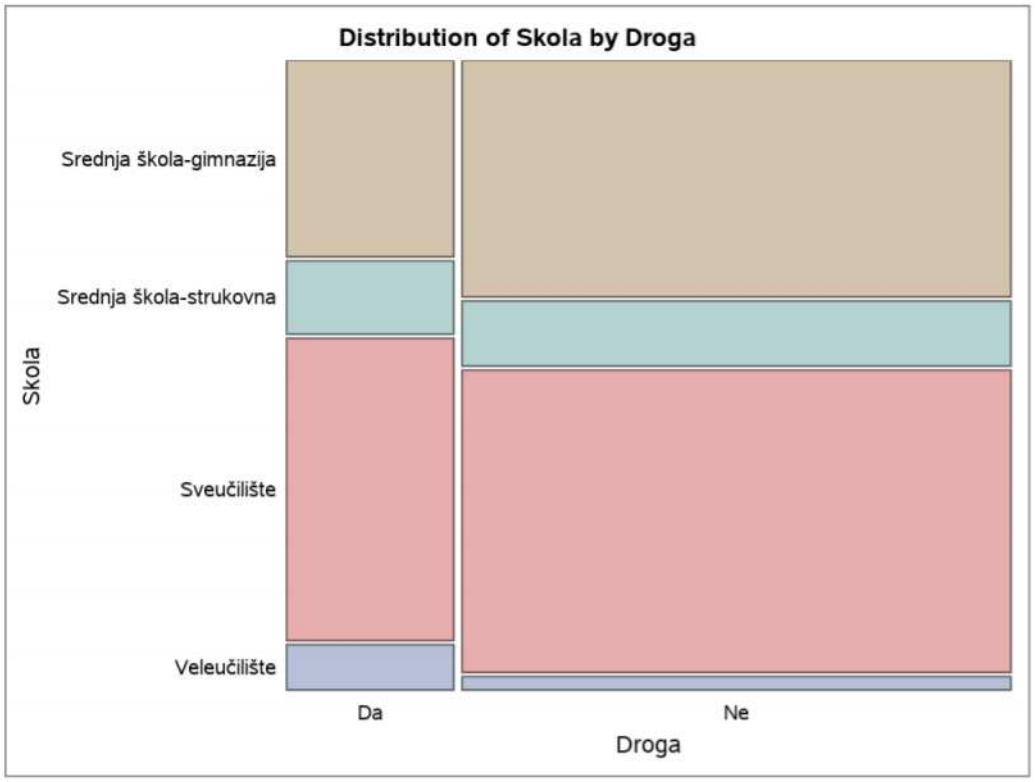
U tablici 2.38 dane su frekvencije konzumacije droga među ispitanicima. Manje od četvrtine ispitanika iz gimnazija, strukovnih škola te sa sveučilišta konzumira droge. Iznenađuje odgovor ispitanika s veleučilišta, gdje je omjer onih koji konzumiraju i ne konzumiraju droge točno 12:12.

Tablica 2.38: Tablica frekvencija ispitanika ovisno o korištenju droga (SAS ispis)

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of Skola by Droga		
	Droga(Droga)		
Skola(Škola)	Da	Ne	Total
Srednja škola-gimnazija	52 7.44 20.16 31.90	206 29.47 79.84 38.43	258 36.91
Srednja škola-strukovna	19 2.72 25.68 11.66	55 7.87 74.32 10.26	74 10.59
Sveučilište	80 11.44 23.32 49.08	263 37.63 76.68 49.07	343 49.07
Veleučilište	12 1.72 50.00 7.36	12 1.72 50.00 2.24	24 3.43
Total	163 23.32	536 76.68	699 100.00

Na slici 2.11 prikazan je odgovarajući mozaik distribucije droga.



Slika 2.11: Mozaik distribucije droga (SAS ispis)

Podatci o zdravstvenim navikama ispitanika dodatno su obrađeni koristeći χ^2-test. Rezultati su sljedeći:

Tablica 2.39: χ^2-test za cigarete (SAS ispis)

Statistics for Table of Skola by Cigarete

Statistic	DF	Value	Prob
Chi-Square	3	14.5783	0.0022
Likelihood Ratio Chi-Square	3	15.2264	0.0016
Mantel-Haenszel Chi-Square	1	11.8261	0.0006
Phi Coefficient		0.1444	
Contingency Coefficient		0.1429	
Cramer's V		0.1444	

Tablica 2.40: χ^2-test za alkohol (SAS ispis)

Statistics for Table of Skola by Alkohol

Statistic	DF	Value	Prob
Chi-Square	3	28.6242	<.0001
Likelihood Ratio Chi-Square	3	28.8711	<.0001
Mantel-Haenszel Chi-Square	1	18.8551	<.0001
Phi Coefficient		0.2024	
Contingency Coefficient		0.1983	
Cramer's V		0.2024	

Tablica 2.41: χ^2-test za droge (SAS ispis)

Statistics for Table of Skola by Droga

Statistic	DF	Value	Prob
Chi-Square	3	11.2289	0.0106
Likelihood Ratio Chi-Square	3	9.7614	0.0207
Mantel-Haenszel Chi-Square	1	3.6615	0.0557
Phi Coefficient		0.1267	
Contingency Coefficient		0.1257	
Cramer's V		0.1267	

Svaki od χ^2-testova ima tri stupnja slobode te su pripadne p-vrijednosti male.

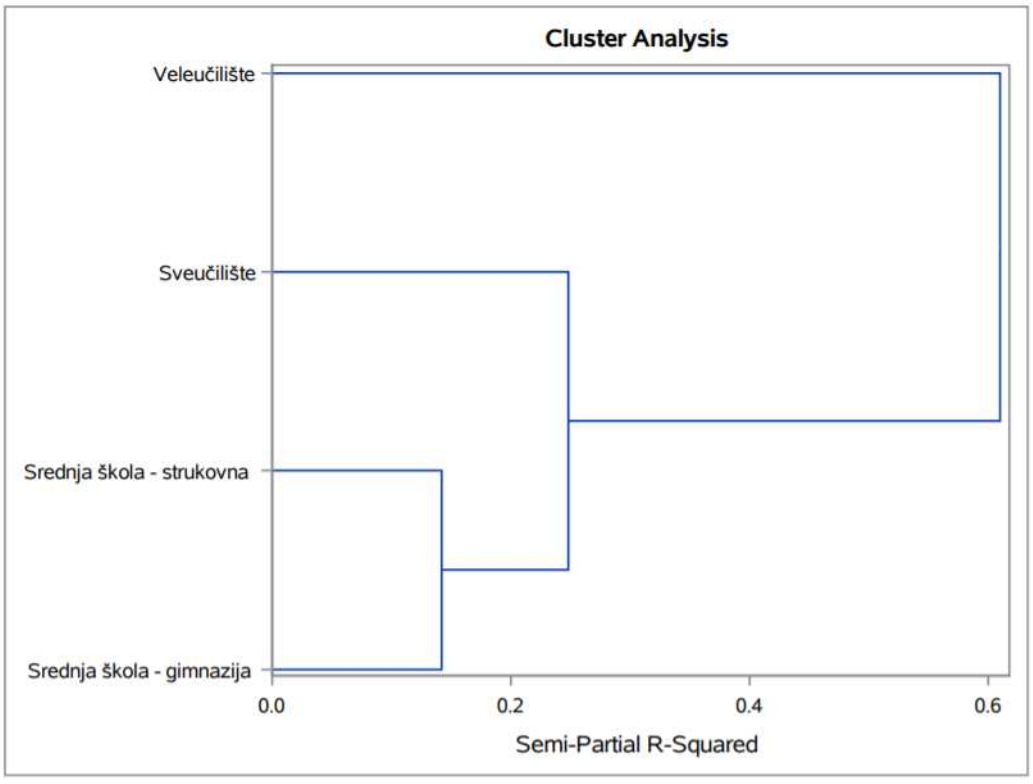
- cigarete: $p = 0.0022$
- alkohol: $p < 0.0001$
- droga: $p = 0.0106$

Dakle, na razini značajnosti $\alpha = 2\%$, za svaku od kategorija možemo odbaciti nultu hipotezu H_0. Razlika između frekvencija ovisno o obrazovnim institucijama je značajna.

Konačno, na podatcima su primijenjeni algoritmi klasteriranja. Hijerarhijskim klasteriranjem uz Ward metodu dobivena je sljedeća podjela:

Tablica 2.42: Podjela na klastere prema zdravstvenim navikama (SAS ispis)

Cluster History						
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Tie
3	Srednja škola - gimnazija	Srednja škola - strukovna	2	0.1420	.858	
2	CL3	Sveučilište	3	0.2482	.610	
1	CL2	Veleučilište	4	0.6099	.000	



Slika 2.12: Dendrogram ispitanika prema podatcima o zdravstvenim navikama (SAS ispis)

Pripadni SAS kod je sljedeći:

```
proc distance data = podaci out = DIST method = Euclid;
    var interval(Cigarete Alkohol Droga);
    id Skola;
run;

proc cluster data = DIST method = Ward outtree = Tree;
    id Skola;
run;
```

Nadalje, na podatcima je pimijenjeno k-means klasteriranje euklidskom mjerom uz ograničenje na maksimalno dva klastera. Dobivena je sljedeća podjela:

Tablica 2.43: Podjela ispitanika na dva klastera prema podatcima o interesima (SAS ispis)

Obs	Skola	CLUSTER
1	Srednja škola - gimnazija	1
2	Srednja škola - strukovna	1
3	Sveučilište	1
4	Veleučilište	2

Gimnazija, strukovna škola te sveučilište pripadaju istom klasteru, a veleučilište je u zasebnom. Analogna podjela dobila bi se „rezanjem" dendrograma 2.12 na razini jedan. U tablici 2.44 dane su aritmetičke sredine odgovora prema klasterima.

Tablica 2.44: Aritmetičke sredine varijabli klastera zdravstvenih navika - 2 klastera (SAS ispis)

Cluster Means			
Cluster	**Cigarete**	**Alkohol**	**Droga**
1	24.31000000	75.16000000	23.05333333
2	29.17000000	79.17000000	50.00000000

Pripadni SAS kod je sljedeći:

```
proc fastclus data = podatci out = klaster2 maxclusters = 2
    nomiss maxiter = 300;
    var Cigarete Alkohol Droga;
run;

proc print data = klaster2;
    var Skola cluster;
run;
```

Dodatno, na podatcima je primijenjen k-means algoritam koristeći euklidsku mjeru sličnosti uz ograničenje na maksimalno tri klastera. Dobivena je sljedeća podjela:

Tablica 2.45: Podjela ispitanika na tri klastera prema podatcima o interesima (SAS ispis)

Obs	Skola	CLUSTER
1	Srednja škola - gimnazija	2
2	Srednja škola - strukovna	2
3	Sveučilište	3
4	Veleučilište	1

Kao i na dendrogramu 2.12, gimnazija i strukovna škola pripadaju istom klasteru, dok su sveučilište i veleučilište u zasebnim klasterima. U tablici 2.46 dane su aritmetičke sredine odgovora prema klasterima.

Tablica 2.46: Aritmetičke sredine varijabli klastera zdravstvenih navika - 3 klastera (SAS ispis)

Cluster Means			
Cluster	Cigarete	Alkohol	Droga
1	29.17000000	79.17000000	50.00000000
2	22.61500000	69.15500000	22.92000000
3	27.70000000	87.17000000	23.32000000

Pripadni SAS kod je sljedeći:

```
proc fastclus data = podatci out = klaster2 maxclusters = 3
    nomiss maxiter = 300;
    var Cigarete Alkohol Droga;
run;

proc print data = klaster2;
    var Skola cluster;
run;
```

Poglavlje 3

Zaključak

Po uzoru na istraživanje [7] pripremljena je anketa s ciljem prikupljanja podataka o mladim ljudima diljem Hrvatske. U periodu od četiri mjeseca 700 ispitanika ispunilo je anketu, a nakon čišćenja podataka, njih 699 ulazi u daljnju analizu. Odgovori su grupirani u šest kategorija: Glazba, Film, Fobije, Hobiji, Interesi i Zdravstvene navike. Na svakoj od kategorija napravljeno je hijerarhijsko, a potom i nehijerarhijsko klasteriranje kako bi se odredilo na koji način se ispitanici grupiraju ovisno kojoj obrazovnoj instituciji pripadaju (gimnazija, strukovna škola, sveučilište, veleučilište).

Na skupu podataka iz kategorije Glazba primijenjeno je hijerarhijsko klasteriranje koristeći 4 metode: metodu maksimuma, minimuma, prosjeka te Ward metodu. Za svaku od metoda sveučilište i veleučilište su u jednom klasteru, a srednje škole u drugom, osim kod metode minimuma gdje su srednje škole odvojene. Na ostalim kategorijama primijenjena je samo Ward metoda. Što se tiče filmova, fakulteti su ponovno u zajedničkom klasteru, a prema visini zajedničkih čvorova na dendrogarmu, gimnazijalci su im sličniji po odgovorima. U kategoriji Fobije, najviše se razlikuje veleučilište, a sveučilište i strukovna škola dio su istog klastera. Analizom hobija, sveučilište i gimnazija imaju najveću sličnost, a ispitanici strukovnih škola najviše se razlikuju. Interesi ispitanika postižu jednaku raspodjelu kao i hobiji. Rezultati su prikazani u tablicama 3.1 i 3.2.

Tablica 3.1: Rezultati hijerarhijskog klasteriranja raznim metodama za kategoriju glazba

	gimnazija	strukovna škola	sveučilište	veleučilište
Ward metoda	CL2	CL2	CL1	CL1
Metoda minimuma	CL2	CL3	CL1	CL1
Metoda maksimuma	CL2	CL2	CL1	CL1
Metoda prosjeka	CL2	CL2	CL1	CL1

Tablica 3.2: Rezultati hijerarhijskog klasteriranja po kategorijama - Ward metoda

	gimnazija	strukovna škola	sveučilište	veleučilište
Film	CL2	CL3	CL1	CL1
Fobije	CL2	CL1	CL1	CL3
Hobiji	CL1	CL3	CL1	CL2
Interesi	CL1	CL3	CL1	CL2
Zdravstvene navike	CL1	CL1	CL2	CL3

Nakon hijerarhijskog klasteriranja, napravljeno je k-means klasteriranje uz ograničenja na dva, a potom i na tri klastera. U kategoriji Glazba, ograničenjem na dva klastera srednje škole su grupirane u jedan klaster, a fakulteti u drugi. Kada se broj klastera povećao na tri, fakulteti su se razdvojili u zasebne klastere. U kategoriji Film, ograničenjem na dva klastera strukovna škola se izdvojila u jedan klaster, a ograničenjem na tri javlja se razlika između fakulteta i gimnazije. Analizom kategorije Fobije, kao i hijerarhijskim klasteriranjem veleučilište se odvaja u poseban klaster. Postavljanjem ograničenja na 3 klastera, sveučilište i strukovna škola ostaju dio istog klastera, kao što je i sugerirano dendrogramom. U kategorijama Hobiji i Interesi, strukovna škola se najviše razlikuje i u oba slučaja pripada zasebnom klasteru, a povećanjem broja klastera, veleučilište se također izdvaja.

Na zdravstvenim navikama ispitanika prvo je primijenjena analiza frekvencija posebno za cigarete, alkohol i drogu, a χ^2-testovi pokazali su da postoji značajna razlika ovisno o obrazovnoj instituciji kojoj ispitanici pripadaju. Hijerarhijskim klasteriranjem srednje škole su se grupirale u jedan klaster, studenti sveučilišta su im najsličniji po odgovorima, a veleučilište najviše odstupa. K-means algoritmom veleučilište se najviše razlikuje i kada se odaberu dva i kada se odaberu tri klastera. Zanimljivo je da za razliku od dendrograma, odabirom tri klastera veleučilište je sličnije po odgovorima srednjim školama nego sveučilište. Ukupni rezultati k-means klasteriranja dani su u tablicama 3.3 i 3.4.

Tablica 3.3: Rezultati k-means klasteriranja po kategorijama - 2 klastera

	gimnazija	strukovna škola	sveučilište	veleučilište
Glazba	CL1	CL1	CL2	CL2
Film	CL1	CL2	CL1	CL1
Fobije	CL1	CL1	CL1	CL2
Hobiji	CL1	CL2	CL1	CL1
Interesi	CL1	CL2	CL1	CL1
Zdravstvene navike	CL1	CL1	CL1	CL2

Tablica 3.4: Rezultati k-means klasteriranja po kategorijama - 3 klastera

	gimnazija	strukovna škola	sveučilište	veleučilište
Glazba	CL1	CL2	CL3	CL3
Film	CL1	CL2	CL3	CL3
Fobije	CL1	CL3	CL3	CL2
Hobiji	CL1	CL2	CL1	CL3
Interesi	CL1	CL2	CL1	CL3
Zdravstvene navike	CL2	CL2	CL3	CL1

Dakle, gotovo za svaku od kategorija hijerarhijsko klasteriranje Ward metodom pružilo je dobar uvid u grupiranje podataka, odnosno k-means klasteriranjem takva je grupacija ponovljena. Moglo se pretpostaviti da će se odgovori ispitanika srodnijih institucija poklapati, kao što je slučaj s kategorijama Hobiji i Interesi. Glazbeni ukus i ukus u filmovima razlikuje se između generacija, a slično je i sa zdravstvenim navikama. Može se zaključiti da su odgovori ispitanika dosljedni i nema razloga sumnjati u njihovu iskrenost.

Poglavlje 4

Anketa

U nastavku slijedi anketa kojom su prikupljeni podatci za izradu rada.

4.1 Istraživanje interesa studenata i srednjoškolaca

Poštovani/Poštovana,

Hvala što ste odvojili vrijeme za sudjelovanje u ovom istraživanju. Ova je anketa dio istraživanja kojim se ispituju interesi, hobiji, zdravstvene navike i fobije srednjoškolaca i studenata. Istraživanje će poslužiti za izradu diplomskih radova Matee Mijatović i Petre Vlaić, studentica na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta u Zagrebu. U anketi nema točnih i netočnih odgovora, zanimaju nas Vaši osobni interesi. Molimo da samostalno označite odgovore koji nabolje odražavaju Vaša mišljenja te da odgovorite na sva pitanja u anketi. Ispunjavanje ankete traje oko 7 minuta i potpuno je anonimno. Znači, nećemo bilježiti tko ste, a Vaši odgovori neće se moći povezati s konkretnom osobom. Molimo Vas da iskreno odgovorite na sva pitanja. Ako poželite, u bilo kojem trenutku možete odustati od daljnjeg ispunjavanja ankete.

Nastavkom potvrđujete da ste pročitali ovaj uvod te da ste obaviješteni o svrsi i postupku ovog istraživanja. Još jednom hvala za Vaš vrijedan doprinos.

4.1.1 Osobni podatci

- Spol: M / Ž
- Godina rođenja (gggg): .....
- Regija Hrvatske u kojoj ste odrasli:

  - Centralna Hrvatska
  - Istočna Hrvatska
  - Planinska Hrvatska
  - Sjeverna Hrvatska obala
  - Južna Hrvatska / Južna Hrvatska obala
- Mjesto u kojem ste odrasli: .....
- Mjesto trenutnog školovanja (npr. Zagreb): .....
- Trenutno pohađam:
  - Srednja škola-gimnazija
  - Srednja škola-strukovna
  - Veleučilište
  - Sveučilište
- Pohađate li privatnu školu/fakultet? DA / NE
- Ako ste student, naznačite kojim se područjem znanosti bavite:
  - Biološke znanosti
  - Društvene znanosti
  - Humanističke znanosti
  - Medicinske znanosti
  - Poljoprivredne znanosti
  - Prirodne znanosti
  - Tehničke znanosti i tehnologija
  - Umjetničke znanosti
- Status veze roditelja:
  - Bez roditelja
  - Razvedeni
  - Samohrani roditelj
  - Udovac/ica
  - Vanbračna zajednica

  - Vjenčani

- Broj djece u obitelji zajedno s Vama (odgovor napišite brojem): .....
- Broj kućnih ljubimaca (odgovor napišite brojem): .....
- Jeste li zaposleni dok se školujete? DA / NE

4.1.2 Glazba

Sljedeći niz pitanja ispituje Vaše interese. Molimo za svako pitanje označite odgovarajućim brojem koliko uživate u određenoj vrsti glazbe, pri čemu je: 1 - ne slušam uopće, 7 - jako volim slušati.

- Volite li slušati glazbu? 1 2 3 4 5 6 7
- Pop: 1 2 3 4 5 6 7
- Rock: 1 2 3 4 5 6 7
- Metal: 1 2 3 4 5 6 7
- Hip hop, Rap: 1 2 3 4 5 6 7
- Jazz: 1 2 3 4 5 6 7
- Narodnjaci: 1 2 3 4 5 6 7
- Techno: 1 2 3 4 5 6 7
- Klasična glazba: 1 2 3 4 5 6 7
- Trash: 1 2 3 4 5 6 7
- House: 1 2 3 4 5 6 7

4.1.3 Film

Molimo za svako pitanje označite odgovarajućim brojem koliko uživate u određenoj vrsti filmova, pri čemu je: 1 - ne gledam uopće, 7 - jako volim gledati.

- Volite li gledati filmove? 1 2 3 4 5 6 7
- Horor: 1 2 3 4 5 6 7
- Triler: 1 2 3 4 5 6 7
- Komedija: 1 2 3 4 5 6 7
- Romantični: 1 2 3 4 5 6 7
- SCI-FI: 1 2 3 4 5 6 7
- Dokumentarni: 1 2 3 4 5 6 7
- Akcijski: 1 2 3 4 5 6 7
- Animirani: 1 2 3 4 5 6 7
- Western: 1 2 3 4 5 6 7

4.1.4 Fobije

Sljedeći niz pitanja ispituje Vaše fobije. Molimo, za svako pitanje označite Vaš odgovor brojem, pri čemu je: 1-nemam strah od navedenog, 7-imam veliki strah od navedenog.

- Let avionom: 1 2 3 4 5 6 7
- Visina: 1 2 3 4 5 6 7
- Vremenske nepogode (grmljavina, oluja, . . .): 1 2 3 4 5 6 7
- Mrak: 1 2 3 4 5 6 7
- Pauci: 1 2 3 4 5 6 7
- Dizalo: 1 2 3 4 5 6 7
- Zmije: 1 2 3 4 5 6 7

- Mali prostori: 1 2 3 4 5 6 7
- Psi: 1 2 3 4 5 6 7
- Javni govor: 1 2 3 4 5 6 7
- Glodavci (štakori, miševi, . . .): 1 2 3 4 5 6 7
- Zubar: 1 2 3 4 5 6 7
- Doktorske igle: 1 2 3 4 5 6 7
- Krv: 1 2 3 4 5 6 7
- Bacili: 1 2 3 4 5 6 7

4.1.5 Hobiji i interesi

Sljedeći niz pitanja ispituje Vaše hobije. Molimo, za svako pitanje označite odgovor brojem 1-7, ovisno o Vašoj uključenosti u sljedeće aktivnosti, pri čemu je: 1- nikad ovo ne radim, 7- redovito se ovim bavim.

- Čitanje: 1 2 3 4 5 6 7
- Strani jezici: 1 2 3 4 5 6 7
- Likovna umjetnost: 1 2 3 4 5 6 7
- Gluma: 1 2 3 4 5 6 7
- Ples: 1 2 3 4 5 6 7
- Sviranje: 1 2 3 4 5 6 7
- Pjevanje: 1 2 3 4 5 6 7
- Koncerti: 1 2 3 4 5 6 7
- Kazalište: 1 2 3 4 5 6 7
- Video igrice: 1 2 3 4 5 6 7
- Automobili: 1 2 3 4 5 6 7

- Shopping: 1 2 3 4 5 6 7
- Profesionalno bavljenje sportom: 1 2 3 4 5 6 7
- Amatersko bavljenje sportom: 1 2 3 4 5 6 7

Molimo, za svako pitanje označite odgovor brojem 1-7, ovisno o Vašoj uključenosti u sljedeće aktivnosti, pri čemu je: 1- ne zanima me, 7- jako me zanima.

- Povijest: 1 2 3 4 5 6 7
- Geografija: 1 2 3 4 5 6 7
- Psihologija: 1 2 3 4 5 6 7
- Politika: 1 2 3 4 5 6 7
- Matematika: 1 2 3 4 5 6 7
- Fizika: 1 2 3 4 5 6 7
- Informatika: 1 2 3 4 5 6 7
- Biologija: 1 2 3 4 5 6 7
- Kemija: 1 2 3 4 5 6 7
- Medicina: 1 2 3 4 5 6 7
- Pravo: 1 2 3 4 5 6 7
- Ekonomija: 1 2 3 4 5 6 7
- Elektrotehnika: 1 2 3 4 5 6 7
- Robotika: 1 2 3 4 5 6 7
- Religija: 1 2 3 4 5 6 7

4.1.6 Zdravstvene navike

Sljedeći (ujedno i posljednji) niz pitanja ispituje Vašu svijest o zdravlju. Molimo odgovorite iskreno.

- Pušite li cigarete? DA / NE
- U prosjeku, koliko dnevno cigareta popušite?
  - Ne pušim cigarete
  - 1-5
  - 6-10
  - 11-20
  - Više od kutije dnevno
- Pijete li alkohol? DA / NE
- Koliko često pijete alkohol?
  - Nikad
  - Na društvenim događanjima
  - Samo vikendom
  - Skoro svakodnevno
- Jeste li ikad konzumirali drogu? DA / NE
- Koliko učestalo konzumirate drogu?
  - Nikad
  - Probao/la
  - U prosjeku jednom tjedno
  - Par puta tjedno
  - Skoro svakodnevno
- Ako konzumirate, u koju skupinu najviše ulazite:
  - Halucinogene droge (marihuana, hašiš, LSD, . . .)
  - Opojne droge (morfij, kodein, heroin, metadon, . . .)
  - Stimulativne droge (kokain i crack, MDMA-ecstasy, . . .)

Hvala na sudjelovanju!

Bibliografija

[1] Michael R. Anderberg, *Cluster analysis for applications*, sv. 1, Academic Press.

[2] B. S. Everitt, S. Landau, Leese M. i D. Stahl, *Cluster Analysis*, Wiley, 2011.

[3] L. P. Fávero i P. Belfiore, *Data Science for Business and Decision Making*, sv. 11, Elsevier.

[4] T. Hastie, R. Tibshirani i Friedman J., *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, 2008.

[5] A. Periklis, *Data Clustering Techniques*, Rapport technique, University of Toronto, Department of Computer Science (2002), http://www.cs.toronto.edu/~periklis/pubs/depth.pdf.

[6] SequentiX Digital DNA Processing, *GelQuest-DNA Fingerprint Analysis Software*, https://www.sequentix.de/gelquest/help/distance_measures.htm.

[7] FSEV UK, *Young people survey*, https://www.kaggle.com/miroslavsabo/young-people-survey.

[8] T. Ungaro, *Klasterska analiza*, Prirodoslovno-matematički fakultet, Matematički odsjek (2016).

[9] R. Xu, *Survey of Clustering Algorithms*, (2005), http://axon.cs.byu.edu/Dan/678/papers/Cluster/Xu.pdf.

Sažetak

Klasterska analiza metoda je grupiranja objekata u klase tako da su u istoj klasi objekti najveće sličnosti.

U ovom radu, klasterska analiza primijenjena je na podatcima o interesima, hobijima i zdravstvenim navikama studenata i srednjoškolaca. Podatci su prikupljeni koristeći online anketu koja je distribuirana putem društvenih mreža. Konačan broj analiziranih ispitanika je bio 699, pri čemu je omjer studenata i srednjoškolaca ravnomjeran. Za primjenu metode, potrebno je izabrati odgovarajuću mjeru sličnosti i algoritam ovisno o vrsti objekata, svojstvima algoritama te konačnim ciljevima istraživanja.

Klasterski algoritmi mogu biti hijerarhijski ili nehijerarhijski (partitivni). Hijerarhijski algoritam je primijenjen koristeći Ward metodu, a rezultati su prikazani dendrogramom. Kao partitivni algoritam izabran je k-means s konačnim ograničenjima na dva, a potom i na tri klastera, dok je mjera sličnosti euklidska.

Summary

Cluster analysis is a method of grouping objects into classes so that the objects with the highest similarity are in the same class.

In this paper, cluster analysis has been applied on a data set which consists of interests, hobbies and health habits of students and high school pupils. The data set was collected using an online survey which was distributed through social media. The final number of analyzed respondents was 699, having the ratio between students and high school pupils relatively equal. In order to apply the method, it is necessary to choose an adequate similarity measure and an algorithm depending on the type of data, features of the algorithm and main goals of the research.

Clustering algorithms are either hierarchical or non-hierarchical (partitive). A hierarchical algorithm has been applied using the Ward method, and the results were shown on dendrograms. As for the type of partitive algorithm, the k-means algorithm was chosen, limiting the final number of clusters on two and on three, with the euclidean similarity measure.

Životopis

Rođena sam 30.09.1995. u Zagrebu. Završila sam OŠ „Split 3" i IV. gimnaziju „Marko Marulić" u Splitu. 2014. upisala sam preddiplomski studij Matematika na Prirodoslovno-matematičkom fakultetu u Zagrebu na kojem sam 2017. godine stekla titulu univ.bacc.math. Te godine upisujem diplomski studij Matematička statistika na istom fakultetu.