UNIVERZITET U NOVOM SADU

KONSTRUKCIJA I ANALIZA KLASTER ALGORITMA SA PRIMENOM U DEFINISANJU BIHEJVIORALNIH FAKTORA RIZIKA U POPULACIJI ODRASLOG STANOVNIŠTVA SRBIJE

DOKTORSKA DISERTACIJA

Mentor: Prof. dr Zorana Lužanin
Prof. dr Eržebet Ač Nikolić

Kandidat: Nataša Dragnić

Novi Sad, 2015. godine

Univerzitet u Novom Sadu

Ključna dokumentacijska informacija

Redni broj: RBR	
Identifikacioni broj: IBR	
Tip dokumentacije: TD	Monografska dokumentacija
Tip zapisa: TZ	Tekstualni štampani materijal
Vrsta rada: VR	Doktorska disertacija
Ime i prezime autora: AU	Nataša Dragnić
Mentor: MN	Prof.dr Zorana Lužanin, redovni profesor Prof. dr Eržebet Ač Nikolić, redovni profesor
Naslov rada: NR	Konstrukcija i analiza klaster algoritma sa primenom u definisanju bihejvioralnih faktora rizika u populaciji odraslog stanovništva Srbije
Jezik publikacije: JP	srpski (latinica)
Jezik izvoda: JI	srpski/engleski
Zemlja publikovanja: ZP	Republika Srbija
Uže geografsko područje: UGP	Vojvodina
Godina: GO	2015
Izdavač: IZ	autorski reprint
Mesto i adresa: MA	21000 Novi Sad, Srbija, Zorana Đinđića 1

Fizički opis rada: FO	Broj poglavlja 10/ stranica 154/ tabela 25/ slika 3/ grafikona 2/ referenci 272/ priloga 3
Naučna oblast: NO	Matematika; Medicina
Naučna disciplina: ND	Matematičko modeliranje; Javno zdravlje
Predmetna odrednica, ključne reči: PO	Klaster analiza; algoritmi; neglatka optimizacija; složenost; kategorijalni podaci; bihejvioralna istraživanja; faktori rizika; odrasli; socioekonomski faktori; demografija
UDK	
Čuva se: ČU	U Centralnoj biblioteci Univerziteta u Novom Sadu, 21000 Novi Sad, Dr Zorana Đinđića 1
Važna napomena: VN	
Izvod: IZ	Klaster analiza ima dugu istoriju i mada se primenjuje u mnogim oblastima i dalje ostaju značajni izazovi. U disertaciji je prikazan uvod u neglatki optimizacioni pristup u klasterovanju, sa osvrtom na problem klasterovanja velikih skupova podataka. Međutim, ovi optimizacioni algoritmi bolje funkcionišu u radu sa neprekidnim podacima. Jedan od glavnih izazova u klaster analizi je rad sa velikim skupovima podataka sa kategorijalnim i kombinovanim (numerički i kategorijalni) tipovima promenljivih. Rad sa velikim brojem instanci (objekata) i velikim brojem dimenzija (promenljivih), može predstavljati problem u klaster analizi, zbog vremenske složenosti. Jedan od načina rešavanja ovog problema je redukovanje broja instanci, bez gubitka informacija. Prvi cilj disertacije je bio upoređivanje rezultata klasterovanja na celom skupu i prostim slučajnim uzorcima sa kategorijalnim i kombinovanim podacima, za različite veličine uzorka i različit broj klastera. Nije utvrđena značajna razlika ($p>0.05$) u rezultatima klasterovanja na uzorcima obima $0.03m, 0.05m, 0.1m, 0.3m$ (gde je m obim posmatranog skupa) i celom skupu. Drugi cilj disertacije je bio konstrukcija

	efikasnog postupka klasterovanja velikih skupova podataka sa kategorijalnim i kombinovanim tipovima promenljivih. Predloženi postupak se sastoji iz sledećih koraka: 1. klasterovanje na prostim slučajnim uzorcima određene kardinalnosti; 2. određivanje najboljeg klasterskog rešenja na uzorku, primenom odgovarajućeg kriterijuma validnosti; 3. dobijeni centri klastera iz ovog uzorka služe za klasterovanje ostatka skupa. Treći cilj disertacije predstavlja primenu klaster analize u definisanju klastera bihejvioralnih faktora rizika u populaciji odraslog stanovništva Srbije, kao i analizu sociodemografskih karakteristika dobijenih klastera. Klaster analiza je primenjena na velikom reprezentativnom uzorku odraslog stanovništva Srbije, starosti 20 i više godina. Izdvojeno je pet jasno odvojenih klastera sa karakterističnim kombinacijama bihejvioralnih faktora rizika: *Bez rizičnih faktora*, *Štetna upotreba alkohola i druge rizične navike*, *Nepravilna ishrana i druge rizične navike*, *Nedovoljna fizička aktivnost*, *Pušenje*. Rezultati multinomnog logističkog regresionog modela ukazuju da ispitanici koji nisu u braku, lošijeg su materijalnog stanja, nižeg obrazovanja i žive u Vojvodini imaju veću šansu za prisustvo višestrukih bihejvioralnih faktora rizika.
Datum prihvatanja teme od strane Senata: DP	11.11.2010.
Datum odbrane: DO	
Članovi komisije: (ime i prezime / titula / zvanje / naziv organizacije / status) KO	predsednik: član: član:

University of Novi Sad

Key word documentation

Accession number: ANO	
Identification number: INO	
Document type: DT	Monograph documentation
Type of record: TR	Textual printed material
Contents code: CC	Ph.D.thesis
Author: AU	Nataša Dragnić
Mentor: MN	Zorana Lužanin, PhD, full profesor Eržebet Ač Nikolić, PhD, full profesor
Title: TI	Construction and analysis of cluster algorithm with application in defining behavioural risk factors in Serbian adult population
Language of text: LT	Serbian
Language of abstract: LA	English / Serbian
Country of publication: CP	Republic of Serbia
Locality of publication: LP	Vojvodina
Publication year: PY	2015
Publisher: PU	Author reprint
Publication place: PP	21000 Novi Sad, Serbia, Dr Zorana Đinđića 1

Physical description: PD	Chapters 10 / pages 154 / tables 25/ pictures 3 / graphics 2 / references 272 / supplements 3
Scientific field SF	Mathematics; Medicine
Scientific discipline SD	Mathematical Modeling; Public Health
Subject, Key words SKW	Cluster Analysis; Algorithms; Nonsmooth optimization; Complexity; Categorical data; Behavioural research, Risk factors, Adult, Socioeconomic factors, Demography
UC	
Holding data: HD	The University of Novi Sad Central Library, Dr Zorana Đinđića 1
Note: N	
Abstract: AB	The cluster analysis has a long history and a large number of clustering techniques have been developed in many areas, however, significant challenges still remain. In this thesis we have provided a introduction to nonsmooth optimization approach to clustering with reference to clustering large datasets. Nevertheless, these optimization clustering algorithms work much better when a dataset contains only vectors with continuous features. One of the main challenges is clustering of large datasets with categorical and mixed (numerical and categorical) data. Clustering deals with a large number of instances (objects) and a large number of dimensions (variables) can be problematic because of time complexity. One of the ways to solve this problem is by reducing the number of instances, without the loss of information. The first aim of this thesis was to compare the results of cluster algorithms on the whole dataset and on simple random samples with categorical and mixed data, in terms of validity, for different number of clusters and for different sample sizes. There were no significant differences ($p>0.05$) between the obtained results on the samples of the size of $0.03m, 0.05m, 0.1m, 0.3m$ (where m is the size of

	the dataset) and the whole dataset. The second aim of this thesis was to develop an efficient clustering procedure for large datasets with categorical and mixed (numeric and categorical) values. The proposed procedure consists of the following steps: 1. clustering on simple random samples of a given cardinality; 2. finding the best cluster solution on a sample (by appropriate validity measure); 3. using cluster centers from this sample for clustering of the remaining data. The third aim of this thesis was to examine clustering of four lifestyle risk factors and to examine the variation across different socio-demographic groups in a Serbian adult population. Cluster analysis was carried out on a large representative sample of Serbian adults aged 20 and over. We identified five homogenous health behaviour clusters with specific combination of risk factors: 'No Risk Behaviours', 'Drinkers with Risk Behaviours', 'Unhealthy diet with Risk Behaviours', 'Smoking'. Results of multinomial logistic regression indicated that single adults, less educated, with low socio-economic status and living in the region of Vojvodina are most likely to be a part of the clusters with a high-risk profile.
Accepted on Senate on: AS	11.11.2010.
Defended: DE	
Thesis Defend Board: DB	president: member: member:

SADRŽAJ

LISTA OZNAKA

$\mathbb{R}$	skup realnih brojeva
$\mathbb{R}_+$	skup nenegativnih realnih brojeva
$\mathbb{R}^n$	$n-$ dimenzionalni Euklidov prostor
$\arg\min f(x)$	tačka u kojoj funkcija f dostiže minimum
m	obim posmatranog skupa
n	broj promenljivih u datom skupu
$n^{(1)}$, $n^{(2)}$	broj neprekidnih, odnosno kategorijalnih promenljivih redom ($n^{(1)}+n^{(2)}=n$)
k	broj klastera
A^j	j-ti klaster skupa A, $A=\bigcup_{j=1}^{k} A^j$
m_j	kardinalnost klastera A^j
$x^i \in \mathbb{R}^n$	centar i-tog klastera, $x^i=\frac{1}{m_i}\sum_{a\in A^i} a$, za numerička neprekidna obeležja
w_{ij}	težinski koeficijent za objekat a^i i j-ti klaster
$z_i \in \mathbb{R}^n$	centar i-tog klastera
$d(x,y)$	mera rastojanja (različitosti) između dva objekta x i y
d_{ij}	mera rastojanja (različitosti, metrike) između dva objekta x_i i x_j, $d_{ij}=d(x_i,x_j)$ $i,j=1,...,n$
s_{ij}	mera sličnosti dva objekta x_i i x_j; $s_{ij}=s(x_i,x_j)=1-d_{ij}$
$d(x,X)$	rastojanje tačke x od skupa X
$D_1,\ldots,D_n$	skup kategorijalnih obeležja
n_j	broj kategorija kategorijalnog obeležja D_j
$dom\,(D_i)$	domen kategorijalnog obeležja D_j, $dom\,(D_j)=\left\{d_j^{(1)},d_j^{(2)},\ldots,d_j^{(n_j)}\right\}$
m_u	obim prostog slučajnog uzorka, izabranog iz skupa kardinalnosti m

t_u	koeficijent (količnik veličine osnovnog skupa i uzorka), $t_u = m/m_u$, $t_u \in (0,1)$
$A^{(k)}$	k-particija dobijena primenom klaster algoritma na celom skupu, $A^{(k)} = \{A^1, A^2, \ldots, A^k\}$
$C_{m_u}^{(k)}$	k-particija dobijena primenom klaster algoritma na uzorku veličine m_u
$\rightarrow$	operator dodeljivanja ($i \rightarrow i+1$ i „uzima novu vrednost" $i+1$)
$f(m) \in O(g(m))$	vremenska (prostorna) složenost algoritma $f(m)$ je reda $g(m)$
$\propto$	ekvivalentnost (međusobna svodljivost) dva problema ē ′ ÷ i ē (ē ′ ∝ ē)

LISTA TABELA

LISTA SLIKA

LISTA GRAFIKONA

1. UVOD

Istraživači se često sreću sa situacijama koje su najbolje rešene definisanjem grupa homogenih objekata, bez obzira da li su u pitanju objekti, ispitanici ili drugo. Potreba za identifikovanjem grupa unutar populacije se sreće u mnogim oblastima, pri čemu je cilj otkrivanje prirodne strukture između observacija. Najčešće korišćena multivarijantna metoda je klaster analiza (eng. *cluster*[1] *analysis*), ili analiza grupisanja. Zadatak ove metode je da maksimizira internu homogenost (unutar klastera) i eksternu heterogenost (između klastera). Problem traženja optimalne particije skupa predstavlja problem globalne optimizacije.

1.1. KLASTER ANALIZA: DEFINICIJA, OSOBINE

"Razumevanje sveta u kome živimo zahteva konceptualizaciju sličnosti i razlika između elemenata koji ga čine" [2]

Klaster analiza predstavlja grupu multivarijantnih tehnika čija je osnovna svrha grupisanje objekata (ispitanika, proizvoda ili drugih objekata) na osnovu njihovih karakteristika. Grupisanje objekata se vrši na taj način da je svaki objekat veoma sličan drugima u klasteru, a nastale grupe treba da imaju osobinu interne homogenosti unutar klastera i visoke eksterne (između klastera) različitosti. Bez obzira koja definicija je u pitanju, generalno važi da je cilj klasterovanja identifikacija „prirodne" strukture u skupu podataka .

Klaster analiza ima značajnu primenu u različitim oblastima, kao što su: inženjerstvo, biologija, medicina, psihologija, sociologija, statistika, astrofizika, ekonomija, marketing, prepoznavanje oblika, istraživanje podataka (eng. *data mining)*, mašinsko učenje, radarsko skeniranje, planiranje razvoja i drugo. U literaturi se sreće pod različitim nazivima u različitim kontekstima, kao što su *nenadgledano učenje* ili *učenje bez učitelja* (eng. *unsupervised learning*) u prepoznavanju oblika (eng. *pattern recognition)*, *numerička taksonomija* (u biologiji, ekologiji), *tipologija* (u socijalnim naukama) i *podela* ili *particija* (u teoriji grafova) [112,238]. *Hartigan* (1975) je obezbedio detaljan pregled velikog broja publikovanih radova u kojima su prikazani rezultati klaster analize [118].

[1] eng. *cluster*-skupina istovrsnih stvari, grozd, sakupiti u hrpu, gomilu
[2] *Tyron* [243]

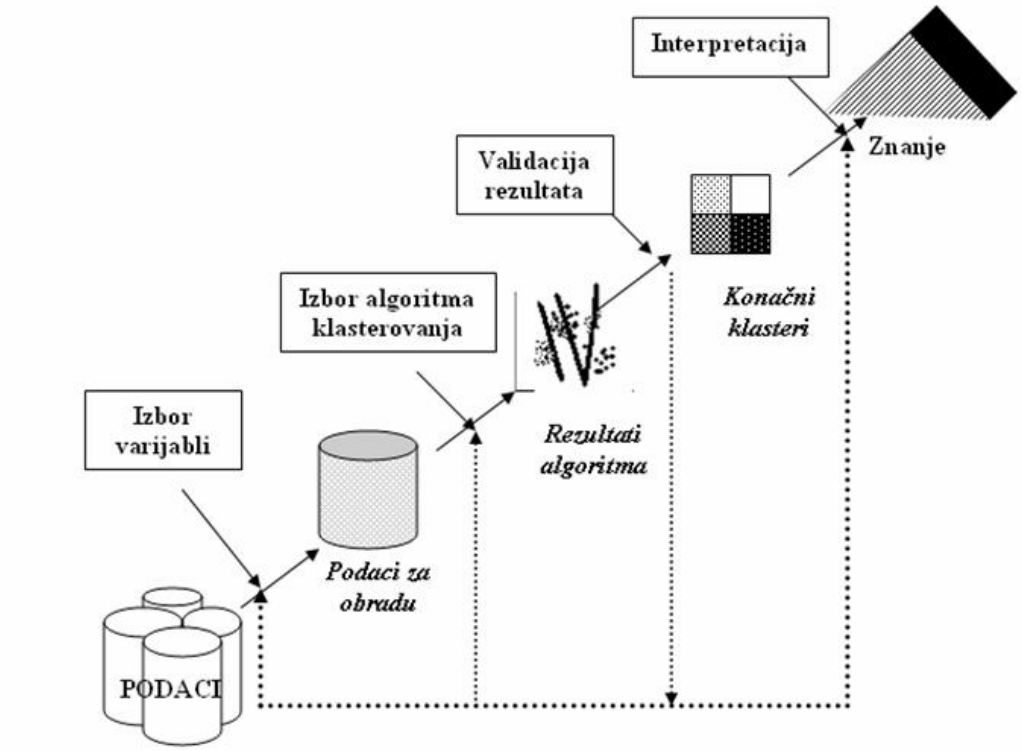
U istraživanjima iz oblasti medicine veoma značajnu ulogu u daljoj analizi može imati klasterovanje oboljenja, ili simptoma oboljenja. *Pardalos* i saradnici [196] daju detaljan pregled razvoja različitih tehnika optimizacije primenjenih u medicini, uključujući dijagnoze obolevanja, predikciju rizika, planiranje tretmana, imidžing i drugo. Klaster analiza se može primenjivati u istraživanju podataka kao samostalan alat za sticanje uvida u raspodelu podataka, uočavanje karakteristika svakog klastera i fokusiranje na određen skup klastera za dalju analizu, pri čemu se i dalje ulažu napori za pronalaženje efikasne metode u radu sa velikim bazama podataka.

Osnovni ciljevi klaster analize su:

- *Istraživanje podataka.* Često ne znamo kako je skup objekata koji se posmatra struktuiran, pa klaster analizom „otkrivamo" nepoznatu strukturu.
- *Redukcija podataka.* Klaster analiza može imati značajnu ulogu u kompresiji informacija koje se nalaze u podacima, što je naročito značajno zbog teškoća prilikom rada sa velikim skupom podataka. Klasterovanje omogućava da se podaci grupišu u „interesantne" klastere, pa se umesto rada na celom skupu podataka kao celini, mogu razmatrati tipični predstavnici ovako dobijenih klastera.
- *Generisanje hipoteza.* Rezultat klaster analize u radu sa podacima nepoznate strukture su klasteri (grupe) čiji broj i sastav može pomoći u definisanju hipoteze o strukturi podataka [144]
- *Testiranje hipoteza.* Klaster analiza se primenjuje u verifikaciji validnosti specifičnih hipoteza, a jedan od načina verifikacije tačnosti je primena klaster analize na reprezentativan skup podataka.
- *Predviđanje.* Rezultujući klasteri su specifični po određenim karakteristikama, osobinama subjekata koji im pripadaju. Na osnovu ovoga, nepoznati subjekt može biti klasifikovan u specifičan klaster na osnovu njegovih „sličnosti" sa karakateristikama tog klastera [112]

Osnovni koraci u klaster analizi su: izbor promenljivih, algoritma klasterovanja (koji uključuje izbor mere sličnosti/ različitosti, kao i kriterijuma klasterovanja), validacija i interpretacija rezultata (slika 1.1).

Za razliku od diskriminantne analize, u kojoj je broj grupa unapred poznat, u klaster analizi nisu poznati broj grupa i karakteristike grupe pre izvođenja samog postupka. Ovde samo pretpostavljamo da objekti pripadaju jednoj od „prirodnih" grupa ili jednostavno želimo izvršiti grupisanje objekata u izvestan manji broj grupa. Grupisanje objekata u manji broj grupa ukazuje da se klaster analiza, slično metodi glavnih komponenata i faktorskoj analizi, može tretirati i kao metoda za redukciju podataka. Međutim, za razliku od ove dve metode, klaster analiza vrši redukciju podataka u odnosu na broj objekata, a ne u odnosu na broj promenljivih.



Slika 1.1. Koraci u postupku klaster analize

Klaster analiza je povezana sa različitim oblastima istraživanja i prisutna je u literaturi već nekoliko decenija [10, 97, 130, 132, 136, 165]. Mada postoje i ranije formulacije klaster analize [80, 272], najveći doprinos i uticaj na njen dalji razvoj imaju trojica istraživača: *Tryon*, *Ward* i *Johnson*. Ovi autori su imali različite pristupe u vezi sa prirodom klaster analize [40]. *Tryon*-ov pristup klaster analizi ima korene u razvoju faktorske analize iz 1930-te. *Tryon* je jedan od prvih istraživača koji se bavio klaster analizom i prvi je upotrebio termin *klaster analiza* (1939. godina). *Ward*-ov pristup je izveden iz analize varijanse, a metod koji je on predložio predstavlja varijantu hijerarhijskih metoda udruživanja. Glavna inovacija koju je *Ward* uveo je optimizacija neke funkcije cilja. Funkcija cilja koju koristi kao primer je greška sume kvadrata unutar klastera. Ovaj metod razmatra minimizaciju varijanse unutar klastera, to jest maksimizaciju varijanse između klastera. Opšte idejno objašnjenje koje je *Johnson* koristio u klaster analizi je multidimenzionalno skaliranje. Ovaj autor je predložio dve metode: *pojedinačno povezivanje* (eng. *single linkage)* i *potpuno povezivanje* (eng. *complete linkage).* On ih opisuje kao „metod minimuma“ i „metod maksimuma“ (1967). *Johnson* je takođe napisao računarski program za njihovo izvođenje za datu matricu sličnosti. Značajnija literatura iz područja klaster analize razvija se od šezdesetih godina XX veka.

Navodimo neke od poželjnih karakteristika, koje treba da ima jedan algoritam klasterovanja [173]:

- *Otkrivanje klastera proizvoljnih veličina i oblika*
- *Identifikovanje klastera visokog kvaliteta u prisustvu šuma (eng. noise)*
- *Sposobnost rada sa različitim tipovima podataka*
- *Neosetljivost na redosled koraka u algoritmu*
- *Merljivost* (eng. *scalability*) *u prisustvu visoko-dimenzionalnih podataka*
- *Minimalni ulazni parametri*
- *Sposobnost rada sa ekstemnim vrednostima*
- *Klasterovanje zasnovano na ograničenjima*
- *Interpretativnost rezultatata i jednostavnost korišćenja*

Veliki broj algoritma klasterovanja dobro radi na malom skupu podataka (koji sadrži manje od 200 objekata) međutim, velike baze podataka mogu sadržati milione podataka. Klasterovanje uzorka iz datog velikog skupa podataka može dovesti do pristrasnih rezultata, te su potrebni merljivi algoritmi klasterovanja za ovakve podatke. Ovakvi podaci mogu biti veoma raštrkani (proređeni) i veoma iskrivljeni.

Većina algoritama klasterovanja zahteva unos određenih parametara, kao što je broj željenih klastera. Sami rezultati klasterovanja su često prilično osetljivi na ulazne parametre, koje je uglavnom teško odrediti, naročito za skupove podataka koji sadrže visokodimenzionalne objekte. Ovo ne samo da predstavlja opterećenje za istraživača, već dovodi i do toga da je teško kontrolisati kvalitet klastera.

Većina stvarnih baza podataka sadrži ekstremne vrednosti (autlajere), to jest objekte koji ne pripadaju ni jednom klasteru ili formiraju klastere veoma male veličine. Neki klaster algoritmi su osetljivi na takve podatke i mogu kao rezultat imati klastere lošeg kvaliteta.

U stvarnom svetu, može biti neophodno klasterovanje skupa podataka uz različite vrste ograničenja. Izazovan zadatak je pronaći klastere podataka sa „dobrim osobinama“, a da zadovoljavaju specifična ograničenja.

1.1.1 Definisanje problema klasterovanja

Pretpostavimo da je A konačan skup tačaka n-dimenzionalnog prostora $\mathbb{R}^n$.

$$A=\left\{a^1,\ldots,a^m\right\}, \text{ gde } a^i\in\mathbb{R}^n, \quad i=1,\ldots,m.$$

Razmatramo „tvrdo“ (eng. *hard*) klasterovanje bez ograničenja, tj. raspodelu tačaka skupa A u k disjunktnih podskupova po unapred definisanom kriterijumu tako da važi:

1. $A^j\neq\varnothing,\ j=1,\ldots,k$
2. $A^i\cap A^l=\varnothing,\ i,\ l=1,\ldots,k,\ i\neq l$
3. $A=\bigcup_{j=1}^{k}A^j$

Skupovi $A^j,\ j=1,\ldots,k$ se nazivaju **klasteri**. Pretpostavimo da se svaki klaster može identifikovati pomoću njegovog centra (ili centroida) $x^j\in\mathbb{R}^n,\ j=1,\ldots,k$. Tada se problem klasterovanja može svesti na sledeći optimizacioni problem[233]:

$$\min \ddot{\ }(C,x), \qquad \text{gde je } \ddot{\ }(C,x)=\frac{1}{m}\sum_{i=1}^{k}\sum_{a\in A^i}\left\|x^i-a\right\|^2 \tag{1.1}$$

$$C\in\bar{C},\ x=(x^1,\ldots,x^k)\in\mathbb{R}^{n\times k},$$

gde je sa $\|\cdot\|$ označena Euklidova normu $\|x\|=\left(\sum_{i=1}^{n}|x_i|^2\right)^{1/2}$, $C=\left\{A^1,\ldots,A^k\right\}$ skup klastera, $\bar{C}$ je skup svih mogućih k-particija skupa A, x^i je centar klastera $A^i,\ i=1,\ldots,k$:

$$x^i=\frac{1}{\left|A^i\right|}\sum_{a\in A^i}a,$$

Problem (1.1) je poznat kao *problem najmanjih kvadrata u klasterovanju***.** Jasno je da se numeričke metode optimizacije ne mogu direktno primeniti za rešavanje ovog problema, pa ga je potrebno preformulisati [26]. Pre toga ćemo se osvrnuti na osnovne pojmove iz teorije optimizacije.

Napomena 1.1. Definisali smo najčešće prisutan problem klaster analize za skupove podataka sa neprekidnim numeričkim obeležjima. Umesto izraza $\sum_{a\in A^i}\left\|x^j-a\right\|^2$ koristi se mera sličnosti, ili različitosti (rastojanja), u zavisnosti od vrste obeležja uključenih u klaster analizu (o čemu će više biti reči u poglavlju 3.1).

1.2 TEORIJSKE OSNOVE

U prvom delu teoretskog uvoda podsetićemo se nekih oznaka, kao i osnovnih rezultata glatke analize. Zatim ćemo uopštiti diferencijalni račun za konveksnu funkciju, koja ne mora biti diferencijabilna. Definisaćemo subgradijente i subdiferencijale [209] i prikazati neke osnovne rezultate. Nakon toga ćemo uopštiti konveksnu diferencijalnu teoriju na lokalne Lipšic neprekidne funkcije [61], uopštićemo klasične uslove optimizacije, to jest navesti potreban uslov da lokalne Lipšic neprekidne funkcije dostižu minimum za slučaj bez ograničenja. Dokazi se mogu pronaći u npr. [167]. U poslednjem, trećem delu ćemo se osvrnuti na neke pojmove iz teorije verovatnoće, koje će se koristiti u opisu pojedinih klaster alagoritama.

1.2.1 Označavanje i definicije

Definicija 1.1 *Skalarni (unutrašnji) proizvod vektora* $x, y \in \mathbb{R}^n$ *definiše se kao*

$$(x, y) = x^T y = \sum_{i=1}^{n} x_i y_i,$$

gde su $x_i, y_i \in \mathbb{R}$ su $i-te$ *komponente vektora* x i y, *redom.*

Definicija 1.2 *Za vektor* $x \in \mathbb{R}^n$ *Euklidova norma je definisana sa*

$$\|x\| = \left(\sum_{i=1}^{n} x_i^2 \right)^2 = \left(x^T x \right),$$

Definicija 1.3 Otvorena lopta sa centrom u $x \in \mathbb{R}^n$ i poluprečnikom $r > 0$ je označena sa

$$B(x, r) = \left\{ y \in \mathbb{R}^n \,\middle|\, \|y - x\| < r \right\}.$$

Definicija 1.4 *Trag matrice* $A \in \mathbb{R}^{n \times n}$ *se označava sa* $\mathrm{tr}(A)$ *i predstavlja sumu dijagonalnih elemenata matrice.*

Trag matrice je jednak sumi njenih svojstvenih vrednosti. Za kvadratne matrice A i B važi $\mathrm{tr}(A+B) = \mathrm{tr}(A) + \mathrm{tr}(B)$.

Definicija 1.5 *Kažemo da je skup* $S \subset \mathbb{R}^n$ ***konveksan*** *ako je*

$$\alpha x + (1-\alpha) y \in S,$$

gde su x i y u S i $\alpha \in [0,1]$.

Ako su $S_i \subset \mathbb{R}^n$ konveksni skupovi za $i = 1, \ldots, m,$ tada je njihov presek $\bigcap_{i=1}^{m} S_i$ takođe konveksan skup.

Linearna kombinacija $\sum_{i=1}^{k} \alpha_i x_i$ se naziva konveksna kombinacija elemenata $x_1, x_2, \ldots, x_k \in \mathbb{R}^n$ ako za svako $\alpha_i > 0$ važi $\sum_{i=1}^{k} \alpha_i = 1.$

Definicija 1.6 *Presek svih konveksnih skupova koji sadrže dati podskup* $S \subset \mathbb{R}^n$ *se naziva* ***konveksan omotač*** *skupa* $S \subset \mathbb{R}^n$, *u oznaci* $\operatorname{conv} S$. *Za proizvoljno* $S \subset \mathbb{R}^n$, $\operatorname{conv} S$ *se sastoji od svih konveksnih kombinacija elemenata iz* S, *to jest*

$$\operatorname{conv} S = \left\{ x \in \mathbb{R}^n \,\middle|\, x = \sum_{i=1}^{k} \alpha_i x_i, \sum_{i=1}^{k} \alpha_i = 1, x_i \in S, \alpha_i \geq 0 \right\}.$$

Konveksni omotač skupa S predstavlja najmanji konveksni skup koji sadrži S, i važi da je S konveksan skup ako i samo ako važi $S = \operatorname{conv} S.$

Definicija 1.7 *Za datu funkciju* $f : \mathbb{R}^n \to \mathbb{R}$ *kažemo da je* ***konveksna*** *ukoliko važi da je*

$$f\left(\alpha x + (1-\alpha) y\right) \leq \alpha f(x) + (1-\alpha) f(y)$$

za svako $x, y \in \mathbb{R}^n$ *i* $\alpha \in [0,1].$

Ukoliko u prethodnom izrazu važi stroga nejednakost za $x \neq y$ i $\alpha \in (0,1)$, kažemo da je funkcija ***strogo konveksna***.

Definicija 1.8 *Neka je dat skup* $Q \subset \mathbb{R}^n$. *Funkcija* $f : \mathbb{R}^n \to \mathbb{R}$ *je* ***Lipšicova na skupu*** Q *sa konstantom* $L_Q > 0$ *ako važi da je*

$$|f(x) - f(y)| \leq L_Q \|x - y\| \text{ za svako } x, y \in Q.$$

Definicija 1.9 *Funkcija* $f : \mathbb{R}^n \to \mathbb{R}$ *je* ***lokalno Lipšicova*** *ako je Lipšicova na svakom ograničenom podskupu skupa* $\mathbb{R}^n$.

Definicija 1.10 *Funkcija* $f : \mathbb{R}^n \to \mathbb{R}$ *je* ***globalno Lipšicova*** *ili samo Lipšicova sa konstantom* $L > 0$ *ako važi da je*

$$|f(x) - f(y)| \leq L \|x - y\| \text{ za svako } x, y \in \mathbb{R}^n.$$

Definicija 1.11 *Funkcija* $f:\mathbb{R}^n \to \mathbb{R}$ *je* ***Lipšicova u tački*** $x \in \mathbb{R}^n$ *ako je Lipšic-ova u nekoj okolini tačke* $x \in \mathbb{R}^n$.

$$\left|f(x)-f(y)\right| \le L\|x-y\| \text{ za svako } x, y \in \mathbb{R}^n.$$

Definicija 1.12 Funkcija $f:\mathbb{R}^n \to \mathbb{R}$ je **diferencijabilna** u $x \in \mathbb{R}^n$ ako postoji vektor $\nabla f(x) \in \mathbb{R}^n$ i funkcija $£:\mathbb{R}^n \to \mathbb{R}$ takva da za svako $d \in \mathbb{R}^n$

$$f(x+d) = f(x) + \nabla f(x)^T d + \|d\| £(d)$$

i $£(d) \to 0$ za $\|d\| \to 0$. Vektor $\nabla f(x)$ je **gradijentni vektor** funkcije

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n} \right)^T,$$

gde su komponente $\frac{\partial f(x)}{\partial x_i}$, za $i = 1, \ldots, n$ parcijalni izvodi funkcije f.

Definicija 1.13 *Ako je funckija diferencijabilna i svi njeni parcijelni izvodi su neprekidni, tada kažemo da je funkcija* ***neprekidno diferencijabilna*** *ili* ***glatka*** *(*$f \in C^1(\mathbb{R}^n)$*).*

Granična vrednost:

$$f'(x;d) = \lim_{\sim \to +0} \frac{f(x+\sim d) - f(x)}{\sim}$$

je izvod funkcije f u odnosu na pravac $d \in \mathbb{R}^n$ za svako x.

Ako je funkcija f diferencijabilna u x, tada postoji njen izvod u svakom pravcu $d \in \mathbb{R}^n$ i važi $f'(x;d) = \nabla f(x)^T d$.

1.2.2 Neglatka analiza

Teorija neglatke analize je zasnovana na konveksnoj analizi. Iz tog razloga ćemo ovaj deo započeti sa prikazom nekih definicija i rezultata koji važe za konveksne funkcije. Cilj ovog dela nisu detaljni opisi neglatke analize (za detalje, pogledati [61,166,209] već prikaz nekih osnovnih definicija i rezultata neophodnih za rad u narednim poglavljima teze.

Definicija 1.14 [209] ***Subdiferencijal*** *konveksne funkcije* $f:\mathbb{R}^n \to \mathbb{R}$ *u* $\boldsymbol{x} \in \mathbb{R}^n$ *je skup* $\partial_c f(x)$ *vektora* $\P \in \mathbb{R}^n$ *tako da važi*

$$\partial_c f(x) = \left\{ \P \in \mathbb{R}^n \,\middle|\, (\forall y \in \mathbb{R}^n)\, f(y) \geq f(x) + \P^T (y - x) \right\}.$$

Svaki vektor $\P \in \partial_c f(x)$ *se naziva* ***subgradijent*** *funkcije* f *u* x.

Subdiferencijal je neprazan, konveksan i kompaktan i važi $\partial_c f(x) \subset B(0; L)$, gde je $L > 0$ Lipšicova konstanta funkcije f u x.

Kako za lokalne Lipšic neprekidne funkcije ne moraju postojati klasični izvodi u pravcu, prvo ćemo definisati **generalisani izvod**. Tada uopštavamo subdiferencijal za nekonveksne lokalno Lipšic neprekidne funkcije.

Definicija 1.15 [61] Neka je funkcija f *lokalno Lipšic neprekidna u* $\boldsymbol{x} \in \mathbb{R}^n$. ***Generalisani izvod (Clarke)*** *funkcije* f *u tački* x *u odnosu na pravac* $d \in \mathbb{R}^n$ *je*

$$f^{\circ}(x, d) = \limsup_{u \to x, \sim \to +0} \frac{f(u + \sim d) - f(u)}{\sim}$$

Definicija 1.16 *Funkcija* f *je* ***Clarke regularna*** *u tački* $x \in \mathbb{R}^n$ *ukoliko je diferencijabilna u odnosu na bilo koji pravac* $d \in \mathbb{R}^n$ *i važi*

$$f'(x, d) = f^{\circ}(x, d)$$

za svako $x, d \in \mathbb{R}^n$ *gde je* $f'(x, d)$ *izvod funkcije* f *u tački* x *u odnosu na pravac* d.

Definicija 1.17 [61] *Neka je* $f:\mathbb{R}^n \to \mathbb{R}$ *lokalno Lipšic neprekidna funkcija u* $\boldsymbol{x} \in \mathbb{R}^n$. *Tada je* ***generalisani (Clarke) gradijent*** *od* f *u* x *skup* $\partial f(x)$ *vektora* $\P \in \mathbb{R}^n$ *takav da važi*

$$\partial f(x) = \left\{ \P \in \mathbb{R}^n \,\middle|\, (\forall d \in \mathbb{R}^n)\, f^{\circ}(x; d) \geq \P^T d \right\}.$$

Svaki vektor $\P \in \partial f(x)$ *se naziva* ***subgradijent*** *od* f u x.

Teorema 1.1 [202] *Neka je* $S \subset \mathbb{R}^n$ *otvoren skup. Funkcija* $f:S \to \mathbb{R}$ *koja je lokalno Lipšic neprekidna na* S *je diferencijabilna skoro svuda na S.*

Razmatramo problem neglatke optimizacije bez ograničenja

$$\min_{x\in\mathbb{R}^n} f(x),$$

gde je ciljna funkcija $f:\mathbb{R}^n\to\mathbb{R}$ lokalno Lipšic neprekidna funkcija u $\boldsymbol{x}$ za svako $x\in\mathbb{R}^n$.

Definicija 1.18 *Tačka* $x\in\mathbb{R}^n$ *je* ***globalni minimum funkcije*** f *ako važi*

$$f(x)\le f(y) \text{ za svako } y\in\mathbb{R}^n.$$

Definicija 1.19 *Tačka* $x\in\mathbb{R}^n$ *je* ***lokalni minimum funkcije*** f *ako postoji* $\pounds>0$ *takvo da je*

$$f(x)\le f(y) \text{ za svako } y\in B(x;\pounds).$$

Potreban uslov da lokalno Lipšic neprekidna funkcija dostigne lokalni minimum u slučaju bez ograničenja je dat sledećom teoremom. Za konveksne funkcije ovi uslovi su takođe i dovoljni i minimum je globalni.

Teorema 1.2 (*Uslov stacionarnosti)* [166] *Neka je* $f:\mathbb{R}^n\to\mathbb{R}$ *lokalno Lipšic neprekidna funkcija u* $\boldsymbol{x}\in\mathbb{R}^n$. *Ukoliko funkcija* f *dostiže lokalni minimum u* x^*, *tada važi*

$$0\in\partial f(x^*).$$

1.2.3 Teorija verovatnoće i statistike

Definicija 1.20 *Ako je X neprekidna slučajna promenljiva sa gustinom raspodele $f(x)$ za $-\infty < x < \infty$, tada je* ***matematičko očekivanje*** *slučajne promenljive X jednako:*

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

Za diskretnu slučajnu promenljivu sa raspodelom verovatnoća

$$\begin{pmatrix} x_1 & x_2 & \cdots \\ p(x_1) & p(x_2) & \cdots \end{pmatrix} \qquad \sum_{i=1}^{n} p_i = 1$$

matematičko očekivanje slučajne promenljive X je dato sa:

$$E(X) = \sum_{i=1}^{n} x_i p_i.$$

Definicija 1.21 ***Disperzija (varijansa)*** *slučajne promenljive X se definiše kao matematičko očekivanje kvadrata odstupanja slučajne promenljive X od matematičkog očekivanja:*

$$D(X) = E\left[X - E(X)\right]^2 = E^2(X) - \left[E(X)\right]^2.$$

Za neprekidnu slučajnu promenljivu, sa funkcijom gustine $f(x)$ važi:

$$D(x) = \int_{-\infty}^{\infty} \left(x - E(x)\right)^2 f(x)dx,$$

dok za diskretnu slučajnu promenljivu važi:

$$D(x) = \sum_{i=1}^{\infty} \left(x_i - E(x_i)\right)^2 p_i.$$

Ako slučajni događaji $G_1, G_2, \ldots, G_n$ označavaju različita stanja nekog fizičkog sistema X, gde je funkcionisanje sistema prelazak iz jednog u drugo stanje, tada se svakom stanju može pridružiti brojevna vrednost i X smatrati slučajnom promenljivom (gde su p_i verovatnoće zauzimanja stanja G_i).

Entropiju $H(x)$ slučajne veličine x možemo interpretirati kao količinu informacija koju sadrži jedna realizacija x, meru neodređenosti u vezi ishoda x [*66*], kao i očekivanu vrednost broja bita neophodnih za opis jedne realizacije x. Uglavnom se pod entropijom podrazumeva ocenjena entropija, poznata kao empirijska entropija.

Definicija 1.22 *Neka je* X *diskretna slučajna promenljiva sa raspodelom verovatnoća*

$$X:\begin{pmatrix} x_1 & x_2 & \cdots \\ p_1 & p_2 & \cdots \end{pmatrix}, \qquad \sum_{i=1}^{n} p_i = 1$$

Entropija *slučajne promenljive* X *je data sa:*

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

Definicija 1.23 *Neka je* $(X_1,\ldots,X_n)$ *prost uzorak obima* n *i neka je* $(x_1,\ldots,x_n)$ *realizovan uzorak. Označimo sa* $f(x;\theta)$ *gustinu raspodele slučajne promenljive* X*, ako je* X *neprekidnog tipa, a sa* $P(\boldsymbol{X}=x;\theta)$, $x\in\{x_1,\ldots,x_n\}$, ako je X *diskretnog tipa. Funkcija verodostojnosti* $L(\theta)$ *se definiše kao*

$$L(\theta)=L(x_1,x_2,\ldots,x_n;\theta)=\begin{cases} f(x_1;\theta)\cdot f(x_2;\theta)\cdot\ldots\cdot f(x_n;\theta), & X \text{ je neprekidnog tipa,} \\ p(x_1;\theta)\cdot p(x_2;\theta)\cdot\ldots\cdot p(x_n;\theta), & X \text{ je diskretnog tipa.} \end{cases}$$

Metoda maksimalne verodostojnosti svodi se na određivanje one ocene $\hat{\theta}$ parametra θ za koju funkcija $L(x_1,x_2,\ldots,x_n;\theta)$ dostiže svoj maksimum. Određivanje maksimuma funkcije L svodi se na anuliranje njenog parcijalnog izvoda po parametru θ. Realno rešenje $\hat{\theta}$ jednačine

$$\frac{\partial L}{\partial \theta}=0,$$

za koju funkcija L dostiže svoj maskimum, predstavlja najefikasniju ocenu parametra θ.

Neka je $\theta=\cdot(x_1,\ldots,x_n)$ *vrednost parametra za koje funkcija* $L(\theta)$ *postiže maksimum. Statistika*

$$\hat{\theta}=\cdot(X_1,\ldots,X_n)$$

je ***ocena maksimalne verodostojnosti*** *parametra* θ.

U statističkom smislu nalaženje optimalnog broja klastera je ekvivalentno fitovanju modela sa registrovanim podacima i optimizacijom nekog kriterijuma. U literaturi postoje različiti kriterijumi koji kombinuju koncepte iz teorije informacija, kao što su *Akaikeov informacioni kriterijum* (AIC) i *Bajesov informacioni kriterijum* (BIC). Statistike za oba kriterijuma su zasnovane na funkciji maksimalne verodostojnosti i broju parametara testirane raspodele i koriste se za poređenje dva modela, to jest testiranje izbora modela.

Definicija 1.24 ***Akaike-ov informacioni kriterijum*** (AIC) [6] *se definiše kao*

$$\mathrm{AIC} = -2L\left(\hat{\vec{\theta}}\right) + 2t$$

gde je $L\left(\hat{\vec{\theta}}\right)$ *maksimum logaritma funkcije verodostojnosti,* $\hat{\vec{\theta}}$ je vektor ocenjenih parametara, *a* t *broj parametara u modelu.*

Definicija 1.25 ***Bayes-ov informacioni kriterijum*** (BIC) [225] *se definiše kao*

$$\mathrm{BIC} = -2L\left(\hat{\vec{\theta}}\right) + t\ln m$$

gde je $L\left(\hat{\vec{\theta}}\right)$ *maksimum logaritma funkcije verodostojnosti,* $\hat{\vec{\theta}}$ je vektor ocenjenih parametara, t *broj parametara u modelu, a* m *veličina uzorka* (*broj observacija*).

1.3 POJAM OPTIMIZACIJE

Teorija optimizacije je relativno nova naučna grana koja pripada oblastima primenjene matematike i operacionih istraživanja i ima široku primenu u nauci, tehnici, poslovnom menadžmentu, vojnoj i kosmičkoj tehnologiji. Optimizacija je postupak nalaženja najboljeg rešenja nekog problema u određenom smislu i pri određenim uslovima. Formulacija optimizacionog problema obuhvata:

1. određivanje jedne ili više optimizacionih promenljivih,
2. izbor funkcije cilja,
3. određivanje skupa ograničenja.

Funkcija cilja i ograničenja mogu biti funkcije jedne ili više optimizacionih promenljivih.

U matematičkom smislu, optimizacija predstavlja minimizaciju ili maksimizaciju date *funkcije cilja* ili *kriterijumske funkcije* (od n promenljivih) u odnosu na data ograničenja nad njenim promenljivim. Označimo sa D skup dopustivih rešenja, ili *dopustivi skup*:

$$D = \left\{ x \in \mathbb{R}^n \middle| g_i(x) \le 0, i = 1, \ldots, s \right\},$$

gde je i indeks ograničenja, s ukupan broj ograničenja i $g_i(x)$ su funkcije ograničenja.

Bez gubitka opštosti posmatraćemo minimizaciju funkcije. Neka je dat dopustivi skup skup $D \subset \mathbb{R}^n$ i funkcija cilja $f(x)$. Opštu formulaciju problema optimizacije, matematički možemo zapisati na sledeći način:

$$\min \left\{ f(x) \middle| x \in D \right\}.$$

Ne postoji univerzalni algoritam za optimizaciju. Umesto toga, postoje brojni algoritmi, od kojih je svaki prilagođen određenom tipu optimizacionog problema. U opštem smislu optimizacione metode se mogu klasifikovati u: **analitičke metode, grafičke metode, eksperimentalne metode** i **numeričke (iterativne) metode.** Problemi optimizacije se mogu klasifikovati u odnosu na broj funkcija cilja (**jednokriterijumski, višekriterijumski**), prirodu funkcije cilja (**linearni, nelinearni; konveksni, nekonveksni**), glatkost funkcije (**diferencijabilni, nediferencijabilni**) itd. Prema eventualnim postavljenim uslovima (ograničenjima) metode optimizacije delimo na:

- **metode uslovne optimizacije (sa ograničenjima)**, gde su ograničenja zadata linearnim jednačinama i (ili) nejednačinama;
- **metode bezuslovne optimizacije (bez ograničenja**).

Metode bezuslovne optimizacije mogu se podeliti u dve klase: *metode sa izračunavanjem izvoda* i *metode bez izračunavanja izvoda*. Prva klasa metoda se može primeniti samo na diferencijabilne funkcije, dok se metode druge klase mogu primeniti u slučaju nediferencijabilnih funkcija, kao i u slučajevima kada je funkcija diferencijabilna, ali je računanje njenih izvoda složeno i zahteva značajne vremenske i memorijske resurse.

Problem traženja optimalne particije skupa je problem globalne optimizacije, a funkcija cilja definisana u (1.1) nije ni konveksna ni diferencijabilna i može imati više lokalnih minimuma. Problem traženja globalnog minimuma funkcije više promenljivih je generalno veoma složen problem [92]. Pokušaj direktnog dobijanja rešenja pretraživanjem svih mogućih particija je računski veoma zahtevan zadatak, koji i iziskuje značajno vreme rada, a u slučaju većeg broja podataka (m) i klastera (k) to postaje skoro nemoguće. Razvoj teorije globalne optimizacije je jedan od najizazovnijih problema u modernoj teoriji optimizacije. Globalne optimizacione probleme je suštinski teško proučavati, kako sa teorijske strane, tako i sa računske, pa je veoma važno razvijanje novih efikasnih metoda globalne optimizacije (detaljan pregled u [*93*]).

1.3.1 Neglatka optimizacija

Klasična teorija optimizacije se uvek oslanjala na diferencijabilnost i stroge pretpostavke o regularnosti [91]. Upravo ove pretpostavke su ponekad onemogućavale praktičnu primenu, zbog neglatkosti prirodnih procesa. Uobičajen postupak dovoljno dobre aprkosimacije neglatkog problema glatkim je doveo do grešaka koje se javljaju kao posledica nedovoljno dobre aproksimacije polazne funkcije. Navedeni problemi su jedan od osnovnih razloga za razvoj **teorije neglatke analize.**

Razmotrimo sledeći problem nelinearne optimizacije sa ograničenjem:

$$\min_{x\in D} \; f(x), \tag{1.2}$$

gde je funkcija cilja $f:\mathbb{R}^n \to \mathbb{R}$ lokalno *Lipšic*ova na dopustivom skupu $D \subseteq \mathbb{R}^n$. Ako je f neprekidno diferencijabilna, onda za (1.2) kažemo da je *problem glatke optimizacije***,** a inače je *problem neglatke optimizacije***.** Ukoliko je $D = \mathbb{R}^n$, onda za problem (1.2) kažemo da je *bez ograničenja*. Ako je f konveksna funkcija i D konveksan skup, onda kažemo da je (1.2) *problem konveksne optimizacije.*

Neglatka ili nediferencijabilna optimizacija se bavi problemom određivanja minimuma (ili maksimuma) realne funkcije na $\mathbb{R}^n$ u odsustvu pretpostavke o diferencijabilnosti funkcije. Numerički algoritmi za neglatku optimizaciju se primenjuju za rešavanje dve vrste problema: neglatkih konveksnih i neglatkih nekonveksnih problema, dakle suštinski je važno da li je funkcija cilja konveksna ili ne. Konveksna analiza obezbeđuje matematičku osnovu za neglatku konveksnu optimizaciju kroz nove koncepte, kao što je *Rockafellar*-ov subdiferencijal [209], kao i osnovu za neglatku nekonveksnu optimizaciju. U slučaju konveksnih funkcija, definicije diferencijabilnosti i glatkosti se poklapaju. U mnogim primenama je utvrđeno da su nekonveksni problemi i nediferencijabilni, što je motivisalo *Frensis Clark-a* da se detaljno bavi razmatranjem lokalno *Lipšicove* neglatke funkcije. On je uveo pojam ***generalisanog gradijenta ili Clarke-ovog subdiferencijala*** [61], koji predstavlja sistematsko proširenje *Rockafellar*-ovog subdiferencijala. *Clarke* koristi subdiferencijal za razvijanje *Karush-Kuhn-Tucker* (KKT) uslova za probleme matematičkog programiranja, kao i za razvoj potrebnih uslova za probleme optimalne kontrole. Ovaj novi pristup u rešavanju neglatkosti u optimizaciji je bio povod za razvoj metoda neglatke analize i neglatke optimizacije. Metode neglatke optimizacije su bazirane na pretpostavkama da je funkcija cilja Lipšic neprekidna i da možemo izračunati vrednost funkcije cilja i vrednost proizvoljnog subgradijenta u svakoj tački. *Clarke*-ov subdiferencijal i *Demyanov-Rubinov* kvazidiferencijal [71,72] imaju ključnu ulogu u neglatkoj, nekonveksnoj optimizaciji, omogućavajući da se uopšte mnoge činjenice iz klasične analize. Poslednjih decenija, mnoge numeričke metode, zasnovane na ovim konceptima, su predložene i proučavane u rešavanju različitih problema nekonveksne optimizacije [27,29, 123,141,217]. Međutim ovakav pristup u velikoj meri zavisi od alata konveksne analize. Dalji pomak od konveksnosti je promovisao *Mordukhovich* razvojem sekvencijalne neglatke analize [183,184].

Dve glavne metode za rešavanje problema neglatke optimizacije bez ograničenja su [168]: **subgradijentne metode** [229] i ***metode koje čuvaju informaciju* (bundle)** [123,140,167]. *Subgradijentne metode* se primenjuju na konveksne funkcije, dok *bundle metode* mogu da se primenjuju i na konveksne i nekonveksne funkcije. Osnovna ideja subgradijentnih metoda je uopštavanje neglatkih metoda korišćenjem subgradijenta umesto gradijenta. *Bundle metode* se smatraju najefektivnijim i najpouzdanijim metodama neglatke optimizacije [166]. Osnovna ideja *bundle* metoda je aproksimacija subdiferencijala (skupa gradijenata) ciljne funkcije prikupljanjem subgradijentnih informacija iz prethodne iteracije u jedan *bundle*. Kao što ćemo videti kasnije, Diskretni gradijentni metod [20,29] koji je predložen za rešavanje neglatkog optimizacionog problema klaster analize predstavlja upravo verziju bundle metode.

1.4 TEORIJA SLOŽENOSTI ALGORITAMA

Analiza algoritma predstavlja postupak kojim se predviđa ponašanje i vrši procena potrebnih resursa algoritma. Grana teorije izračunljivosti u računarstvu, **teorija složenosti** proučava vreme i memorijski prostor koje zahteva algoritam, u funkciji veličine ulaznih podataka. Ovaj pristup se koristi za poređenje različitih algoritama za rešavanje istog problema. Teorija složenosti je relativno mlada oblast, sa prvobitnim radovima koji datiraju iz 1971-1972. godine [63,134]. Postoji nekoliko različitih teorija računske složenosti algoritamski rešivog problema, a najpoznatija je ona zasnovana na Tjuringovim mašinama, koje predstavljaju apstraktne matematičke modele računara. Za razvijanje teorije složenosti, najadekvatnije je iskazati sve probleme kao **probleme odlučivanja**, kod kojih je izlazni podatak binaran (da/ne, T/ $\perp$).

1.4.1 Vremenska i prostorna složenost algoritma

Vremenska složenost algortima je funkcija T, gde je $T(m)$ maksimalan broj koraka algoritma koji je potreban za rešavanje problema, ako su ulazni podaci veličine m. **Prostorna složenost** se definiše analogno za memorijski prostor koji je potreban. Uobičajena je praksa razmatranje složenosti za najgori mogući slučaj. Tačnu vrednost funkcije $T(m)$ moguće je odrediti samo za jednostavnije algortime, dok je u najvećem broju slučajeva dovoljno utvrditi asimptotsko ponašanje $T(m)$ za velike vrednosti m (tzv. **O-notacija** [117]), imajući u vidu da je $T(m)$ monotono neopadajuća funkcija.

Definicija 1.26 *Neka su* $f:\mathbb{N}\to\mathbb{R}_+$ *i* $g:\mathbb{N}\to\mathbb{R}_+$ *proizvoljne realne funkcije. Kažemo da je* $g(m)$ ***asimptotska gornja granica funkcije*** $f(m)$, *u oznaci* $f(m)=O(g(m))$ *ako postoje pozitivne konstante* c *i* m_0 *tako da za svako* $m>m_0$ *važi* $f(m)\le cg(m)$.

Kažemo i da je $f(m)$ **istog reda** kao i $g(m)$. Ukoliko je vremenska ili prostorna složenost algoritma $O(g(m))$, kažemo da je složenost algoritma $O(g(m))$. Oznaka $O(g(m))$ se odnosi na samu klasu funkcija, a jednakost $f(m)=O(g(m))$ je uobičajena oznaka za inkluziju $f(m)\in O(g(m))$. Važe sledeće osobine:

$O(f(m))+O(g(m))=O(f(m)+g(m))$,

$O(f(m))O(g(m))=O(f(m)g(m))$.

Teorija složenosti deli sve probleme u dve grupe: „lake“ i „teške“ za rešavanje, u zavisnosti od toga koliko je složena (dakle koliko je brza ili spora) računska procedura za taj problem. Zbog toga uvodimo sledeću definiciju.

Definicija 1.27 *Algoritam je* ***polinomnog vremena*** **(*polinomne složenosti*)** *ukoliko postoji polinom p takav da je*

$$T(m) \le p(m), \quad \forall m \in \mathbb{Z}^+.$$

Najčešće se koristi termin **polinomni algoritam**. Polinomni algoritmi, tj. algoritmi složenosti $O(m^j)$ su značajni u smislu da se sa stanovišta vremena smatraju dobrim, vremenski efikasnim. Označimo sa **P** klasu svih problema za koje postoje polinomni algoritmi, tj. koji su rešivi u polinomnom vremenu. Probleme koje pripadaju klasi **P** smatramo „lakim“.

Navodimo neke primere klasifikacije složenosti algoritama (gde je m veličina ulaza):

- konstantna složenost: $O(1)$. Algoritam ne zavisi od dimenzije ulaza.
- sublinearna složenost: $O(\log m)$, $O(\sqrt{m})$. Klasa veoma efikasnih algoritama
- linearna složenost: $O(m)$, superlinearna složenost: $O(m \log m)$, $O(m^2)$. Efikasni algoritmi
- eksponencijalna složenost: $O(2^m), O(m^m)$. Nisu efikasni za veće dimenzije ulaza.

1.4.2 Klasifikacija teških problema

Pre nego što pređemo na klasifikaciju problema odlučivanja, definisaćemo svodljivost algoritma (eng. *reducibilty*). Neka su $\bar{e}$ i $\bar{e}'$ dva problema odlučivanja. Za problem $\bar{e}'$ kažemo da je **svodljiv** ili se **polinomno redukuje** u problem $\bar{e}$ (u oznaci $\bar{e}' \propto \bar{e}$) ako postoji redukciona funkcija izračunljiva u determinističkom polinomnom vremenu, koja svaku instancu I' problema $\bar{e}'$ transformiše u ekvivalentnu instancu problema $\bar{e}$. Za dva problema kažemo da su **ekvivalentna** ako su su uzajamno svodljiva (ili se jednostavno mogu redukovati) jedan na drugi.

Sledeća teorema pojednostavljuje problem klasifikacije algoritma, to jest ako su $\bar{e}$ i $\bar{e}'$ uzajamno svodljivi problemi i $\bar{e}$ pripada klasi **P** tada i $\bar{e}'$ pripada klasi **P.**

Teorema 1.3 [83] *Ako je* ē $\in$ **P**, tada važi ē′ $\propto$ ē $\Rightarrow$ ē′ $\in$ **P**.

Na koji način se možemo utvrditi da je problem zapravo „težak“?. Počinjemo sa definisanjem šire klase problema, koja uključuje probleme **P** i takođe sve druge teške probleme sa kojima se uopšte srećemo. Označimo sa $V_{A_1}(m)$ maksimalno vreme za algoritam A_1 koje je potrebno za potvrdu da dato rešenje uspostavlja odgovor DA za bilo koju instancu dužine m.

Definicija 1.28 [83] *Za algoritam* Ã *kažemo da je* ***nedeterministički polinomnog vremena****, ako postoji polinom* p *takav da za svaki ulaz dužine* m*, sa odgovorom DA važi*

$$V_{\tilde{A}}(m) \le p(m). \quad \forall m \in \mathbb{Z}^+.$$

NP predstavlja klasu svih problema koji mogu biti rešeni nedeterminističkim algoritmom za polinomijalno vreme. Jedno od najznačajnijih i najintrigantnijih otvorenih pitanja moderne matematike i teorijskog računarstva je problem utvrđivanja odnosa klasa **P** i **NP**, koji je poznat u literaturi kao "problem **P=NP**". Kako inkluzija **P** $\subseteq$ **NP** važi, problem se svodi na pitanje da li se svaki nedeterministički algoritam može determinizovati tako da pri tome ostane sačuvana polinomna vremenska složenost.

Za problem ē se kaže da je **NP-težak** ako je svaki problem iz klase **NP** polinomijalno svodljiv na ē. Za problem ē se kaže da je **NP-kompletan** ako je ē $\in$ **NP** i ē je **NP**-težak. Zaključujemo da su **NP**-kompletni problemi najteži problemi klase NP.

Korišćenje polinomne redukcije (svodljivosti) nam bitno olakšava utvrđivanje složenosti problema, to jest klasifikaciju problema, o čemu govori i sledeća teorema.

Teorema 1.4 [83] *Neka je* ē **NP**-*kompletan problem. Ako je* ē′ $\propto$ ē *i* ē′ $\in$ **NP***, tada je i problem* ē′ **NP**-*kompletan.*

Da bismo primenili ovaj kriterijum, neophodno je da znamo listu problema za koje je poznato da pripadaju klasi **NP**-kompletnih problema. Veoma značajan doprinos je dao S.A.Cook 1971. godine [63] koji je dokazao da postoje **NP**-kompletni problemi (problem zadovoljivosti iskaznih formula). Detaljan pregled **NP**-kompletnih problema je prikazan u [99], za različite oblasti istraživanja.

2. PRIMENA NUMERIČKIH METODA OPTIMIZACIJE U REŠAVANJU PROBLEMA KLASTEROVANJA

Problem određivanja klastera matematički se modelira problemom nelinearne optimizacije sa ograničenjima, koji se u najvećem broju slučajeva, može rešiti samo primenom metoda numeričke optimizacije.

2.1 NEGLATKI OPTIMIZACIONI PRISTUP U KLASTER ANALIZI

Numeričke metode optimizacije se ne mogu direktno primeniti za rešavanje problema (1.1), pa ga je potrebno preformulisati. Problem (1.1) je ekvivalentan sledećem problemu matematičkog programiranja:

$$\min \cdot \ (x,w), \qquad \cdot \ (x,w) = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k} w_{ij} \left\| x^j - a^i \right\|^2, \tag{2.1}$$

za koje važi

$$x = (x^1,\ldots,x^k) \in \mathbb{R}^{n\times k},$$

$$\sum_{j=1}^{k} w_{ij} = 1, \qquad i = 1,\ldots,m,$$

$$w_{ij} \in \{0,1\}, \quad i = 1,\ldots,m, \quad j = 1,\ldots,k,$$

gde je w_{ij} dodeljen težinski koeficijent za objekat a^i i j-ti klaster, na sledeći način:

$$w_{ij} = \begin{cases} 1, & a^i \in A^j, \ \forall i = 1,\ldots,m, \ j = 1,\ldots,k \\ 0, & a^i \notin A^j \end{cases}$$

$$x^j = \frac{\sum_{i=1}^{m} w_{ij} a^i}{\sum_{i=1}^{m} w_{ij}}, \qquad j = 1,\ldots,k,$$

$W = \left[w_{ij} \right]_{m\times k}$ je matrica reda $m \times k$.

Funkciju f nazivamo ***klaster funkcija***. Postoje različiti pristupi za rešavanje problema klasterovanja [82,131,169]. Problem klasterovanja (2.1) predstavlja globalni optimizacioni problem, a različiti algoritmi matematičkog programiranja mogu biti primenjeni za njegovo rešavanje:

dinamičko programiranje, *branch* i *bound*, *cutting planes*, k*-means* i drugo [115]. Dinamičko programiranje se efikasno primenjuje u klasterovanju, kada je broj slučajeva $m \leq 20$, što najčešće nije slučaj kod realnih skupova podataka. *Branch* i *bound* su efikasni kada skup sadrži stotine podataka i broj klastera k nije veliki (<5). Zbog svega ovoga neophodno je korišćenje lokalnih tehnika i različite heuristike za rešavanje velikih problema klasterovanja. Jedna od najpoznatijih tehnika u klaster analizi je $k-means$ algoritam koji služi za traženje lokalnog minimuma za problem (2.1) [233]. Ovaj algoritam je brz (u smislu vremenske složenosti) i daje dobre rezultate za mali broj klastera, međutim rezultati nisu zadovoljavajući za veliki broj klastera. Rezultati numeričkih eksperimenata ukazuju da metaheurististički algoritmi za globalnu optimizaciju, kao što je simulirano kaljenje (eng. *simulated annealing), Tabu search* i genetski algoritmi daju bolje rezultate (u smislu kvaliteta samih klastera) u odnosu na $k-means$ algoritam [206]. Međutim navedeni algoritmi su manje efikasni (u smislu vremenske složenosti), tj. zahtevaju 500 puta više vremena u odnosu na $k-means$ algoritam, za broj slučajeva $m<100$ i broj klastera $k<5$ [9]. Za relativno velike baze podataka, ova razlika se uvećava, što metaheurističke algoritme čini neefikasnim u rešavanju mnogih problema klasterovanja. Problem klaster analize može da se redukuje u problem linearnog programiranja [169]. Međutim, u tom slučaju broj promenljivih u linearnom programiranju je $k \times n \times m$, što je veoma velik broj u radu sa velikim skupovima podataka.

Problem (2.1) je globalni optimizacioni problem i funkcija cilja ima velik broj lokalnih minimuma. Međutim, primena tehnika globalne optimizacije u rešavanju većine problema klasterovanja zahteva dosta vremena. Zbog toga je važno razviti algoritme klasterovanja zasnovane na tehnikama optimizacije koje računaju lokalne minimume koji su blizu globalnog minimuma ciljne funkcije. Opisaćemo algoritam klasterovanja koji je zasnovan na neglatkom optimizacionom pristupu. Ovaj algoritam omogućava izračunavanje klastera korak po korak, postepeno povećavajući broj klastera, dok ne bude ispunjen zadati kriterijum.

Problemi (1.1) i (2.1) se mogu preformulisati na sledeći način [19]:

$$\min_{x \in \mathbb{R}^{n \times k}} f(x), \qquad x = \left(x^1, \ldots, x^k\right) \tag{2.2}$$

gde je

$$f\left(x^1, \ldots, x^k\right) = \frac{1}{m} \sum_{i=1}^{m} \min_{j=1,\ldots,k} \left\| x^j - a^i \right\|^2. \tag{2.3}$$

Za $k > 1$, ciljna funkcija (2.3) u problemu (2.2) je **neglatka i nekonveksna.** Poslednje dve decenije počeo je razvoj numeričkih metoda za probleme neglatke i nekonveksne optimizacije [46, 47, 92, 140, 145, 216, 228].

Problemi (2.1) i (2.2) su ekvivalentni [42], ali postoje određene razlike između ove dve formulacije:

- Broj promenljivih u optimizacionom problemu (2.1) iznosi $(n+m)\times k$, dok u problemu (2.2) iznosi $n\times k$ i ne zavisi od broja slučajeva. U realnim životnim problemima, broj slučajeva m je značajno veći od broja obeležja n.
- U problemu klasterovanja (2.1) koeficijenti w_{ij} su celobrojni, tj. problem sadrži celobrojne i neprekidne promenljive. U neglatkoj optimizacionoj formulaciji, promenljive su neprekidne.
- Formulacija klaster problema zasnovana na neglatkoj optimizaciji omogućava jednostavno razmatranje različitih mera sličnosti.

Sve navedene osobine možemo smatrati prednostima neglatke optimizacione formulacije problema klasterovanja (2.2). Ukoliko su broj klastera (k) i broj promenljivih (n) veliki, tada imamo globalni optimizacioni problem velikih dimenzija (eng. *large-scale*). Funkcija cilja je složena i direktna primena metoda globalne optimizacije nije dovoljna za rešavanje ovog problema. Prema tome, da bismo primenili neglatki optimizacioni pristup za rešavanje problema klasterovanja, veoma je važno prepoznavanje i korišćenje metoda lokalne optimizacije. Jasno je da takav pristup ne garantuje globalno optimalno rešenje problema (2.2). Sa druge strane ovaj pristup obezbeđuje rešenje koje je blizu globalnog minimuma ciljne funkcije, čime se obezbeđuje dovoljno dobar opis klasterovanja skupa podataka.

2.2 ALGORITAM K-SREDINA I GLOBALNI ALGORITAM K-SREDINA

Jedan od najpoznatijih i najviše korišćenih nehijerarhijskih metoda klasterovanja je algoritam k***-sredina*** (k*-means*) (*Ball and Hall* 1965.godina [34]; *MacQueen* 1967. godina [165]; Anderberg 1973.godina [10]). Ovaj algoritam, zajedno sa njegovim varijacijama je poznat kao brz algoritam (u smislu vremenske složenosti) i primenjiv je na velike skupove podataka. Algoritam k-sredina se primenjuje u radu sa neprekidnim tipovima obeležja. Klasteri su opisani pomoću centroida koji predstavlja aritmetičku sredinu objekata koji se nalaze u klasteru. Počinje se od k klastera (određuju se proizvoljno, ili na osnovu prethodnog klasterovanja), a objekti se razvrstavaju u one klastere čiji centroid im je najbliži.

Označimo sa A konačan skup tačaka n-dimenzionalnog prostora $\mathbb{R}^n$:

$$A=\left\{a^1,\ldots,a^m\right\}, \text{ gde } a^i\in\mathbb{R}^n, \quad i=1,\ldots,m.$$

Opisaćemo postupak algoritma k-sredina.

Algoritam 2.1 ***Algoritam*** k ***- sredina***
Korak 1. Izabrati početno rešenje koje se sastoji od k *centara (ne moraju pripadati skupu* A*.*

Korak 2. Dodeliti tačke $a^i\in A$ *najbližem centru čime je dobijena* k*-podela skupa* A*.*

Korak 3. Ponovo odrediti centre za ovu novu podelu i vratiti se na korak 2, sve dok se ne poklope centri klastera u poslednje dve iteracije.

Razlozi velike primene algoritma k*-sredina* su sledeći:

- Vremenska složenost je $O(m\times k\times l)$, gde je m broj slučajeva, k je broj klastera, a l broj iteracija algoritma. Kako su k i l unapred fiksirani, vremenska složenost ovog algoritma je linearna u odnosu na veličinu uzorka.
- Prostorna složenost je $O(k+m)$.
- Algoritam je nezavisan od redosleda. Za dati inicijalni skup centara klastera, generiše istu podelu bez obzira na redosled slučajeva.

Nedostaci ovog algoritma su sledeći:

- mora se unapred odrediti (zadati) broj klastera k
- moraju se pronaći početni centroidi da bi startovao algoritam. *Hartigan* and *Wong* (1979) sugerišu korišćenje aktuelnih objekata kao početnih centara za klastere. Oni mogu biti izabrani na slučajan način

- često konvergira ka lokalnom optimumu
- osetljivost na autlajere i šum
- težnja ka pronalaženju sferičnih klastera jednake veličine

Algoritam *k-sredina* je efikasan u pronalaženju dobre početne tačke za neglatku optimizaciju. Glavna mana ovog algoritma je što je veoma osetljiv na izbor početnih tačaka. Jedan od načina za izbegavanje ovog problema je korišćenje višestrukog restartovanja algoritma *k-sredina*. Međutim, sa povećanjem broja podataka i broja klastera, potrebno je više početnih tačaka za dobijanje bliskog globalnog rešenja za problem klasterovanja. *Chan* i saradnici uopštavaju ovaj algoritam, uvodeći težinske koeficijente za svako obeležje u svakom klasteru [55], što je detaljno opisano u poglavlju 2.7.

Algoritam *k-sredina* konvergira ka lokalnom minimumu i ovi lokalni minimumi se mogu značajno razlikovati od globalnih rešenja, kako se broj klastera povećava. U cilju prevazilaženja ovog nedostatka algoritma, *Likas* i saradnici [158] su kreirali ***globalni algoritam k-sredina***.

Algoritam 2.2 ***Globalni algoritam k-sredina***

Korak 1. Izračunati centar x^1 *skupa* A*:*

$$x^1 = \frac{1}{m}\sum_{i=1}^{m} a^i, \quad a^i \in A, \ i = 1,...,m$$

i neka je $q = 1$.

Korak 2. Neka je $q \to q+1$. *Neka su* $x^1, x^2, ..., x^{q-1}$ *centri klastera iz prethodne iteracije.*

Korak 3. Razmatrati svaku tačku $a \in A$ *kao startnu tačku za centar* q*-tog klastera. Odavde se dobija* m *početnih rešenja* $(x^1, ..., x^{q-1}, a)$. *Primeniti algoritam* k*-sredina na svako od ovih rešenja. Odrediti najbolju* k*-particiju skupa* A*, a odgovarajuće centre označiti sa* $(x^1, ..., x^{q-1}, x^q)$.

Korak 4. Ako je $q = k$, *zaustaviti se; inače se vratiti na korak 2.*

Ova verzija algoritma nije primenjiva za klasterovanje srednjih i velikih skupova podataka. Predložena *su* dva postupka za redukovanje ovog problema [158], pri čemu navodimo jedan od njih. Označimo sa d_{k-1}^{i} kvadrat rastojanja između $a^i \in A$ i najbližeg klaster centra od $k-1$ centara $x^1, x^2, ..., x^{k-1}$:

$$d_{k-1}^{i} = \min\left\{\left\|x^1 - a^i\right\|^2, \ldots, \left\|x^{k-1} - a^i\right\|^2\right\}. \tag{2.4}$$

Za svako $a^i \in A$, računamo:

$$r_i = \sum_{j=1}^{m} \min\left\{0, \left\|a^i - a^j\right\|^2 - d_{k-1}^{j}\right\}$$

i uzimamo tačku $a^l \in A$ za koju je

$$l = \arg\min_{i=1,\ldots,m} r_i,$$

kao startnu tačku za k-ti klaster centar.
Zatim se primenjuje algoritam k-*sredina*, za pronalaženje centara k klastera, pri čemu se kao početna tačka uzima $\left(x^1, x^2, \ldots, x^{k-1}, a^l\right)$.

2.3 KLASTER ALGORITAM ZASNOVAN NA NEGLATKOJ OPTIMIZACIJI

U ovom delu ćemo opisati dva sekvencijalna klaster algoritma bazirana na neglatkom optimizacionom pristupu. U prvom algoritmu (Algoritam 2.3) se primenjuju tehnike neglatke optimizacije za pronalaženje početne tačke za centar k-tog klastera. Ovaj algoritam predstavlja modifikaciju globalnog algoritma k-*sredina*. Drugi algoritam (Algoritam 2.4) je klaster algoritam zasnovan na optimizaciji.

Prvo ćemo opisati algoritam za pronalaženje početne tačke za centar k-tog klastera [28]. Pretpostavimo da su centri $\left(x^1, x^2, \ldots, x^{k-1}\right)$ za k-1 klastera poznati. Uvodimo sledeću funkciju:

$$\bar{f}^k(y) = \frac{1}{m}\sum_{i=1}^{m} \min\left\{d_{k-1}^i, \left\|y - a^i\right\|^2\right\} \tag{2.5}$$

gde je $y \in \mathbb{R}^n$ centar k-tog klastera, a d_{k-1}^i je definisano u (2.4). Posmatramo skup

$$\bar{D} = \left\{y \in \mathbb{R}^n : \left\|y - a^i\right\|^2 \geq d_{k-1}^i\right\}.$$

Za navedeni skup važi da rastojanje između bilo koje njegove tačke y i bilo koje tačke $a^i \in A$ nije manje od rastojanja između ove tačke i centra odgovarajućeg klastera za tu tačku. Takođe, razmatramo i sledeći skup

$$D_0 = \mathbb{R}^n \setminus \bar{D} \equiv \left\{y \in \mathbb{R}^n : \exists I \subset \{1, \ldots, m\}, I \neq \varnothing : \left\|y - a^i\right\|^2 < d_{k-1}^i\right\}.$$

Funkcija $\bar{f}^k(y)$ je konstantna na skupu $\bar{D}$:

$$\bar{f}^k(y) = d_0 = \frac{1}{m}\sum_{i=1}^{m} d_{k-1}^i, \text{ za svako } y \in \bar{D}.$$

Jasno je da $x^j \in \bar{D}$ za svako $j = 1, \ldots, k-1$ i $a^i \in D_0$ za svako $a^i \in A$, $a^i \neq x^j$, $j = 1, \ldots, k-1$. Takođe važi $\bar{f}^k(y) < d_0$ za svako $y \in D_0$.

Bilo koja tačka iz $y \in D_0$ može biti uzeta kao početna tačka za centar k-tog klastera. Mnogo bolji kandidat za početnu tačku je globalni minimizator funkcije $\bar{f}^k(y)$. Međutim, ova funkcija je nekonveksna i neglatka, pa minimizacija ove funkcije predstavlja težak zadatak. Analiziramo postupak za pronalaženje lokalnog minimuma ove funkcije.

Za proizvoljno $y \in D_0$ posmatramo sledeći skup:

$$S_2(y) = \left\{ a^i \in A : \left\| y - a^i \right\|^2 < d_{k-1}^i \right\}.$$

Važi da je $S_2(y) \neq \varnothing$ za svako $y \in D_0$. Sada ćemo opisati algoritam za pronalaženje početne tačke [28] za centar k-tog klastera.

Algoritam 2.3 ***Algoritam za pronalaženje početne tačke***

Korak 1. Za proizvoljno $a^i \in D_0 \cap A$ *izračunati skup* $S_2(a^i)$, *centroid* c^i *za dati skup i odrediti vrednost funkcije funkcije* $\overline{f}^k$ *u ovoj tački,* $\overline{f}^k(c^i)$.

Korak 2. Izračunati

$$\overline{f}_{\min}^k = \min_{a^i \in D_0 \cap A} \overline{f}^k(c^i),$$
$$a^j = \arg\min_{a^i \in D_0 \cap A} \overline{f}^k(c^i),$$

i odgovarajući centar c^j *i skup* $S_2(c^j)$.

Korak 3. Ponovo odrediti skup $S_2(c^j)$ *i izračunati njegov centar sve dok ne bude više tačaka skupa koje se premeštaju.*

Izbor broja klastera je veoma važan korak u klaster analizi. Veoma je teško unapred odrediti optimalan broj klastera koji predstavlja skup A. Neophodno je razmatrati različit broj klastera, počevši od od određenog malog broja k. *Bagirov* i saradnici [26] su predložili optimizacioni algoritam zasnovan na primeni tehnika neglatke optimizacije, u kome se vrši izračunavanje klastera *korak po korak,* postepeno povećavajući broj klastera do ispunjenja kriterijuma zaustavljanja. Ukoliko rešenje odgovarajućeg optimizacionog problema (2.2) nije zadovoljavajuće, nastavlja se sa razmatranjem za $k+1$ i tako dalje. Dakle neophodno je da se ponavlja rešavanje globalnog optimizacionog problema za različite vrednosti broja klastera k.

Algoritam 2.4 ***Klaster algoritam zasnovan na neglatkoj optimizaciji***

*Korak 1. (**Inicijalizacija**). Izabrati $\varepsilon>0$. Izračunati centroid $x^{1*} \in \mathbb{R}^n$ skupa A. Neka je f^{1*} odgovarajuća vrednost funkcije cilja za problem (2.3). Neka je $k=1$.*

*Korak 2. (**Izračunavanje centra sledećeg klastera**). Postaviti $k \to k+1$. Neka su $x^1,\ldots,x^{k-1}$ centri za $k-1$ klastera. Primeniti Algoritam 2.3 za pronalaženje početne tačke $\bar{y} \in \mathbb{R}^n$ za centar k-tog klastera.*

*Korak 3. (**Ažuriranje svih centara klastera**). Izabrati $\left(x^1,\ldots,x^{k-1},\bar{y}\right)$ kao novu početnu tačku, primeniti algoritam k-sredina za rešavanje problema k particije. Neka je $\left(y^1,\ldots,y^k\right)$ rešenje ovog problema i f^k odgovarajuća vrednost ciljne funkcije (2.3).*

*Korak 4. (**Kriterijum zaustavljanja**). Ako je*

$$\frac{f^{k-1}-f^k}{f^1}<£,$$

zaustaviti se, inače postaviti $x^i=y^i,\ i=1,\ldots,k$ i vratiti se na korak 2.

U koraku 1 izračunat je centar celog skupa. U koraku 2 računa se centar k-og klastera, pod pretpostavkom da su poznati centri prethodnih k-1 klastera.

Jasno je da za svako $k \geq 1$ važi $f^k \geq 0$ i niz $\left\{f^k\right\}$ je opadajući :

$$f^{k+1} \leq f^k \ .$$

Odavde sledi da će nakon konačnog broja iteracija kriterijum zaustavljanja u koraku 4 biti ispunjen. Veoma je važan izbor granice $£>0$. Velike vrednosti dovode do pojave velikih klastera, dok male vrednosti mogu proizvesti male i veštačke klastere. Rezultati numeričkih eksperimenata [21] ukazuju da je optimalna vrednost $£ \in \left[10^{-1},10^{-2}\right]$.

2.4 REŠAVANJE OPTIMIZACIONOG PROBLEMA

Numeričke metode globalne optimizacije zahtevaju dosta vremena i ne mogu se primenjivati kod visoko-dimenzionalnih problema nekonveksne optimizacije. Ovo i predstavlja razlog korišćenja različitih kombinacija tehnika globalnog i lokalnog pretraživanja. Generalno važi da su lokalne metode veoma osetljive na izbor početne tačke. Uglavnom se koriste sledeća dva tipa kombinacija lokalnih i globalnih metoda optimizacije:

1. Lokalne tehnike se koriste za dobijanje stacionarne tačke (lokalnog minimuma), a zatim se primenjuju globalne tehnike sa ciljem pronalaženja nove tačke koja će biti korišćena kao početna vrednost za novi krug lokalnog pretraživanja [23,121].
2. Tačke dobijene pomoću globalnih tehnika koriste se kao početne tačke za lokalno pretraživanje [22].

Ovakvi pristupi su dovoljno dobri za lokalnu minimizaciju funkcije koja ima nekoliko stacionarnih tačaka, međutim ne funkcioniše u slučaju funkcije koja ima mnogo lokalnih minimuma, kao što je klaster funkcija. Zbog svega toga, više nas interesuje lokalni mimimum koji je blizak globalnom minimumu.

Ciljna funkcija (2.3) u problemu (2.2) i funkcija (2.5) su nekonveksne i neglatke, za $k > 1$. Osim toga, ove funkcije nisu regularne, pa je izračunavanje čak i subgradijenata prilično teško. Bagirov je predložio metodu bez izračunavanja izvoda, **Diskretni gradijentni metod** [20, 29] za rešavanje problema (2.3). Ovaj metod predstavlja verziju *metode koja čuva informacije* (eng. *bundle)*, gde su subgradijenti funkcije cilja zamenjeni njihovim diskretnim gradijentima. Diskretni gradijentni metod obuhvata sledeće etape: računanje subgradijenata klase neregularnih funkcija, dokaz da diskretni gradijenti mogu aprokismirati subdiferencijale takvih funkcija i algoritam za određivanje pravca opadanja neglatkih funkcija korišćenjem diskretnih gradijenata. *Andramonov* i saradnici (1999), *Bagirov* i *Rubinov* (2001) su predložili *cutting angle* metod (CAM) koji je efikasan metod za rešavanje globalnog optimizacionog problema [11,24]. Neke modifikacije CAM i kombinacija CAM algoritma sa lokalnim pretraživanjem se uspešno primenjuju u klasifikaciji [22].

2.5 REDUKOVANJE SLOŽENOSTI ZA VELIKE SKUPOVE PODATAKA

Dve karakteristike datog skupa podataka mogu značajno uticati na rezultate klasterovanja: broj objekata i broj promenljivih. U mnogim situacijama da bi rad sa skupom podataka bio efikasan neophodna je redukcija obe karakteristike, bez gubitka informacija. Analiziramo redukovanje broja objekata.

Visoko-dimenzionalan skup podataka uobičajeno sadrži veliki broj tačaka smeštenih u ograničenom skupu. Mnoge tačke iz ovoga skupa su veoma blizu jedna drugoj. Neka je $A \subset \mathbb{R}^n$ konačan skup. Pretpostavimo da neka mala okolina tačke $b \in \mathbb{R}^n$ sadrži m_b tačaka iz A. Možemo aproksimirati svaku od ovih tačaka pomoću b i zamenjujući odgovarajući deo klaster funkcije izrazom $m_b \|x_i - b\|$.

Definicija 2.1 [21] Dati su skupovi $A \subset \mathbb{R}^n$, $B \subset \mathbb{R}^n$ *i data je granica tolerancije* $\pounds$, *tako da za svako* $a \in A$ *postoji* $b \in B$ *i važi* $\|a-b\| < \pounds'$. *Kažemo da* $(A_b)_{b\in B}$ *skup podskupova skupa* A *predstavlja* $\pounds'$**–*disjunktno pokrivanje*** *skupa* A, *ako je*

$$\|a-b\| < \pounds',\ (a \in A_b),\ A_b \cap A_{b'} = \varnothing\ (b \neq b'),\ A = \underset{b\in B}{\cup} A_b.$$

Definicija 2.2 [21] *Neka je* $(A_b)_{b\in B}$ $\pounds'$*–disjunktno pokrivanje skupa* A *i kardinalnost skupa* $|A_b| = m_b$. *Zamenjujući svako* $a \in A_b$ *sa* b *u klaster funkciju* f, *dobijamo*

$$\tilde{f}\left(x^1,\ldots,x^k\right) = \frac{1}{m}\sum_{b\in B} m_b \min\left(\left\|x^1-b\right\|,\ldots,\left\|x^k-b\right\|\right).$$

Dobijena funkcija je ***generalisana klaster funkcija.***

Tvrđenje 2.1 [21] *Neka je* $(A_b)_{b\in B}$ $\pounds'$*–disjunktno pokrivanje skupa* A *i* $\tilde{f}$ *generalisana klaster funkcija koja odgovara ovom pokrivanju. Važi da je*

$$\left| f\left(x^1,\ldots,x^k\right) - \tilde{f}\left(x^1,\ldots,x^k\right)\right| < \pounds' \quad \text{za svako } \left(x^1,\ldots,x^k\right) \in \left(\mathbb{R}^n\right)^k.$$

Dokaz. Videti u [101].

Tvrđenje 2.1. omogućava da se dati skup A zameni manjim skupom B. Minimizacija generalisane klaster funkcije koja odgovara ovom skupu daje nam neke tačke $\left(x^1,\ldots,x^k\right)$. Možemo ove tačke razmatrati kao centre k klastera skupa A. Tada klaster A_j sa odgovarajućim centrom x^j možemo opisati kao uniju skupova A_b, $A_j = \cup A_b$ za svako b, za koje važi da je:

$$\left\|b - x^j\right\| \le \min_{i \ne j} \left\|b - x^i\right\|.$$

Opisaćemo postupak [21] za konstrukciju $£'-$disjunktnog pokrivanja skupa A sa datom granicom tolerancije $£'$. Neka je $A = \{a_i\}_{i=1,\ldots,m}$ dati skup podataka. Neka je $D = \left(d_{ij}\right)_{i,\,j=1,\ldots,\mathrm{m}}$ simetrična matrica, gde je $d_{ij} = \left\|a^i - a^j\right\|$. Koraci postupka su sledeći:

1. izaberemo prvi vektor a^1, eliminišemo iz skupa sve vektore za koje važi $d_{1j} \le £'$, i dodeljujemo ovom vektoru broj elimisanih vektora m_1. Označimo $a^1 = b^1$.

2. izaberemo sledeći preostali vektor b^2 i ponavljamo navedeni postupak za ovaj vektor itd.

Kao rezultat ovakvog postupka, dobijamo podskup $B = \left\{b^j\right\}$, $j = 1,\ldots,l$ datog skupa A i skup $\left(m_j\right)$, gde je m_j broj eliminisanih vektora u koraku j. Kardinalnost l skupa B može biti značajno manja od kardinalnosti m skupa A. U cilju pronalaženja klastera skupa A primenjujemo generalisanu klaster funkciju

$$\tilde{f}\left(x^1,\ldots,x^k\right) = \frac{1}{m}\sum_{j=1}^{l} m_j \min\left(\left\|x^1 - b^j\right\|,\ldots,\left\|x^k - b^j\right\|\right).$$

Izabran vektor b^j je predstavnik skupa A^j svih tačaka koje su eliminisane u koraku j.

Opisaćemo jedan od mogućih načina odabira parametra $£'$ [21].

Za svako i, $i = 1,\ldots,m$,

$$r_i' = \min_{j \ne i} d_{ij},$$

gde je d_{ij} već definisano . Neka je

$$r_0 = \frac{1}{m}\sum_{i=1}^{m} r_i'.$$

Izabrati $£' = cr_0$, $c > 0$.

Rezultati numeričkih eksperimenata Bagirova i saradnika [21], za različite baze podataka ukazuju da opisani postupak omogućava značajnu redukciju broja objekata u datom skupu podataka i da se optimalne vrednosti za parametar c nalaze u intervalu $[1.5, 2]$.

Napomena 2.1 Navedeni opis odabira parametra £′ je baziran na minimalnom rastojanju između tačaka. Predloženi pristup redukovanja složenosti se odnosi na velike skupove podataka sa numeričkim neprekidnim obeležjima. Međutim, kako je jedan od glavnih problema u klasterovanju, a ujedno i jedan od ciljeva ove disertacije, rešavanje problema klasterovanja u radu sa velikim skupovima podataka i kategorijalnim, odnosno kombinovanim obeležjima, primenićemo pristup zasnovan na korišćenju prostih slučajnih uzoraka umesto rada na celom skupu podataka. Ovaj pristup omogućava da se smanji vreme izvršavanja klaster algoritma, što ujedno i predstavlja veliki problem u radu sa velikim skupovima podataka. Detaljan opis postupka je dat u glavi 5.

2.6 OPTIMIZACIONI ALGORITAM KLASTEROVANJA SA TEŽINSKIM MERAMA RAZLIKE

Rezultati klasterovanja postaju manje precizni ukoliko su u analizu uključena obeležja koja nisu relevantna za pojedine klastere. Odabir relevantnih obeležja uglavnom predstavlja predkorak u postupku klasterovanja. *Chan* i saradnici [55] su predložili novi pristup u rešavanju ovog problema, korišćenjem težinskih mera razlike za objekte, pri čemu algoritam pronalazi težinu za svako obeležje u svakom klasteru.

Neka je $\boldsymbol{X}$ skup od m objekata opisanih pomoću n obeležja. Klaster algoritam za težinska obeležja, koji grupiše skup $\boldsymbol{X}$ u k klastera, zasniva se na minimizaciji funkcije cilja:

$$F(W, Z, \Lambda) = \sum_{l=1}^{k} \sum_{j=1}^{m} \sum_{i=1}^{n} w_{l,j}{}^{\mathsf{a}}{}_{l,i} d\left(z_{l,i}, x_{j,i}\right) \tag{2.6}$$

tako da važi

$$w_{l,j} \in \{0,1\}, \qquad 1 \le l \le k,\ 1 \le j \le m, \tag{2.7}$$

$$\sum_{l=1}^{k} w_{l,j} = 1, \quad 1 \le j \le m, \tag{2.8}$$

$$0 < \sum_{i=1}^{m} w_{l,j} < m, \quad 1 \le l \le k, \tag{2.9}$$

$$\mathsf{a}_{l,j} \ge 0, \qquad 1 \le l \le k,\ 1 \le i \le n, \tag{2.10}$$

$$\sum_{i=1}^{n} \mathsf{a}_{l,j} = 1, \quad 1 \le l \le k, \tag{2.11}$$

gde je $k(\le m)$ broj klastera, <1, $W = \left[w_{ij} \right]$ matrica celih brojeva reda $k \times m$, $Z = \left[z_1, z_2, \ldots, z_k \right] \in \mathbb{R}^{n \times k}$ sadrži centre klastera, $\Lambda = \left[\mathsf{a}_{l,i} \right]$ matrica realnih brojeva reda $k \times n$ i $d(z_{l,i}, x_{j,i}) \ge 0$ mera razlike između centra z_l i objekta x_j za $i-$to obeležje (detaljniji opis mera dat u poglavlju 3.1). Glavna ideja optimizacionog problema je minimizacija mere razlike između centara klastera i objekata. Mera razlike je definisana pomoću n težinskih obeležja. Tako definisana funkcija cilja nam omogućava da razmatramo težinu za svako obeležje u svakom klasteru.

Minimizacija funkcije $F(\cdot,\cdot,\cdot)$ u jednakosti (2.6) sa ograničenjima (2.7)-(2.11) formira klasu nelinearnih jednačina sa ograničenjima, čija su rešenja nepoznata. Uobičajeni metod optimizacije ove funkcije $F(\cdot,\cdot,\cdot)$ u (2.6) je korišćenje parcijalne optimizacije za Z, Λ i W. Prvo fiksiramo Z i Λ i nađemo potrebne uslove za W za minimizaciju $F(\cdot,\cdot,\cdot)$. Tada fiksiramo W i Λ i minimiziramo $F(\cdot,\cdot,\cdot)$ u odnosu na Z. Zatim fiksiramo W i Z i minimiziramo $F(\cdot,\cdot,\cdot)$ u odnosu na Λ. Postupak se ponavlja sve dok se ne postigne poboljšanje funkcije cilja.

Algoritam 2.5 Optimizacioni algoritam klasterovanja sa težinskim merama razlike

Korak 1. Izabrati početnu matricu $Z^{(1)} \in \mathbb{R}^{n\times k}$ *i neka je* $\Lambda^{(1)}$ *matrica reda* $k\times n$ *sa svim elementima jedankim* $1/n$. *Neka je* $t=1$.

Korak 2. Odredititi $W^{(t+1)}$ *takvo da je* $F(W^{(t+1)}, Z^{(t)}, \Lambda^{(t)})$ *minimizovana. Ukoliko je*

$$F(W^{(t+1)}, Z^{(t)}, \Lambda^{(t)}) = F(W^{(t)}, Z^{(t)}, \Lambda^{(t)}),$$

zaustaviti se; u suprotnom preći na korak 3.

Korak 3. Odredittii $Z^{(t+1)}$ *tako da je* $F(W^{(t+1)}, Z^{(t+1)}, \Lambda^{(t)})$ *minimizovana. Ukoliko je*

$$F(W^{(t+1)}, Z^{(t+1)}, \Lambda^{(t)}) = F(W^{(t+1)}, Z^{(t)}, \Lambda^{(t)}),$$

zaustaviti se; u suprotnom preći na korak 4.

Korak 4. Odrediti $\Lambda^{(t+1)}$ *tako da je* $F(W^{(t+1)}, Z^{(t+1)}, \Lambda^{(t+1)})$ *minimizovana. Ukoliko je*

$$F(W^{(t+1)}, Z^{(t+1)}, \Lambda^{(t+1)}) = F(W^{(t+1)}, Z^{(t+1)}, \Lambda^{(t)}),$$

zaustaviti se; u suprotnom vratiti se na korak 2.

Matrice Z, Λ i W se računaju u skladu sa sledećim teoremama.

Teorema 2.1 [55] *Neka su* $\tilde{Z}$ *i* $\tilde{\Lambda}$ *fiksirani. Minimizator matrice* $\tilde{W}$ *za optimizacioni problem*

$$\min_{W} F(W, \tilde{Z}, \tilde{\Lambda}), \quad \textit{pri čemu važi } (2.7)\text{-}(2.9)$$

je dat pomoću

$$\hat{w}_{l,j} = \begin{cases} 1, & \sum_{i=1}^{n} \tilde{a}_{l,i} \left(\tilde{z}_{l,i} - x_{j,i} \right)^2 \le \sum_{i=1}^{n} \tilde{a}_{h,i} \left(\tilde{z}_{h,i} - x_{j,i} \right)^2, \ 1 \le h \le k \\ 0, & \textit{inače} \end{cases}$$

Teorema 2.2 [55] *Neka su* $\tilde{W}$ *i* $\tilde{\Lambda}$ *fiksirani. Minimizator* $\tilde{Z}$ *optimizacionog problema*

$$\min_{Z} F(\tilde{W}, Z, \tilde{\Lambda})$$

je dat sa

$$\hat{z}_{l,i} = \frac{\sum_{j=1}^{m} \tilde{w}_{l,j} x_{i,j}}{\sum_{j=1}^{m} \tilde{w}_{l,j}}, \ \textit{gde je } 1 \le l \le k,$$

kada je i-ta promenljiva numerička, ili

$$\hat{z}_{l,i} = d_i^{(r)} \in DOM(D_i), \quad \textit{gde je}$$

$$\left| \left\{ \tilde{w}_{l,j} \middle| x_{i,j} = d_i^{(r)}, \tilde{w}_{l,j} = 1 \right\} \right| \ge \left| \left\{ \tilde{w}_{l,j} \middle| x_{i,j} = d_i^{(t)}, \tilde{w}_{l,j} = 1 \right\} \right|,$$
$$\forall t \in DOM\left(D_i\right),$$

kada je i-ta promenljiva kategorijalna. Sa $|Y|$ *je označena kardinalnost skupa* Y.

Teorema 2.3 [55] *Neka su* $\tilde{W}$ *i* $\tilde{Z}$ *fiksirani. Minimizator matrice* $\tilde{\Lambda}$ *za optimizacioni problem*

$$\min_{\Lambda} F(\tilde{W}, \tilde{Z}, \Lambda), \quad \textit{za koji važi (2.10) i (2.11)}$$

je dat pomoću

$$\hat{a}_{l,i} = \begin{cases} \dfrac{1}{n_i} & \sum_{j=1}^{m} \tilde{w}_{l,j} \left(\tilde{z}_{l,i} - x_{j,i} \right)^2 = 0, \\ & n_i = \left| \left\{ t : \sum_{j=1}^{m} \tilde{w}_{l,j} \left(\tilde{z}_{l,t} - x_{j,t} \right)^2 = 0 \right\} \right|, \\ 0 & \sum_{j=1}^{m} \tilde{w}_{l,j} \left(\tilde{z}_{l,i} - x_{j,i} \right)^2 \neq 0, \\ & (\exists t) \sum_{j=1}^{m} \tilde{w}_{l,j} \left(\tilde{z}_{l,t} - x_{j,t} \right)^2 = 0, \\ \dfrac{1}{\sum_{t=1}^{n} \left[\dfrac{\sum_{j=1}^{m} \tilde{w}_{l,j} \left(\tilde{z}_{l,i} - x_{j,i} \right)^2}{\sum_{j=1}^{m} \tilde{w}_{l,t} \left(\tilde{z}_{l,t} - x_{j,t} \right)^2} \right]^{1/(\beta - 1)}} & \sum_{j=1}^{m} \tilde{w}_{l,j} \left(\tilde{z}_{l,t} - x_{j,t} \right)^2 \neq 0, \ \forall 1 \leq t \leq n. \end{cases}$$

Algoritam 2.5 predstavlja klaster algoritam za težinska obeležja, u kome se matrica W određuje u skladu sa Teoremom 2.1, centri klastera Z u svakoj iteraciji u skladu sa Teoremom 2.2, a matrica težina za obeležja Λ u skladu sa Teoremom 2.3.

Teorema 2.4 [55] *Optimizacioni klaster algoritam* (*Algoritam 2.5*) *konvergira u konačnom broju iteracija.*

Na osnovu teoreme 2.4, ovaj algoritam za ponderisana obeležja konvergira. Međutim, zaustavlja se u lokalnom minimumu [36]. Vremenska složenost algoritma iznosi $O(l \times m \times n \times k)$, a prostorna složenost $O(m \times (n+k) + 2k \times n)$ gde je l ukupan broj iteracija, k je broj klastera, n broj obeležja i m broj objekata u posmatranom skupu. Dakle, predloženi algoritam je prilagođen za rad sa velikim skupovima podataka.

3. ANALIZA KLASTER ALGORITAMA

Postupak klaster analize se sastoji iz dva osnovna koraka, izbora odgovarajuće mere udaljenosti (sličnosti) i izbora algoritma klasterovanja, to jest niza procedura za grupisanje objekata tako da postoje male razlike unutar klastera, a velike razlike između klastera. Podaci se iz posmatranog skupa grupišu u klastere na osnovu mera sličnosti (ili udaljenosti) između dva različita objekta. Ne postoji slaganje oko toga koja mera udaljenosti je najadekvatnija za primenu u klasterovanju.

3.1 MERE SLIČNOSTI I RAZLIČITOSTI IZMEĐU OBJEKATA

Bitna stavka kod grupisanja podataka jeste poznavanje koliko su objekti međusobno bliski, odnosno koliko su oni udaljeni. Razmatramo m objekata $\{x_1, x_2, \cdots, x_m\}$ koji su opisani sa n obeležja.

Za meru $d_{ij} = d(x_i, x_j)$ kažemo da predstavlja **meru različitosti (mera rastojanja, mera metrike)** objekata x_i i x_j, gde je $i, j \le n$, ako zadovoljava sledeće osobine:

1. $d_{ij} > 0$, ako se objekti i i j razlikuju, a $d_{ij} = 0$, samo ako su objekti identični (*uslov nenegativnosti*)
2. $d_{ij} = d_{ji}$ (*uslov simetričnosti*)
3. $d_{ij} \le d_{ik} + d_{kj}$ za sve objekte i, j i k (*uslov triangularnosti*)

Za meru $s_{ij} = s(x_i, x_j)$ kažemo da predstavlja **meru sličnosti** objekata x_i i x_j, gde je $i, j \le n$, ako zadovoljava sledeće osobine:

1. $0 \le s_{ij} \le 1$, za sve objekte i i j (*uslov normiranosti*)
2. $s_{ij} = 1$, samo ako su objekti i i j identični
3. $s_{ij} = s_{ji}$ (*uslov simetričnosti*).

Kada su sve promenljive neprekidne, najčešće se koriste mere rastojanja (metrike). Veliki broj ovih mera se može generisati pomoću L_r-norme (*Minkovski*), $r \ge 1$

$$d_{ij} = \left\| x_i - x_j \right\|_r = \left\{ \sum_{s=1}^{n} \left| x_{is} - x_{js} \right|^r \right\}^{1/r}.$$

Ovde x_{is} označava vrednost $s-$te promenljive za objekat x_i. Najčešće korišćene mere rastojanja, *Euklidova metrika (mera)* i *Apsolutna (blok, Manhattan) metrika* su specijalni slučajevi rastojanja *Minkowskog*, za $r=2$ i $r=1$, respektivno. U tabeli 3.1 su prikazane najčešće korišćene mere rastojanja za podatke sa numeričkim obeležjima.

Tabela 3.1. Mere rastojanja (metrike) i sličnosti za numerička obeležja

Euklidsko rastojanje	$d_{ij}=\left(\sum_{s=1}^{n}\left(x_{is}-x_{js}\right)^2\right)^{1/2}$
Manhattan rastojanje	$d_{ij}=\sum_{s=1}^{n}\left\|x_{is}-x_{js}\right\|$
Minkowski rastojanje	$d_{ij}=\left(\sum_{s=1}^{n}\left(x_{is}-x_{js}\right)^r\right)^{1/r} \quad r\geq 1$
Canberra rastojanje	$d_{ij}=\begin{cases}0 & za\ \ x_{is}=x_{js}=0\\ \sum_{s=1}^{n}\left\|x_{ik}-x_{js}\right\|/\left(\left\|x_{is}\right\|+\left\|x_{js}\right\|\right) & za\ \ x_{is}\neq 0\ ili\ x_{js}\neq 0\end{cases}$
Pearson-ova korelacija	$d_{ij}=\left(1-¤_{ij}\right)/2$, gde je $¤_{ij}=\frac{\sum_{s=1}^{n}\left(x_{is}-\bar{x}_{i\cdot}\right)\left(x_{js}-\bar{x}_{j\cdot}\right)}{\left(\sum_{s=1}^{n}\left(x_{is}-\bar{x}_{i\cdot}\right)^2\sum_{s=1}^{n}\left(x_{js}-\bar{x}_{j\cdot}\right)^2\right)^{1/2}}$ $\bar{x}_{i\cdot}=\frac{1}{n}\sum_{s=1}^{n}x_{is}$
Uglovna separacija	$¢_{ij}=\left(1-¤_{ij}\right)/2$, gde je $¤_{ij}=\frac{\sum_{s=1}^{n}x_{is}x_{js}}{\left(\sum_{s=1}^{n}x_{is}^{\ 2}\sum_{s=1}^{n}x_{js}^{\ 2}\right)^{1/2}}$

Kada su sve promenljive kategorijalne (više od dve kategorije), a ne koristi se pristup transformisanja ovih promenljivih u binarne, tada se obično koriste mere različitosti. Označimo sa x_{is} vrednost s-tog kategorijalnog obeležja $(s=1,\ldots,n)$ za objekat x_i $(i=1,\ldots,m)$. Huang [127] je uveo meru različitosti d_{ij} za dva objekta x_i i x_j sa n kategorijalnih obeležja koja predstavlja ukupan broj nepoklapanja vrednosti dva objekta:

$$d_{ij} = \sum_{s=1}^{n} ¢_{i,j,s},$$

gde je

$$¢_{i,j,s} = \begin{cases} 0, & x_{i,s} = x_{j,s}, \\ 1, & x_{i,s} \neq x_{j,s}. \end{cases}$$

Ova mera različitosti ima veliku primenu, kao što ćemo kasnije videti kod algoritma *k-modusa* [127] i njegovih brojnih modifikacija.

Dalje ćemo analizirati slučaj kada kategorijalna promenljiva ima dva nivoa, to jest slučaj binarne promenljive. U cilju merenja sličnosti objekata uvek upoređujemo parove observacija, (x_i, x_j), gde je $x_i^{\mathrm{T}} = (x_{i1}, \ldots, x_{in})$, $x_j^{\mathrm{T}} = (x_{j1}, \ldots, x_{jn})$, i $x_{is}, x_{js} \in \{0,1\}$. Moguća su četiri slučaja: $x_{is} = x_{js} = 1$, $x_{is} = 0, x_{js} = 1$, $x_{is} = 1, x_{js} = 0$, $x_{is} = 0, x_{js} = 0$, pa definišemo:

$$a_1 = \sum_{s=1}^{n} I(x_{is} = x_{js} = 1),$$

$$a_2 = \sum_{s=1}^{n} I(x_{is} = 0, x_{js} = 1),$$

$$a_3 = \sum_{s=1}^{n} I(x_{is} = 1, x_{js} = 0),$$

$$a_4 = \sum_{s=1}^{n} I(x_{is} = x_{js} = 0).$$

Uočimo da svako a_l, $l = 1, \ldots 4$, zavisi od para (x_i, x_j) (tabela 3.2).

Tabela 3.2 Tabela kontigencije za par (x_i, x_j)

		Objekat i		
Objekat j		1	0	Ukupno
	1	a_1	a_3	$a_1 + a_3$
	0	a_2	a_4	$a_2 + a_4$
	Ukupno	$a_1 + a_2$	$a_3 + a_4$	$a_1 + a_2 + a_3 + a_4$

Popularna grupa mera koja se koristi za binarne podatke je poznata pod zajedničkim nazivom *koeficijenti poklapanja* i može se prikazati kao:

$$s_{ij} = \frac{a_1 + ¢\, a_4}{a_1 + ¢\, a_4 + ª(a_2 + a_3)},$$

gde su ¢ i ª težinski koeficijenti. U tabeli 3.3 su prikazane najčešće korišćene mere sličnosti (za date vrednosti težinskih koeficijenata).

Tabela 3.3 Mere (koeficijenti) sličnosti za binarne promenljive

Naziv mere	¢	a	Definisanje
Jaccard [129]	0	1	$\frac{a_1}{a_1+a_2+a_3}$
Rogers and Tanimoto [210]	1	2	$\frac{a_1+a_4}{a_1+2(a_2+a_3)+a_4}$
Simple Matching (M) [231]	1	1	$\frac{a_1+a_4}{a_1+a_2+a_3+a_4}$
Russel and Rao (RR) [218]	-	-	$\frac{a_1}{a_1+a_2+a_3+a_4}$
Dice [74]	0	0.5	$\frac{2a_1}{2a_1+(a_2+a_3)}$
Kulczynski [149]	-	-	$\frac{a_1}{a_2+a_3}$
Gower and Legendre [103]	1	0.5	$\frac{a_1+a_4}{a_1+0.5(a_2+a_3)+a_4}$
Sokal and Sneath [232]	0	2	$\frac{a_1}{a_1+2(a_2+a_3)}$

Postoje brojni pristupi za kreiranje mera sličnosti za kombinovane podatke, to jest za podatke sa neprekidnim i kategorijalnim promenljivim. Jedna od mogućnosti je da se kreira mera razlike za svaki tip promenljive, a zatim ih kombinuje (sa ili bez njihovih težina) u jedan koeficijent. Drugačiji pristup daje *Gower* [104]. Mera koju on predlaže je korisna za kombinaciju neprekidnih i binarnih tipova podataka, kao i u slučaju rada sa nedostajućim podacima. Gouerova (*Gower*) uopštena mera sličnosti je data sledećim koficijentom:

$$s_{ij} = \frac{\sum_{l=1}^{n} w_{ijl} s_{ijl}}{\sum_{l=1}^{n} w_{ijl}},$$

gde je s_{ijk} sličnost između i-tog i j-tog objekta za k-tu promenljivu, a w_{ijk} je 0 ili 1. Važi da je $w_{ijk} = 0,$ ako vrednost k-te promenljive nedostaje za bar jedan od objekata, ili je k-ta promenljiva binarna i važi $x_{il} = x_{jl} = 0$. Za binarne promenljive i kategorijalne promenljive sa više od dve kategorije, s_{ijk} uzima vrednost 1 kada objekti imaju istu vrednost, a 0 inače. Za neprekidne promenljive *Gower* mera sličnosti ima sledeći oblik:

$$s_{ijl} = 1 - \frac{\left| x_{il} - x_{jl} \right|}{R_l},$$

gde je R_l opseg vrednosti za l-tu promjenljivu.

3.2 KLASIFIKACIJA ALGORITAMA KLASTEROVANJA

Postoje različiti algoritmi za rešavanje problema klasterovanja [82, 105,131,204,211,267]. *Halkidi* i saradnici [112] navode klasifikaciju algoritama u odnosu na:

- tipove podataka koji se unose u algoritam
- kriterijum klasterovanja koji definiše sličnost između podataka
- teorijske i fundamentalne koncepte na kojima su zasnovane tehnike klaster analize (npr. *fuzzy* teorija, statistika).

Klaster algoritmi se mogu klasifikovati u odnosu na tipove promenljivih:

- ***Statističke***
- ***Konceptualne***

Statistički algoritmi klasterovanja su bazirani na konceptima statističke analize. Ovi algortimi koriste mere sličnosti za podelu objekata i ograničeni su isključivo na numeričke podatke. *Konceptualni algoritmi klasterovanja* se koriste za klasterovanje kategorijalnih podataka. Ovaj vid klasterovanja je zasnovan na zajedničkim osobinama, tako da klasteri dele neku zajedničku osobinu ili predstavljaju pojedinačni koncept.

Još jedna klasifikacija algoritama klasterovanja je na osnovu prisutnog preklapanja klastera:

- **Rasplinuto *(eng. fuzzy)* klasterovanje**
- **Tvrdo *(eng. hard, crisp)* klasterovanje**

Rasplinuto klasterovanje koristi *fuzzy* tehnike, pri čemu važi da jedan objekat može biti klasifikovan u više od jednog klastera. Najpoznatiji ovakav algoritam je *Fuzzy C-Means* [37]. *Tvrdo klasterovanje* razmatra nepreklapajuću podelu, što znači da svaka tačka pripada tačno jednom klasteru. Većina prisutnih algoritama klasterovanja pripada ovoj kategoriji algoritama.

U literaturi je zastupljena i sledeća podela metoda klasterovanja:

- **Hijerarhijske metode**
- **Nehijerarhijske metode (metode raščlanjivanja)**
- **Metode zasnovane na modelu**
- **Metode bazirane na gustini**
- **Metode bazirane na mreži**

Najčešće se koristi podela na **hijerarhijske metode** i **nehijerarhijske metode klasterovanja.**

3.2.1 Hijerarhijske metode

Hijerarhijske metode daju niz sukcesivnih particija skupa na klastere, pri čemu se prvo vrše izračunavanja udaljenosti svih jedinica međusobno, a zatim se klasteri formiraju pomoću tehnika spajanja ili razdvajanja. Formira se skup ugneždenih klastera organizovanih u obliku drveta, koji se najčešće prikazuje pomoću dijagrama- *dendrograma.* Hijerarhijske metode mogu da se klasifikuju u **metode udruživanja** ili **sakupljajuće metode** (eng. *agglomerative*) i **metode deobe** ili **razdvajajuće metode** (eng. *divisiv*e), u zavisnosti od toga kako je formirana hijerarhijska dekompozicija.

Najćešće se koriste hijerarhijske metode udruživanja, koje počinju od m objekata (tj. svaka tačka je jedan klaster) i sekvencijalno se spajaju u veće klastere, a ređe hijerarhijske metode deobe kada se polazi od jednog klastera (ceo posmatrani skup), koji se zatim deli na manje klastere. Metode udruživanja se razlikuju prema načinu na koji se procenjuje udaljenost između klastera u sukcesivnim koracima. Najčešće se koristi *Lance-Wiliams*ova grupa metoda:

- ***Metod centroida***. Udaljenost između klastera je predstavljena pomoću udaljenosti između centroida. Dve klastera se udružuju ukoliko su njihovi centroidi najmanje udaljeni međusobno u odnosu na međusobno rastojanje svih parova klastera koje postoje na posmatranom nivou udruživanja.
- ***Metod jednostrukog (prostog) povezivanja*** (*eng. single linkage),* poznata i kao *metoda najbližeg suseda*. Mera rastojanja između dva klastera predstavlja minimalno rastojanje između parova obekata koji pripadaju ovim klasterima.
- ***Metod potpunog (kompletnog) povezivanja*** (*eng. complete linkage),* poznata i kao *metoda najdaljeg suseda*. Rastojanje između dva klastera predstavlja maksimalno rastojanje između parova objekata koji pripadaju tim klasterima.
- ***Metod prosečnog povezivanja,*** ili ***metod proseka*** (*eng. average linkage*). Rastojanje se određuje prema prosečnom rastojanju svih objekata koji pripadaju dvema grupama.
- ***Metod Ward-a,*** poznata i kao ***metoda minimalne varijanse***. Kao i ostakle metode povezivanja, kreće se od m klastera (svaki klaster sadrži jedan objekat), ali se ne računa udaljenost između klastera, već se maksimizira homogenost unutar klastera. Ukupna suma kvadrata unutar klastera (SSE) se računa u cilju utvrđivanja koje se dve grupe spajaju u svakom koraku algoritma. Suma kvadrata greške (SSE) je definisana kao:

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \left(x_{ij} - x^i \right)^2,$$

gde je x_{ij} je j-ti objekat u i-tom klasteru, k je broj klastera, $x^i = \overline{x}_i$ centar i-tog klastera, a m_i je broj objekata u i-tom klasteru. Korišćenje Ward-ove metode ne zahteva pretpostavke o multivarijantnoj normalnoj raspodeli promenljivih.

Prednost hijerarhijskih metoda je to što nisu potrebne pretpostavke o broju klastera. Nedostatak ovih metoda predstavlja to da jednom spojena dva objekta ne mogu biti ponovo razdvojena, niti razdvojeni objekti mogu biti ponovo spojeni. Nedostaci ovih metoda su i osetljivost na šum i autlajere (elemente van granica), rad sa klasterima različite veličine, kao i rad sa klasterima konveksnog oblika. Prostorna složenost ovih algoritama iznosi $O(m^2)$, a vremenska složenost iznosi $O(m^3)$, gde je m broj slučajeva. Najpoznatiji hijerarhijski algoritmi klasterovanja su: BIRCH (*Balanced Iterative Reducing and Clustering*) [270,271], CURE (*Clustering Using REpresentatives*) [107], ROCK (*RObust Clustering using linKs)* [108].

3.2.2 Nehijerarhijske metode

Za dati skup od m objekata, nehijerarhijske metode konstruišu samo jednu optimalnu podelu (particiju) skupa podataka na k grupa (klastera). Nehijerahijski metod klasterovanja funkcioniše na sledeći način: za dati broj klastera, izvrši se inicijalna podela; objekti se zatim premeštaju između klastera sa ciljem poboljšanja funkcije cilja. Ove metode su poznate i pod nazivom k-grupisanje. Generalno važi da svaki klaster sadrži bar jedan objekat i svaki objekat pripada tačno jednom klasteru (nepreklapajuće klasterovanje). Nasuprot hijerarhijskim metodama klasterovanja, nehijerarhijske metode ne podrazumevaju grafički prikaz podataka pomoću stabla. Za razliku od hijerarhijskih metoda, ovde se dozvoljava premeštanje objekata iz ranije formiranih grupa. Nehijerarhijske metode klasterovanja su brže, pouzdanije od hijerarhijskih, pretpostavlja se da je broj klastera poznat unapred, ili kao kod nekih metoda, varira tokom postupka klasterovanja. Uglavnom primena nehijerarhijskih algoritama klasterovanja podrazumeva korišćenje ili modifikaciju jedne od dve najpopularnije heurističke metode: algoritam k-*sredina*, gde je svaki klaster predstavljen pomoću prosečne vrednosti objekata u klasteru i algoritam k-*medoida* [136], gde je centar klastera realan objekat u klasteru. Ovi algoritmi dobro funkcionišu u pronalaženju klastera sferičnog oblika, kao i prilikom rada sa malim do umereno velikim bazama podataka. Za pronalaženje klastera sa kompleksnijim oblicima i za klasterovanje velikog skupa podataka, potrebno je proširenje ovih metoda.

3.2.3 Metode zasnovane na gustini

Osnovna ideja ovih metoda je da se grupišu susedni objekti iz skupa podataka u klastere na osnovu uslova gustine. Ovakav metod može biti koristan u filtriranju šuma i otkrivanju klastera proizvoljnog oblika. Najpoznatije su DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) [84] i OPTICS (*Ordering Points To Identify the Clustering Structure*)[14].

3.2.4 Metode zasnovane na mreži

Ove metode kvantifikuju prostor objekata u konačan broj ćelija koje formiraju rešetkastu (mrežnu) strukturu. Glavna prednost ovog pristupa je njegovo brzo vreme izvođenja, koje je nezavisno od broja objekata i zavisi samo od broja ćelija u svakoj dimenziji prostora. Tipičan predstavnik ove metode je algoritam STING (*STatistical INformation Grid*) [248].

3.2.5 Metode zasnovane na modelu

Ove metode pretpostavljaju model za svaki klaster i pronalaze najbolje fitovanje podataka za dati model. Algoritam klasterovanja baziran na modelu može locirati klastere pomoću konstruisane funkcije gustine koja odražava prostornu distribuciju podataka (tačaka). Takve metode su često bazirane na pretpostavci da su podaci generisani pomoću kombinovanja osnovnih raspodela verovatnoća. Metode klasterovanja zasnovane na modelu slede dva glavna pristupa: statistički pristup ili pristup zasnovan na neuronskim mrežama.

3.3 KLASTEROVANJE VELIKIH SKUPOVA PODATAKA SA KATEGORIJALNIM I KOMBINOVANIM TIPOVIMA OBELEŽJA

Problem koji se javlja kod primene različitih algoritama klasterovanja je rad sa velikim brojem podataka i velikim brojem obeležja. Kada se govori i velikom skupu podatka, sama definicija "veliki skup" je prilično neodređena. U poglavlju 2.3 opisan je algoritam k*-sredina*, za klasterovanje velikih skupova podatka sa neprekidnim numeričkim promenjivima.

U literaturi postoje različiti prostupi za redukovanje velikih skupova podataka pomoću manjih reprezentativnih podskupova, a u cilju smanjenja vremenske kompleksnosti klaster algoritama [131, 132, 267]. Navešćemo samo neke od ovih pristupa, bez obzira na vrstu promenljivih uključenih u analizu. *Kaufman* i *Rousseeuw* (1990) su predložili CLARA (*Clustering LARge Applications*) algoritam za klasterovanje velikih skupova podataka [136], koji predstavlja kombinaciju postupka uzorkovanja i PAM (*Partitioning Around Medoids*) algoritma klasterovanja. CLARA algoritam primenjuje mali uzorak iz velikog skupa podataka, koristi PAM za generisanje k-medoida iz uzorka, koji dalje služe za klasterovanje ostatka skupa. Računska složenost CLARA algoritma u pojedinačnoj iteraciji iznosi $O\left(km_u^{\,2}+k(m-k)\right)$, gde je m obim skupa, k broj klastera, a $m_u = 40+2k$ obim uzorka. CLARANS **(***Clustering Large Applications based on RANdom Search*) je algoritam koji su predložili *Ng* i *Han* (1994) kao način da poboljšaju CLARA metod [189]. Ovaj metod identifikuje kandidate za centroide klastera korišćenjem ponovljenih slučajnih uzoraka iz originalnog skupa podataka. Autori tvrde da on obezbeđuje bolje klastere pomoću malog broja "traženja" (poređenje k alternativnih objekata kao predstavnika klastera). Vremenska složenost ovog algoritma iznosi O(m). Osim ovih algoritama koji su zasnovani na kombinaciji uzorkovanja i primene klaster algoritma na ovim uzorcima, postoje i drugačiji pristupi u rešavanju problema klasterovanja velikih skupova podataka [131,132]. *Algoritmi klasterovanja zasnovani na razlaganju* ili postupku „podeli pa vladaj" (eng. *divide and conquer*) rekurzivno razbijaju problem na dva ili više potproblema dok oni ne postanu dovoljno jednostavni da se mogu direktno rešiti [12]. Ova tehnika se sastoji u podeli skupa podataka dimenzije $m\times n$ u p disjunktnih blokova i zatim odvojeno klasterovanje tih skupova, pri čemu se optimalan broj blokova p može odrediti na osnovu algoritma *Murty and Krishna* [185]. Konačni skup klastera predstavlja uniju klastera ovih skupova ili se dobija kombinovanjem i/ili prečišćavanjem odvojenih skupova klastera.

Sa stanovišta ciljnog skupa podataka u klaster analizi, postojeći algoritmi se mogu svrstati u tri kategorije: *numerički, kategorijalni, i kombinovani*. Većina algoritama klasterovanja je fokusirana na numeričke podatke čije geometrijske osobine mogu biti prirodno iskorišćene za

definisanje funkcije udaljenosti između podataka, kao što su algortimi *DBSCAN, BIRCH, CURE, CHAMELEON* [135]. Numerički algoritmi klasterovanja ne odgovaraju kategorijalnim obeležjima, te je dakle lako zaključiti da oni takođe nisu pogodni za klasterovanje obeležja koja su kombinovanog tipa. Poslednjih godina predloženi su neki efikasni algoritmi za klasterovanje kategorijalnih podataka. Međutim, svi ovi algoritmi su predviđeni za kategorijalna obeležja i nisu poznate njihove potencijalne mogućnosti u klasterovanju kombinovanih tipova obeležja.

Tradicionalni način tretiranja kategorijalnih obeležja kao numeričkih ne daje uvek značajne rezultate, zato što mnogi kategorijalni domeni nisu ordinalni. *Ralambondrainy* (1995) je prikazao pristup koristeći algoritam *k-sredina* u klasterovanju kategorijalnih podataka (sa nominalnom skalom merenja) [203]. On je transformisao kategorijalna obeležja u binarna obeležja (dodeljujući vrednost 0 ukoliko je kategorija odsutna, odnosno 1, ako je prisutna) i tretirao binarna obeležja kao numerička u algoritmu *k-sredina*. Prvi nedostatak ovog pristupa je da on podrazumeva rad sa velikim brojem binarnih obeležja, jer skupovi podataka u istraživanju podatka imaju stotine ili hiljade kategorija, što neminovno povećava računarske i prostorne troškove algoritma. Drugi nedostatak ovog pristupa je da centri, tj. aritmetičke sredine klastera, čije vrednosti su realni brojevi između 0 i 1, ne pokazuju stvarne karakteristike klastera.

Huang je predložio dva algoritma, algoritam *k-modusa* [125] i *k-prototip* [126,127], koji proširuju algoritam *k-sredina* na podatke sa kategorijalnim obeležjima, odnosno kombinovanim tipovima obeležja. *Chaturvedi* i saradnici [56] navode sledeće tehnike koje se primenjuju u klasterovanju kategorijalnih podataka:

1. Transformacija kategorijalnih promenljivih u veštačke (*dummy*) promenljive i korišćenje hijerarhijskih algoritama, odnosno algoritma *k*-sredina
2. Primena analize korespodencije za dobijanje prostornih koordinata za svaki subjekat, a zatim primena algoritma *k-sredina* na dobijene koordinate
3. Primena procedura za identifikaciju latentnih klasa na osnovu tabela kontigencije
4. Korišćenje *Hartigan-Ditto* algoritma za kategorijalne podatke

LIMBO [13] je hijerarhijski skalabilni algoritam za kategorijalne podatke, koji koristi informacioni teorijski koncept za definisanje kvaliteta klasterovanja. Kao hijerarhijski algoritam, LIMBO ima prednost da proizvodi klastere različite veličine, ali nije tako brz kao nehijerarhijski algoritmi. ROCK algoritam se primenjuje, kako za numeričke, tako i za kategorijalne podatke. Osnovna ideja je da su podaci slični ukoliko imaju dovoljno zajedničkih suseda (tj. veza). Ovakakv koncept veza koristi više globalnu

informaciju o prostoru klastera u poređenju sa merom sličnosti, rastojanja, gde se samo razmatra lokalno rastojanje između dve tačke. Nedostatak mu je osetljivost na selekciju parametara, kao i prisustvo šuma.

U literaturi se sreću i različiti algoritmi za klasterovanje kombinovanih tipova obeležja. Chiu i saradnici [59] su predložili *Dvostepeni klaster algoritam (Twostep*) u radu sa kombinovanim tipovima promenljivih. *Li* i *Biswas* su predložili SBAC algoritam (*Similarity Based Agglomerative Clustering*), zasnovan na *Goodall* meri sličnosti [155]. Mada dobro funkcioniše u radu sa kombinovanim tipovima obeležja, ovaj algoritam je računski veoma skup. *Ahmad and Dey* koriste klaster algoritam zasnovan na algoritmu k*-sredina* koji prevazilazi nedostatke k*-prototip* algoritma [4,5].

3.3.1 Algoritam k*-modusa*

Algoritam k*-modusa* predstavlja modifikaciju algoritma k*-sredina* za klasterovanje kategorijalnih podataka korišćenjem:

1. mere različitosti za kategorijalne objekte
2. modusa umesto sredina klastera
3. metode zasnovane na frekvenciji za ažuriranje modusa.

Pretpostavimo da je skup m kategorijalnih objekata $X = \{x_1, x_2, ..., x_m\}$ definisan pomoću skupa n kategorijalnih obeležja $D_1, D_2, ..., D_n$. Svako obeležje D_j je opisano konačnim domenom vrednosti, $DOM(D_j)$, za koji važi: za svako $a, b \in DOM(D_j)$, $a = b$, ili $a \neq b$. Svaki objekat skupa X možemo prikazati kao konjukciju parova vrednosti $[D_1 = x_1] \wedge [D_2 = x_2] \wedge \cdots \wedge [D_n = x_n]$, gde je $x_j \in DOM(D_j)$ za $1 \leq j \leq n$. Objekat x_i je predstavljen sa $[x_{i,1}, x_{i,2}, ..., x_{i,n}]$, gde je $x_{i,j} \in DOM(D_j)$ i važi $x_i = x_s$, ukoliko je $x_{i,j} = x_{s,j}$ za $i, s = 1, \ldots, m$, $1 \leq j \leq n$. Mera različitosti dva kategorijalna objekta $x = [x_1, x_2, ..., x_n]$ i $y = [y_1, y_2, ..., y_n]$ se može definisati ukupnim brojem nepoklapanja između odgovarajućih kategorija obeležja za dva objekta [136]:

$$d(x, y) = \sum_{j=1}^{n} ¢(x_j, y_j), \text{ gde je} \tag{3.1}$$

$$¢(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases} \tag{3.2}$$

Definisaćemo prvo modus skupa.

Definicija 3.1 *Modus skupa* $X = \{x_1, x_2, ..., x_m\}$ *je vektor* $q = [q_1, q_2, ..., q_n]$ *takav da minimizira:*

$$d(X, q) = \sum_{i=1}^{m} d(x_i, q), \tag{3.3}$$

gde je mera $d(x_i, q)$ *definisana u (3.1).*

Modus q ne mora biti element skupa X.

Sada ćemo opisati način za pronalaženje modusa skupa. Neka je sa $m_{c_{k,j}}$ označen broj objekata koji imaju k-tu kategoriju $c_{k,j}$ obeležja D_j i neka je $f_r(D_j = c_{k,j} | X) = \frac{m_{c_{k,j}}}{m}$ relativna frekvencija kategorije $c_{k,j}$ u X.

Teorema 3.1 [127] *Funkcija* $d(X, q)$ definisana u (3.3) *dostiže svoj minimum ako i samo ako važi da je*

$$f_r(D_j = q_j | X) \geq f_r(D_j = c_{k,j} | X) \text{ za } q_j \neq c_{k,j}, \text{ za sve } j = 1, \ldots, n.$$

Algoritam k*-modusa* se sastoji iz sledećih koraka:

1. *Izabrati inicijalnih* k*-modusa, jedan za svaki klaster*

2. *Dodeliti objekt u klaster čiji je modus najbliži, na osnovu mere* $d(x, y)$ *u (3.1). Ažurirati moduse klastera, u skladu sa Teoremom 3.1.*

3. *Nakon dodeljivanja svih objekata u klastere, ponovo testirati udaljenost objekata u odnosu na moduse klastera. Ukoliko se utvrdi da postoji objekat takav da njegov najbliži modus pripada drugom klasteru, dodeliti ga tom drugom klasteru i ponovo odrediti moduse za oba klastera*

4. *Ponavljati korak 3, sve dok se ne poklope centri poslednje dve iteracije*

Teorema 3.1 definiše način određivanja modusa za dati skup podataka, što dalje omogućava korišćenje definisanog algoritma k-sredina u slučaju kategorijalnih podataka.

Klaster algoritam k-*modusa*, koji grupiše skup od m objekata u k klastera, zasniva se na minimizaciji funkcije cilja:

$$F(W,Z)=\sum_{l=1}^{k}\sum_{i=1}^{m}w_{li}d(z_l,x_i) \tag{3.4}$$

tako da važi

$$w_{li}\in\{0,1\},\quad 1\le l\le k,\ 1\le i\le m, \tag{3.5}$$

$$\sum_{l=1}^{k}w_{li}=1,\ \ 1\le i\le m, \tag{3.6}$$

$$0<\sum_{i=1}^{m}w_{li}<m,\ \ 1\le l\le k, \tag{3.7}$$

gde je $k(\le m)$ poznat broj klastera, $W=[w_{li}]$ matrica $\{0,1\}$ reda $k\times m$, $Z=[z_1,z_2,...,z_k]$ i $z_i\in\mathbb{R}^n$ je centar i-tog klastera sa kategorijalnim obeležjima $D_1,D_2,\ldots,D_n$.

Kao i algoritam k-*sredina*, ovaj algoritam takođe proizvodi lokalna optimalna rešenja koja zavise od početnih modusa i redosleda objekata u skupu podataka. Minimizacija funkcije F date u (3.4) sa ograničenjima (3.5)-(3.7) predstavlja problem nelinearne optimizacije sa ograničenjima čija su rešenja nepoznata. Uobičajeni metod optimizacije podrazumeva korišćenje parcijalne optimizacije za Z i za W. Ovaj postupak se može prikazati na sledeći način:

Algoritam 3.1 A*lgoritam* $k-$*modusa*

Korak 1. Izabrati početnu tačku $Z^{(1)}\in\mathbb{R}^{mk}$. *Odrediti* $W^{(1)}$ *takvo da je* $F(W,Z^{(1)})$ *minimalno. Neka je* $t=1$

Korak 2. Odrediti $Z^{(t+1)}$ *takvo da je* $F(W^{(t)},Z^{(t+1)})$ *minimalno. Ukoliko je* $F(W^{(t)},Z^{(t+1)})=F(W^{(t)},Z^{(t)})$, *zaustaviti se; u suprotnom preći na korak 3.*

Korak 3. Odrediti $W^{(t+1)}$ *takvo da je* $F(W^{(t+1)},Z^{(t+1)})$ *minimalno. Ako je* $F(W^{(t+1)},Z^{(t+1)})=F(W^{(t)},Z^{(t+1)})$, *zaustaviti se; u suprotnom* $t\to t+1$ *i vratiti se na korak 2.*

Napomena 3.1. Za razliku od algoritma 2.5. koji se primenjuje u klasterovanju numeričkih i kategorijalnih obeležja i koristi težinske mere razlike, algoritam 3.1 se primenjuje u klasterovanju kategorijalnih podataka i ne koriste se težinska obeležja, te nemamo parcijalnu optimizaciju u odnosu

na matricu težina Λ. Vremenska složenost algoritma iznosi $O(l \times k \times m)$, gde je l broj iteracija, k broj klastera, a m broj objekata u skupu podataka.

Problem optimizacije u klasterovanju pomoću algoritma *k-modusa* se može rešiti iterativno rešavanjem sledeća dva problema minimizacije:
1. Problem P_1: Fiksirati $Z = \hat{Z}$, rešiti redukovani problem $F(W, \hat{Z})$
2. Problem P_2: Fiksirati $W = \hat{W}$, rešiti redukovani problem $F(\hat{W}, Z)$.
Problemi P_1 i P_2 su rešeni u skladu sa sledećim teoremama redom:

Teorema 3.2 [188] *Neka je* $\hat{Z}$ *fiksirano i razmatramo problem:*

$$\min_W F(W, \hat{Z}) \quad \textit{za koji važi (3.5), (3.6) i (3.7).}$$

Minimum $\hat{W}$ *je dat sa:*

$$w_{il} = \begin{cases} 1, & d(\hat{z}_l, x_i) \le d(\hat{z}_h, x_i), \quad 1 \le h \le k, \\ 0, & d(\hat{z}_l, x_i) > d(\hat{z}_h, x_i), \quad 1 \le h \le k. \end{cases}$$

Teorema 3.3 [188] *Neka je* X *skup kategorijalnih objekata opisanih pomoću* n *kategorijalnih obeležja* $D_1, D_2, \ldots, D_n$ *i* $DOM(D_j) = \left\{ d_j^{(1)}, d_j^{(2)}, \ldots, d_j^{(n_j)} \right\}$, *gde je* n_j *broj kategorija obeležja* D_j, *za* $1 \le j \le n$. *Neka su centri klastera* z_l *dati sa* $\left[z_{l,1}, z_{l,2}, \ldots, z_{l,n} \right]$ *za* $1 \le l \le k$. *Tada,* $\sum_{l=1}^{k} \sum_{i=1}^{m} w_{li} d(z_l, x_i)$ *dostiže svoj minimum ako i samo ako je*

$$z_{l,j} = d_j^{(r)} \in DOM(D_j), \quad \textit{gde je}$$

$$\left| \left\{ w_{li} \,\middle|\, x_{i,j} = d_j^{(r)}, w_{li} = 1 \right\} \right| \ge \left| \left\{ w_{li} \,\middle|\, x_{i,j} = d_j^{(t)}, w_{li} = 1 \right\} \right|, \quad 1 \le t \le n_j \quad \textit{za } 1 \le j \le n.$$

Korišćenje mere različitosti u algoritmu *k-modusa* može prouzrokovati probleme prilikom dodeljivanja objekata. U rešavanju problema P_1 za algoritam *k*-modusa, objekti su dodeljeni u skladu sa teoremom 3.2, to jest svaki objekat je dodeljen najbližem klasteru. Međutim, mera različitosti je 0 ili 1, što ne reprezentuje uvek stvarnu udaljenost između objekta i klastera. Različiti istraživači se bave problemom pronalaženja početnih modusa u algortimu *k-modusa* [30,36,50]. *Huang* (1998) je predložio dve metode za selekciju početnih modusa za ovaj algoritam i pokazao da korišćenje različitih početnih modusa dovodi do boljih rezultata klasterovanja [127]. Eksperimenti *Sun* i saradnika [237] su pokazali da korišćenje prečišćenih početnih tačaka u ovom algoritmu dovodi do mnogo pouzdanijih rezultata nego metoda slučajnog izbora, bez prečišćavanja.

3.3.1.1 Modifikovani algoritam k-modusa

U literaturi, problem odabira važnih obeležja je rešen pomoću tehnika izbora promenljivih, što predstavlja pretprocesirajući korak u postupku klasterovanja. Međutim, svaki klaster može imati različite skupove važnih obeležja i svaki klaster može sadržati neka nevažna obeležja. *He* i saradnici [119] i *San* i saradnici [221] su nezavisno jedan od drugog, predložili optimizacioni algoritam klasterovanja sa težinskim obeležjima koji uopštava algoritam k*-modusa* definisanjem mere različitosti za funkciju cilja. Glavna ideja je modifikacija algoritma k*-modusa* dodeljivanjem težine za svako obeležje u svakom klasteru.

Kao i kod standardnog algoritma k*-modusa*, cilj klasterovanja objekata u k klastera je pronalaženje Z i W tako da se minimizira funkcija

$$F(W,Z)=\sum_{l=1}^{k}\sum_{i=1}^{m} w_{li} d_1(z_l, x_i) \tag{3.8}$$

sa istim uslovima kao u (3.5), (3.6) i (3.7).

Mera različitosti $d_1(z_l, x_i)$ je definisana na sledeći način:

$$d_1(z_l, x_i)=\sum_{j=1}^{n} ¤(z_{l,j}, x_{i,j}) \tag{3.9}$$

gde je

$$¤(z_{l,j}, x_{i,j})=\begin{cases} 1 & , \quad z_{l,j} \neq x_{i,j} \\ 1-\dfrac{|c_{l,j,r}|}{|c_l|} & , \quad z_{l,j} = x_{i,j} \end{cases}$$

gde je $|c_l|$ broj objekata u l -tom klasteru dat sa

$$|c_l| = \left|\{ i \,|\, w_{li}=1 \}\right|,$$

i $|c_{l,j,r}|$ je broj objekata sa kategorijom $d_j^{(r)}$ j -tog atributa u l -tom klasteru, data sa $|c_{l,j,r}| = \left|\{ w_{ls} \,|\, z_{l,j} = x_{s,j} = d_j^{(r)}, w_{ls}=1 \}\right|$.

Teorema 3.4 [188] *Neka je* X *skup kategorijalnih objekata opisanih pomoću kategorijalnih obeležja* $D_1, D_2, \ldots, D_n$ *i* $DOM(D_j) = \left\{ d_j^{(1)}, d_j^{(2)}, \ldots, d_j^{(n_j)} \right\}$, *gde je* n_j *broj kategorija obeležja* D_j, $1 \le j \le n$. *Neka su centri klastera* z_l *predstavljeni sa* $\left[z_{l,1}, z_{l,2}, \ldots, z_{l,n} \right]$ *za* $1 \le l \le k$. *Tada* $\sum_{l=1}^{k} \sum_{i=1}^{m} w_{li} d_1(z_l, x_i)$ *dostiže minimum ako i samo ako je*

$$z_{l,j} = d_j^{(r)} \in DOM(D_j), \text{ gde je}$$

$$\left| \left\{ w_{li} \middle| x_{i,j} = a_j^{(r)}, w_{li} = 1 \right\} \right| \ge \left| \left\{ w_{li} \middle| x_{i,j} = a_j^{(t)}, w_{li} = 1 \right\} \right|, \quad 1 \le t \le n_j \quad za \; 1 \le j \le n \; .$$

Teorema 3.5 [188] *Algoritam* k*-modusa sa merom različitosti definisanom u* (*3.9*) *konvergira u konačnom broju iteracija.*

Eksperimentalni rezultati su potvrdili veću efikasnost ovako modifikovanog algoritma k*-modusa* u klasterovanju kategorijalnih podataka u odnosu na standardni algoritam k*-modusa*. U literaturi postoje različite modifikacije algoritma k*-modusa* [31,119,221], u koje spadaju i *fuzzy algoritam* k*-modusa* [124], GKMODE algoritam [98], gde je korišćen genetski algoritam integrisan sa algoritmom k*-modusa*.

3.3.2 Dvostepeni klaster algoritam

Dvostepeni klaster algoritam (Twostep, dalje u tekstu *TSCA*) se primenjuje u radu sa neprekidnim, kategorijalnim promenljivama (sa nominalnom skalom merenja), odnosno kombinovanim tipovima promenljivih [59]. Dvostepeni klaster algoritam se sastoji iz sledećih koraka:

1. pre-klasterovanje
2. rešavanje atipičnih vrednosti (autlajeri)-opcionalno
3. klasterovanje

Korak preklasterovanja grupiše subjekte u nekoliko malih klastera koji se dalje koriste kao ulazni parametri za grupisanje u veće klastere. Postupak je zasnovan na dobro definisanoj statistici i može automatski da odredi optimalan broj klastera za date ulazne promenljive.

Mera rastojanja je potrebna u oba koraka, u koraku preklasterovanja i u koraku klasterovanja, a dve mere su na raspolaganju: prirodni logaritam funkcije verodostojnosti i Euklidovo rastojanje. Za rad sa kombinovanim tipovima promenljivih kao mera rastojanja koristi se logaritam verodostojnosti. Označimo sa $n^{(1)}$ ukupan broj neprekidnih promenljivih, $n^{(2)}$ ukupan broj kategorijalnih promenljivih ($n^{(1)} + n^{(2)} = n$), a sa n_i broj kategorija i-te kategorijalne promenljive ($i = 1,\ldots,n^{(2)}$). Svaki objekat $x_i \in \mathbb{R}^n$ iz skupa $X = \{x_1, x_2, \ldots, x_m\}$ možemo prikazati kao $\left[x_{i,1}^{(1)}, x_{i,2}^{(1)}, ..., x_{i,n^{(1)}}^{(1)}, x_{i,1}^{(2)}, x_{i,2}^{(2)}, ..., x_{i,n^{(2)}}^{(2)} \right]$, gde su $x_{i,l}^{(1)}$, $l = 1,2,...,n^{(1)}$ neprekidne promenljive, a $x_{i,j}^{(2)}$, $j = 1,2,...,n^{(2)}$ kategorijalne promenljive.

Model pretpostavlja da neprekidne promenljive unutar i-tog klastera imaju normalnu raspodelu $N(«_{ij}, \pm_{ij}^2)$ sa sredinom $«_{ij}$ i varijansom $\pm_{ij}^2$, a da kategorijalne promenljive unutar klastera i imaju multinomnu raspodelu sa verovatnoćama $®_{ijl}$, gde je (jl) indeks za $l-$tu kategoriju ($l = 1,\ldots,n_j$) j-te promenljive. Važi da je $®_{ijl} = m_{ijl} / m_i$, gde je m_i ukupan broj objekata u i-tom klasteru, a m_{ijl} broj objekata u $i-$tom klasteru čija $j-$ta kategorijalna promenljiva ima $l-$tu kategoriju. Takođe se podrazumeva da su promenljive međusobno nezavisne [59].

Kao meru rastojanja, analiziramo logaritam verodostojnosti. Ova mera rastojanja je izvedena iz modela verovatnoće da je rastojanje između dva klastera ekvivalentno smanjenju verovatnoće u logaritmu funkcije verodostojnosti koje nastaje kao rezultat spajanja klastera. Rastojanje između klastera A^i i A^s možemo definisati na sledeći način:

$$d(A^i, A^s) = \P_{\langle i,s \rangle} - \left(\P_i + \P_s\right), \tag{3.10}$$

gde je

$$\P_i = m_i \left(\sum_{j=1}^{n^{(1)}} \frac{1}{2} \log\left(\pounds_{ij}^2 + \pounds_j^2\right) + \sum_{j=1}^{n^{(2)}} H_{ij} \right), \tag{3.11}$$

$$\P_s = m_s \left(\sum_{j=1}^{n^{(1)}} \frac{1}{2} \log\left(\pounds_{sj}^2 + \pounds_j^2\right) + \sum_{j=1}^{n^{(2)}} H_{sj} \right), \tag{3.12}$$

$$\P_{\langle i,s \rangle} = m_{\langle i,s \rangle} \left(\sum_{j=1}^{n^{(1)}} \frac{1}{2} \log\left(\pounds_{\langle i,s \rangle j}^2 + \pounds_j^2\right) + \sum_{j=1}^{n^{(2)}} H_{\langle i,s \rangle j} \right). \tag{3.13}$$

$\P_i$ u (3.11) predstavlja oblik ocenjene varijanse unutar i-tog klastera. Prvi deo izraza, $m_i \sum_{j=1}^{n^{(1)}} \frac{1}{2} \log\left(\pounds_{ij}^2 + \pounds_j^2\right)$ predstavlja varijansu neprekidne promenljive x_j unutar i-tog klastera. $\pounds_j^2$ je dodato u (3.11) je dodat da bi se izbegao slučaj kada je $\pounds_{ij}^2 = 0$. Unutrašnja suma u drugom delu izraza (3.11), $H_{ij} = -\sum_{l=1}^{n_j} \®_{ijl} \log(\®_{ijl})$ predstavlja entropiju $j-$te kategorijalne promenljive u $i-$tom klasteru, to jest meru disperzije za kategorijalne varijable.

Klasteri sa najmanjim rastojanjem $d(A^i, A^s)$ se spajaju u svakom koraku, što je slično postupku korišćenom u aglomerativnim hijerarhijskim algoritmima, a logaritam verodostojnosti za korak sa k klastera se računa kao

$$L(k) = \sum_{t=1}^{k} ,_t .$$

Ovu funkciju $L(k)$ možemo smatrati merom varijanse unutar klastera. Ukoliko imamo samo kategorijalne promenljive, $L(k)$ predstavlja entropiju unutar k klastera.

Broj klastera u TSCA se određuje tokom dve faze ocenjivanja. U prvoj fazi se računa *Akaike informacioni kriterijum* $\left(AIC_k\right)$ ili *Bajesov informacioni kriterijum* $\left(BIC_k\right)$:

$$AIC_k = 2r_k - 2L(k),$$

$$BIC_k = r_k \log m - 2L(k),$$

gde je sa r_k označen broj nezavisnih parametara u modelu:

$$r_k = k\left(2n^{(1)} + \sum_{j=1}^{n^{(2)}}\left(n_j - 1\right)\right).$$

Oba kriterijuma $\left(AIC_k \text{ i } BIC_k\right)$ daju dobre ocene za maksimalan broj klastera [59]. Promena vrednosti *BIC* je razlika vrednosti *BIC* kriterijuma između datog modela sa $k+1$ klastera i sledećeg manjeg sa k klastera:

$$dBIC_k = BIC_{k+1} - BIC_k$$

Maksimalan broj klastera je broj klastera za koje važi da je količnik *BIC* promene za tekući klaster i promene *BIC* za prelazak sa 2 na 1 klaster manji od konstante c_1 (najčešće se koristi vrednost $c_1 = 0.04$, [18]):

$$d_k = d\,BIC_k / dBIC_1 < c_1.$$

U drugoj fazi koristi se količnik mere rastojanja za model sa k klastera:

$$R(k) = \frac{d_{k-1}}{d_k},$$

gde je d_k rastojanje ukoliko je $k+1$ klastera spojeno u k klastera, a na isti način se definiše rastojanje d_{k-1}.

U zavisnosti od toga da li se koristi kriterijum AIC_k ili BIC_k dobijaju se različita rešenja za optimalan broj klastera. Dobijeni broj klastera je ono rešenje gde imamo veliki skok vrednosti količnika $R(k_1)/R(k_2)$. Ovaj količnik se računa kao $R(k_1)/R(k_2)$ za dve najveće vrednosti $R(k)$ ($k = 1, 2, ..., k_{\max}$; $k_{\max}$ je dobijeno iz prvog koraka). Ukoliko je količnik promene veći od unapred definisane vrednosti c_2 ($c_2 = 1.15$, [18]), broj klastera je jednak k_1, inače je broj klastera jednak rešenju sa maksimalnom vrednosti $\max(k_1, k_2)$.

3.4 OCENA VALIDNOSTI REZULTATA KLASTER ANALIZE

U opštem smislu, algoritmi klasterovanja definišu podelu skupa podataka zasnovanu na određenim pretpostavkama, pri čemu ovo ne mora biti bezuslovno „najbolja“ podela koja fituje skup podataka. Kako se na osnovu algoritama klasterovanja izdvajaju klasteri koji nisu poznati „a priori“, konačna podela skupa podataka u većini aplikacija zahteva neki vid procene. Na primer, pitanja kao što su: „koliko ima klastera u datom skupu podataka?“, „da li rezultujuća šema klasterovanja fituje naš skup podatka?“, „da li postoji bolja podela za naše podatke?“ imaju za cilj kvantitativnu procenu rezultata algoritama klasterovanja i poznati su pod zajedničkim imenom *metode klaster validacije***.** Na osnovu jedne od definicija, pod validacijom klasterovanja se podrazumeva postupak ocenjivanja koliko se dobro podela datog skupa slaže sa osnovnom strukturom podataka [112].

Finalni klasteri zahtevaju postupak procene koji uključuje rešavanje niza problema, koji se mogu definisati i na sledeći način:

- određivanje optimalnog broja klastera
- ispitivanje kvaliteta klastera
- procena da li se rezultujuća podela dobro slaže sa osnovnom strukturom podataka [113]

U realnim životnim situacijama sa kojima se sreće istraživač, najvažnija odluka u primeni klaster analize je izbor odgovarajuće metode klasterovanja i određivanje optimalnog broja klastera, jer uspeh dalje analize veoma zavisi od ove odluke. Danas su u literaturi prisutni različiti postupci (mere) za određivanje optimalnog broja klastera [52,147,148,176,178,236].

Evaluacija (procenjivanje) da li je određeno klasterovanje dobro, predstavlja težak zadatak, a još je *Bonner* (1964. godina) istakao da ne postoji univerzalna definicija „dobrog“ klasterovanja [43]. Mere (kriterijumi) validnosti su obično podeljeni u tri kategorije**:**

- *Mere eksterne validnosti* se koriste za procenu stepena slaganja između dve podele (U i V), gde je podela U rezultat postupka klasterovanja, a podela V je formirana na osnovu a priori informacije, nezavisno od particije U (kao što je klasifikacija). *Halkidi* i saradnici su dali pregled nekih od ovih mera [112]. U ovu grupu mera spadaju *tačnost*, *preciznost*, *odziv*, *entropija.* Glavni nedostatak eksternih mera je da se ne mogu uvek primenjivati, jer u realnom skupu podataka a priori informacije nisu uvek poznate.
- *Mere interne validnosti* koriste informacije dobijene unutar postupka klaster analize i ne zahtevaju dodatne informacije o podacima. Interni

kriterijumi mere homogenost unutar klastera, razdvojenost između klastera ili njihovu kombinaciju i predstavljaju slaganje, to jest fitovanje (*goodness-of-fit*) ulaznih podataka i rezultata grupisanja podataka putem klaster analize.

- *Mere relativne validnosti* [246], kod kojih se vrši poređenje particija dobijenih primenom istog algortima, ali korišćenjem različitih parametara, ili različitih podskupova podataka. Ove mere takođe ne zahtevaju dodatne informacije o podacima.

Danas postoji veliki broj različitih tehnika validacije klasterovanja, a koji uključuje: slaganje sa postojećom klasifikacijom, ponovljivost, slaganje sa ekspertskom intuicijom, slaganje sa različitim multivarijantnim metodama, testovi značajnosti, Monte Karlo metode, kontrola interne konzistentnosti i drugo [75].

3.4.1 Mere interne validnosti

Milligan i *Cooper* su izvršili veoma sveobuhvatnu uporednu analizu 30 različitih mera za određivanje optimalnog broja klastera [175]. Rezultati njihove studije su zasnovani na malim skupovima podataka (veličine oko 50). Koristili su 108 sintetičkih baza podataka sa različitim brojem nepreklapajućih klastera (2, 3, 4 ili 5) i različitim brojem promenljivih (4, 6 i 8). Pri tome su kao najznačajnije izdvojene sledeće mere validnosti: *Calinski - Harabasz* [49] *i Gamma* [33]. Od svih analiziranih metoda, globalni metod koji su predložili *Calinski-Harabasz* je pokazao najbolje rezultate u odnosu na druge analizirane mere. Danas u literaturi postoje brojni radovi koji se bave analiziranjem različitih mera interne validnosti klastera, kao i modifikacijom postojećih mera [75,15,109,219], a mi ćemo navesti samo neke od njih.

Calinski-Harabasz [49], poznat i kao ***Fisher-wise kriterijum,*** se računa kao:

$$CH(k) = \frac{\mathrm{tr}(B)/(k-1)}{\mathrm{tr}(W)/(m-k)},$$

gde je k broj klastera, m obim datog skupa podataka, sa *tr* je označen trag matrice, a B i W su matrice disperzije između klastera, odnosno unutar klastera. Trag matrica B i W se računa kao

$$\mathrm{tr}\left(B\right) = \sum_{i=1}^{k} \left|A^i\right| \left(z_i - z\right)^T \left(z_i - z\right),$$

$$\mathrm{tr}(W)=\sum_{i=1}^{k}\sum_{x\in A^i}(x-z_i)^T(x-z_i),$$

gde je z aritmetička sredina (centroid) celog skupa, a z_i centroid klastera A^i. Maksimalna vrednost ovog indeksa se koristi za izbor najbolje particije.

Gamma indeks [33] se računa kao:

$$G=\frac{S_+-S_-}{S_++S_-}\in[-1,1]$$

S_+ predstavlja broj konkordantnih parova objekata, a S_- predstavlja broj diskordantnih parova objekata:

$$S_+=\frac{1}{2}\sum_{l=1}^{k}\sum_{\substack{x_i,x_j\in A^l\\ x_i\neq x_j}}\frac{1}{2}\sum_{m=1}^{k}\sum_{\substack{x_p\in A^m\\ x_q\notin A^m}}¢\left(\|x_i-x_j\|<\|x_p-x_q\|\right),$$

$$S_-=\frac{1}{2}\sum_{l=1}^{k}\sum_{\substack{x_i,x_j\in A^l\\ x_i\neq x_j}}\frac{1}{2}\sum_{m=1}^{k}\sum_{\substack{x_p\in A^m\\ x_q\notin A^m}}¢\left(\|x_i-x_j\|>\|x_p-x_q\|\right).$$

Par rastojanja (različitosti) je konkordantan (diskonkordantan) ukoliko je rastojanje unutar klastera striktno manje (striktno veće) nego rastojanje između klastera. Važi da je $¢(\cdot)=1$, ukoliko je zadovoljena nejednakost u zagradi, odnosno $¢(\cdot)=0$, ukoliko ne važi nejednakost. Bolja particija se očekuje za veće vrednosti S_+, manje vrednosti S_-, to jest veće vrednosti indeksa G.

Silhouette indeks [214]. *Kaufman* i *Rousseeuw* su predložili ovaj indeks za ocenjivanje optimalnog broja klastera u podacima. Označimo sa $X=\{x_1,x_2,\ldots,x_m\}$ skup od m objekata grupisanih u k klastera $A^1,\ldots,A^k$. Neka je $A^j=\left\{x_1^j,x_2^j,\ldots,x_{m_j}^j\right\}$ j-ti klaster, $j=1,\ldots,k$, gde je $|A^j|=m_j$. Označimo sa $d(x_i^j,x_s^j)$ rastojanje između i-tog objekta iz klastera A^j i s-tog objekata u istom klasteru. Definišemo prvo prosečno rastojanje a_i^j između i-tog objekta iz klastera A^j i svih drugih objekata u istom klasteru:

$$a_i^j=\frac{1}{m_j-1}\sum_{s=1}^{m_j}d(x_i^j,x_s^j),\qquad i=1,\ldots,m_j.$$

Minimalno prosečno rastojanje između i-tog objekta u klasteru A^j i svih drugih objekata u klasteru A^s, $s=1,\ldots,k$, $s \neq j$ je definisano na sledeći način:

$$b_i^j = \min_{\substack{l=1,\ldots,k \\ l \neq j}} \left\{ \frac{1}{m_l} \sum_{s=1}^{m_l} d(x_i^j, x_s^l) \right\}, \qquad i=1,\ldots,m_j.$$

Silhouette širina $i-$tog objekta, koji pripada j-tom klasteru A^j, se računa kao:

$$s_i^j = \frac{b_i^j - a_i^j}{\max\left\{a_i^j, b_i^j\right\}} \in [-1,1]..$$

Silhouette klastera A^j se definiše kao

$$S_j = \frac{1}{m_j} \sum_{i=1}^{m_j} s_i^j \ .$$

Sillhouette indeks je definisan kao prosečna *sillhouette širina* za sve objekte u datom skupu:

$$S = \frac{1}{k} \sum_{j=1}^{k} S_{j.}$$

Ovaj indeks odražava kompaktnost (gustinu) unutar klastera i razdvojenost između klastera. Za obe mere (Silhouette klastera i globalni Silhouette indeks) važi da se nalaze u intervalu $[-1,1]$. Optimalna vrednost broj klastera k je izabrana tako da maksimizira vrednost S.

Bajesov informacioni kriterijum (definisan u delu 1.2.3) takođe spada u meru interne validnosti klasterovanja. Maksimiziranje logaritma funkcije verodostojnosti je ekvivalentno minimiziranju Bajesovog informacionog kriterijuma. Za dva data modela fitovana za isti skup podataka, model sa manjom vrednosti informacionog kriterijuma se smatra boljim.

3.4.2 Mere eksterne validnosti

Za klaster kažemo da je „čist“ ukoliko svi elementi pripadaju istoj klasi. Za merenje efikasnosti predložene metode klasterovanja, koristićemo **tačnost** (eng. *accuracy*), koja je definisana na sledeći način:

$$r = \frac{1}{m}\sum_{l=1}^{k} \max_i (m_l^i),$$

gde je sa m označen broj elemenata u skupu, sa k broj klastera, a sa m_l^i broj objekata iz $i-$te klase, koji pripada $l-$tom klasteru. Greška klasterovanja e je definisana kao $e = 1 - r$.

Entropija (definisana u delu 1.2.3) takođe spada u mere eksterne validnosti. Neka je $C = \{A^1, \ldots, A^k\}$ skup disjunktnih klastera za posmatrani skup A, to jest važi $A = \bigcup_{j=1}^{k} A^j$. Cilj je minimiziranje takozvane ukupne entropije za skup klastera, to jest **očekivane entropije** [157]:

$$H(C) = \sum_{i=1}^{k} \frac{m_i}{m} H(A^i),$$

gde je $|A^i| = m_i$ kardinalnost i-tog klastera, m je obim skupa A, a $H(A^i)$ je entropija klastera A^i. Ukoliko je broj klastera $k = 1$, dobijemo $H(C) = H(A)$, dok za $k = m$ (svaka tačka u svom sopstvenom klasteru) važi $H(C) = 0$.

3.5 PRIMENA KLASTER ANALIZE

Problemi sa kojima se sreću istraživači u klaster analizi su veliki broj pokazatelja sličnosti (rastojanja), veliki broj metoda, određivanje skupa relevantnih promenljivih, nedostajući podaci, određivanje optimalnog broja klastera, validnost rešenja [240]. Značajan izazov u klaster analizi predstavlja rad sa velikim skupovima podataka i velikim brojem obeležja, posebno rad sa kategorijalnim, odnosno kombinovanim tipovima obeležja. Važno je napomenuti da mnogi od problema povezanih sa klaster analizom, generalno predstavljaju probleme u multivarijantnoj statistici, a to su: izbor odgovarajuće mere, izbor promenljivih, unakrsna validacija i eksterna validnost [201].

Izbor odgovarajuće metode zavisi od više elemenata, kao što su na primer, osetljivost na ekstreme. Neke metode teže da daju klastere jednake veličine, a neke ne. Pojedine metode razvijene i primenjive u okviru određenih naučnih disciplina, u drugima nisu od većeg značaja. Različite metode klasterizacije mogu da dovedu do različitih konačnih rešenja. Neke metode teže ka malim kompaktnim klasterima, a druge ka velikim, razuđenim. Nažalost, ne postoji jednostavno i univerzalno uputstvo za rešavanje ovih problema.

Određivanje skupa relevantnih promenljivih je kao i kod većine multivarijantnih metoda jedna od najvažnijih odluka, jer sama tehnika klaster analize ne razlikuje relevantne od nerelevantnih promenljivih. Što je više promenljivih uključeno u klaster analizu i što su one više međusobno nezavisne, teže je pronaći odgovarajući obrazac za grupisanje jedinica posmatranja. Uključivanje jedne irelevantne promenljive povećava verovatnoću uticaja autlajera, što može značajno da utiče na rezultate. Mora se voditi računa o multikolinearnosti promenljivih.

Najjednostavniji pristup rešavanju problema nedostajućih podataka je korišćenje samo onih objekata koji imaju kompletne vrednosti promenljivih. Naravno, ovo može dovesti do redukovanog broja objekata u analizi pa se koriste i druge metode. Često se koristi Gouerov (*Gower*) koeficijent sličnosti za konstrukciju matrice sličnosti za objekte koji imaju vrednost bar jedne promenljive. Sledeća mogućnost je da se koriste odgovarajući postupci za ocenu nedostajućih vrednosti. Ovakav pristup se uglavnom ne preporučuje u klaster analizi, jer je većina postojećih metoda zasnovana na prosečnim vrednostima promenljive, što bi moglo da dovede do pristrasnosti rezultata. Pristrasnost rezulatata je posledica činjenice da se ovim postupkom svakoj nedostajućoj vrednosti jedne promenljive dodeljuje ista vrednost, pri čemu je jasno da ove vrednosti teže da se klasteruju zajedno.

Većina algoritama klasterovanja je ograničena na rad sa skupovima podataka koji sadrže neprekidna obeležja. Međutim, u realnim životnim situacijama često imamo velike skupove podataka sa kategorijalnim, kao i kombinovanim tipovima obeležja. Ovo predstavlja veliki izazov sa matematičkog stanovišta, u smislu kreiranja novog efikasnog pristupa u klasterovanju ovakvih podataka.

Verovatno najteži problem sa kojim se sreće istraživač u praksi pri primeni tehnika klaster analize je procena stabilnosti i validnosti dobijenih klastera. Ocena validnosti je veoma važan korak u klaster analizi, jer metode klasterovanja proizvode klastere, čak i za prilično homogene skupove podataka. Problem određivanja optimalnog broja klastera se smatra fundamentalnim problemom validnosti klaster analize. Postoje različiti pokazatelji koji treba da ukažu koje rešenje je najbolje, ali nijedan nije univerzalan i opšte prihvaćen. I pored različitih mera/kriterijuma za procenu validnosti uopšte prisutnih u literaturi i dalje se veoma mali broj formalnih metoda primenjuje u praksi, pri radu sa bazama podataka i istraživanju podataka [113]. Prilikom interpretacije rezultata primenjenih tehnika za određivanje optimalnog broja klastera, treba biti veoma obazriv: neke metode daju "tačan" broj klastera, pri čemu su zasnovane na "lošoj" klasifikaciji. Činjenica je da različiti kriterijumi klasterovanja i metode za određivanje najboljeg broja klastera mogu dati različite rezultate kada se primene na isti skup podataka. *Saitta* i saradnici u svom radu [219] detaljno analiziraju prednosti i mane postojećih indeksa za ocenu validnosti klastera. *Hardy* predlaže korišćenje nekoliko tehnika klaster analize i metoda za određivanje optimalnog broja klastera i analiziranje svih rezultata u cilju dobijanja više informacija o klasterima: veličini, obliku, konveksnosti, gustini, razdvajanju i uzeti u obzir da ove informacije određuju najbolju klasifikaciju, te ih interpretirati pažljivo [116]. Neke od alternativnih metoda validacije klasterovanja su indeksi zasnovani na homogenosti i/ili separaciji, poređenje različitih metoda klasterovanja na istom skupu podataka, vizuelna validacija klastera, testovi homogenosti skupa u poređenju sa alternativnim klasterovanjem i korišćenje eksternih informacija. Jedan od načina procene validnosti klasterskih rešenja obuhvata testiranje razlika između klastera na promenljivama korišćenim u postupku klaster analize. Ovaj pristup podrazumeva korišćenje različitih statističkih tehnika, u zavisnosti od vrste i broja obeležja, kao i broja klastera. Međutim nedostatak ovakvog pristupa predstavlja to što objekti nisu svrstani u klastere po slučaju, već na osnovu maksimiziranja rastojanja između klastera po korišćenim promenljivama.

Mada pretpostavke kao što su linearnost, normalnost raspodele i homoskedastičnost nemaju veliki značaj u klaster analizi, postoje i drugi statistički aspekti koji se moraju rešiti: da li uzorački podaci reprezentuju populaciju, da li postoji multikolinearnost, kao i postojanje autlajera [111]. Pouzdanost rezultata klasterovanja zavisi od reprezentativnosti uzorka. Kao rezultat klaster analize dobijaju se klasteri, čak i kada ne postoji bilo kakva

struktura u podacima. Rešenja nisu jedinstvena, a dobijeni klasteri zavise od mnogo elemenata u samom postupku, izbora promenljivih, algoritma klasterovanja. Sa druge strane, ukoliko se "pravilno" koristi, klaster analiza ima potencijal da otkrije strukturu i povezanost koja se ne može utvrditi drugim standardnim metodama.

4. CILJEVI I HIPOTEZE ISTRAŽIVANJA

4.1 CILJEVI ISTRAŽIVANJA

1. Upoređivanje efikasnosti klaster algoritma primenjenog na prostim slučajnim uzorcima i klaster algoritma primenjenog na celom skupu, za kategorijalne i kombinovane tipove podataka

2. Primena numeričkih metoda u kreiranju modifikovanog postupka klaster analize za velike skupove podataka sa kategorijalnim i kombinovanim tipovima obeležja.

3. Utvrditi da li su klasteri definisani u odnosu na bihejvioralne faktore rizika (pušenje, zloupotreba alkohola, nepravilna ishrana i nedovoljna fizička aktivnost) kod odraslog stanovništva Srbije karakteristični i u odnosu na sociodemografske karakteristike.

4.2 HIPOTEZE ISTRAŽIVANJA

1. Modifikovani postupak klasterovanja je efikasan u radu sa velikim skupovima podataka sa kategorijalnim, odnosno kombinovanim tipovima obeležja.

2. Dobijeni klasteri bihejvioralnih faktora rizika kod odraslog stanovništva Srbije se statistički značajno razlikuju u odnosu na njihove sociodemografske karakteristike (pol, starost, bračni status, materijalno stanje, nivo obrazovanja).

4.3 METODOLOGIJA

Prvi deo metodologije sadrži opis podataka korišćenih u medicinskom delu disertacije (*Istraživanje zdravlja stanovnika Srbije 2006. godine*), kao i delu rezultata za velike skupove podataka sa kombinovanim tipovima promenljivih. Ovaj deo se sastoji iz opisa uzorka, upitnika korišćenih u istraživanju, kao i obeležja korišćenih u klaster analizi. Drugi deo predstavlja opis veštačke baze podataka (baza *Mushrooms*) korišćene u delu rezultata za velike skupove podataka sa kategorijalnim promenljivima.

4.3.1 Istraživanje zdravlja stanovnika Srbije

Istraživanje zdravlja stanovnika Srbije 2006. godine je realizovalo Ministarstvo zdravlja, uz finansijsku i stručnu pomoć Svetske banke, Regionalne kancelarije Svetske zdravstvene organizacije za Evropu, kancelarije za Srbiju, Instituta za javno zdravlje "Dr Milan Jovanović Batut", kao i mreže instituta i zavoda sa teritorije Republike Srbije. Ovakvo istraživanje je prvi put sprovedeno 2000. godine, te je po ovoj metodologiji urađeno ponovljeno istraživanje (eng. *follow up*) 2006. godine.

Ciljna populacija je stanovništvo starosti 20 i više godina (isključene su osobe iz posebnih institucija-starački domovi, socijalne ustanove, zatvori, psihijatrijske institucije). Uzorački okvir čine sva domaćinstva popisana u okviru Popisa stanovništva 2002 godine. Korišćen je stratifikovani dvostepeni uzorak. U Srbiji je identifikovano 6 geografskih oblasti koji predstavljaju glavne stratume u uzorku: Vojvodina, Beograd, centralna, zapadna, istočna i jugoistočna Srbija. Dalja podela stratuma je bila na gradska i ostala područja. Dvoetapno uzorkovanje je sprovedeno tako što su u prvoj etapi izdvojeni popisni krugovi odabrani pomoću „uzorkovanja sa verovatnoćom proporcionalnoj veličini" (eng. *probability proportional sampling)*. U drugoj etapi su odabrana domaćinstva (10 domaćinstava i 3 rezervna sa spiska) pomoću prostog slučajnog uzorka bez ponavljanja. Uzorak odraslog stanovništva Srbije je obuhvatio 14522 odrasle osobe iz 6156 domaćinstava [179]. Nešto više od petine odraslih ispitanika nije analizirano u procesu klasterovanja (22.2%), zbog nedostajućih podataka za neku od ulaznih promenljivih (navike u ishrani, pušenje, upotreba alkohola, nivo fizičke aktivnosti), pa je konačan uzorak koji smo koristili u daljoj analizi obuhvatio 11300 ispitanika.

Istraživanje zdravlja stanovništva Srbije 2006. godine je sprovedeno putem intervjua i merenja telesne visine, telesne mase i arterijskog krvnog pritiska. Upitnici koji su pri tome korišćeni odgovaraju standardima koji se koriste u ovakvim istraživanjima (*WHO Health Survey 2002*, SF-36),

relevantnim iskustvima i preporukama iz sličnih populacionih istraživanja sprovedenih u drugim zemljama (*FINBALT*[3] 2000, 2002, *CINDI*[4] *program*) i specifičnim potrebama naše zemlje. Izvršene su određene izmene i dopune protokola i upitnika iz 2000. godine, koje nisu ugrozile uporedivost sa podacima iz te godine. One su metodološki unapredile istraživanje i obezbedile dobijanje odgovora na standardnizovana pitanja koja se koriste u istraživanjima u Evropskoj Uniji, kao i za dobijanje podataka za indikatore sadržane u bazi podataka Svetske zdravtsvene organizacije „Zdravlje za sve"[5] i indikatora koji su preporučeni za Zdravstvene indikatore Evropske unije (*ECHI*[6]*-2*) radi obezbeđenja uporedivosti naših pokazatelja zdravlja sa indikatorima drugih zemalja [179]. Upitnici su uključili pitanja iz različitih oblasti: sociodemografske karakteristike ispitanika, higijenske navike, navike u ishrani, korišćenje slobodnog vremena, fizička aktivnost, mentalno zdravlje, emocionalno zdravlje, samoprocena zdravlja, pušenje, upotreba alkohola, rizici i znanja o zdravlju, korišćenje zdravstvene službe i drugo (u Prilogu).

4.3.1.1 Klasifikacija bihejvioralnih faktora rizika

Kako je jedan od ciljeva ovog rada primena klaster analize u definisanju populacionih grupa u odnosu na bihejvioralne faktore rizika kod odraslog stanovništva Srbije, kao i utvrđivanje sociodemografskih karakteristika izdvojenih klastera, korišćenjem upitnika su izvedena sledeća dva seta promenljivih: bihejvioralni faktori rizika (tabela 4.1) i sociodemografske karakteristike ispitanika (tabela 4.2).

Klasifikaciju *pušačkog statusa* kod odraslog stanovništva Srbije smo izvršili podelom u tri kategorije, na sledeći način: nepušači (nikada nisu pušili), bivši pušači, pušači (povremeni ili svakodnevni). U rizičnu kategoriju-pušače smo svrstali sve sadašnje pušače, jer postoje indikacije da čak i povremeni pušači imaju obrasce rizičnog ponašanja koji su slični onima kod redovnih pušača [214].

Klasifikaciju odraslog stanovništva Srbije u odnosu na *unos alkohola* smo izvršili na sledeći način: 1. nikada nisu pili alkohol, 2. pio, ali više ne, 3. piju alkohol (nerizično ponašanje) i 4. piju alkohol (rizično ponašanje). U ovu poslednju rizičnu kategoriju spadaju muškarci koji unesu dve ili više alkoholnih jedinica dnevno, odnosno žene koje unesu jednu ili više alkoholnih jedinica dnevno [54] ili prisustvo bar 12 „epizoda pijanstva" (unos 6 ili više alkoholnih jedinica u jednom danu tokom prethodne godine). Pod alkoholnom jedinicom se podrazumeva 8g ili 10ml čistog alkohola (ekvivalent polovini standardne čaše-175 ml crvenog vina) [2].

[3]FINBALT zdravstveni monitoring je kolaborativni sistem za monitoring zdravstvenih navika (pušenje, upotreba alkohola, navike u ishrani i fizička aktivnost) kod odraslog stanovništva Estonije, Finske, Latvije i Litvanije; [4]CINDI-*Countrywide Integrated Noncommunicable Disease Intervention*; [5] Baza podataka za pretraživanje svih uzroka smrti i bolesti, uključujući povrede i nasilje; [6] ECHI-*European Core Health Indicators*

Nepravilna ishrana je bazirana na nedovoljnom unosu voća/povrća, tako da su svi ispitanici koji ne konzumiraju bar jednom dnevno sveže voće ili povrće klasifikovani u grupu sa rizičnim navikama u pogledu načina ishrane.

Nivo fizičke aktivnosti u slobodno vreme je klasifikovan u četiri kategorije na osnovu upitnika, kojeg su kostruisali Saltin i Grimbi [220] uz minorne modifikacije. Nivo aktivnosti je definisan na sledeći način: 1. *sedentarni tip* (čitanje, gledanje televizije) 2. *lagana fizička aktivnost* (šetnja, biciklizam, ribolov itd) najmanje 4 časa nedeljno 3. *umerena fizička aktivnost* (trčanje, bazen, igranje lopte, težak rad u bašti itd) najmanje 4 časa nedeljno i 4. *naporna fizička aktivnost* koja uključuje redovne vežbe jačeg intenziteta ili sportske treninge nekoliko puta nedeljno. Nivo 1 je definisan kao nedostatak fizičke aktivnosti (*nedovoljna fizička aktivnost*).

Mada su neka istraživanja zdravstveno-rizičnih navika kod odraslog stanovništva uključila i gojaznost (ili BMI) kao bihejvioralni faktor rizika, ovo obeležje nismo uključili u klaster analizu, jer je gojaznost posledica drugih navika (nedovoljna fizička aktivnost, nepravilna ishrana i drugo), a ne zdravstvena navika. U analiziranju četiri navedena bihejvioralna faktora rizika koristili smo dihotomne (binarne) varijable, koje imaju dve kategorije: 0 = ne postoji rizik, 1 = prisutan rizik. U tabeli 4.1 su definisane rizične kategorije (prisustvo rizika) za svaki od bihejvioralnih faktora rizika.

Tabela 4.1 . Indikatori zdravstveno-rizičnog ponašanja

Bihejvioralni faktori rizika	Indikatori zdravstveno-rizičnog ponašanja
Pušenje	-pušenje (povremeno ili redovno)
Štetna upotreba alkohola	-dve ili više alkoholnih jedinica dnevno (muškarci), jedna ili više alkoholnih jedinica dnevno (žene) ili prisustvo bar 12 „epizoda pijanstva“[1] tokom prethodne godine
Nepravilna ishrana	-neredovna upotreba svežeg voća, ili povrća u ishrani
Nedovoljna fizička aktivost	-slobodno vreme podrazumeva sedentarni tip aktivnosti (čitanje, gledanje televizije)

[1] konzumiranje 6 ili više alkoholnih jedinica u jednom danu tokom prethodne godine

4.3.1.2 Sociodemografske karakteristike stanovništva

Sociodemografske karakteristike odraslog stanovništva Srbije korišćene u daljoj analizi, prikazane su u tabeli 4.2. Konstrukcija Indeksa blagostanja-*DHS Wealth Index (Demographic and Health Survey)* se sastojala iz nekoliko koraka: određivanja promenljivih indikatora, dihotomizacije, izračunavanja pondera indikatora i vrednosti indeksa i izračunavanja prosečnih tačaka intervala. Promenljive uključene u računanje indeksa blagostanja se odnose

na posedovanje različitih trajnih dobara: broj spavaćih soba po domaćinstvu, materijal od koga je napravljen pod, krov i zidovi stambenog prostora, vrsta vodosnabdevanja i sanitarija; vrsta goriva koja se koristi za grejanje; posedovanje televizora u boji, mobilnog telefona, frižidera, mašine za pranje veša, mašine za pranje sudova, kompjutera, klima uređaja, centralnog grejanja i automobila. Detaljan opis konstrukcije indeksa blagostanja i kvintila prikazan je u [179].

Tabela 4.2. Sociodemografske karakteristike odraslog stanovništva Srbije

Varijabla	Kategorije varijable	Kategorije transformisane varijable
Pol	1 = ženski 2= muški	
Tip naselja	1 = grad 2 = selo	
Region	1=Beograd 2=Vojvodina 3=Centralna Srbija	
Starost	1=20-34 2=35-44 3=45-54 4=55-64 5 =65 i više godina	1=20-34 2=35-54 3 =55 i više godina
Nivo obrazovanja (završena škola)	1= bez škole, nepotpuna osnovna škola 2= osnovna škola 3= srednja škola 4 = viša škola, fakultet	1= osnovna škola i niže 2= srednja škola 3 = viša škola, fakultet
Bračno stanje	1= neoženjen/neudata 2=oženjen/udata 3 = vanbračna zajednica 4=razveden/a 5=udovac/a	0 = žive sami (neoženjen, neudata; razveden/a; udovac/a) 1 = u braku (oženjen/udata; vanbračna zajednica)
Materijalno stanje	1=najsiromašniji 2=drugi 3=srednji 4=četvrti 5= peti	1= lošije (najsiromašniji+drugi) 2 =srednje/bolje (srednji+četvrti+peti)

4.3.2 Baza *Mushrooms*

U analiziranju klasterovanja velikih skupova podataka sa kategorijalnim obeležjima, koristićemo veštačku bazu podataka *Mushrooms*, dostupnu na UCI *Machine Learning Respository* [244]. Ova baza podataka uključuje opis hipotetičkih uzoraka koji obuhvataju 23 vrste gljiva iz familije *Agaricus* i *Lepiota*. Dat je opis gljiva u smislu fizičkih karakteristika i klasifikacije. Svaka vrsta je identifikovana kao jestiva, ili otrovna. Baza podataka sadrži 8124 objekta i 22 promenljive, od čega je 18 nominalnih (više od dve kategorije) i 4 binarne promenljive. Dalje smo izvršili transformaciju nominalnih promenljivih, tako da je svaka promenljiva sa l kategorija transformisana u l binarnih promenljivih, te konačna baza sadrži 125 binarnih promenljivih. Ulazna matrica podataka sadrži 8124×125 podataka, pa bazu možemo smatrati velikim skupom podataka. Ne postoji univerzalno najbolji algoritam klasterovanja za kategorijalne podatke, u smislu kvaliteta rešenja i vremenske složenosti. Stvarna klasifikacija na celom skupu podataka (2 klase gljiva: jestive/otrovne), omogućava da utvrdimo koji je od primenjenih algoritama najbolji na celom skupu, pri čemu će se koristiti kriterijum eksterne validnosti-tačnost (r).

4.4 STATISTIČKE METODE

U numeričkom delu rezultata klasterovanja (poglavlje 5), analizirana su klaster rešenja za velike skupove podataka sa kategorijalnim i kombinovanim tipovima promenljivih. Klaster algoritmi za velike baze sa kategorijalnim promenljivim su analizirani korišćenjem baze *Mushroom*, gde je unapred poznata klasifikacija na celom skupu. Primenjeni su različiti algoritmi klasterovanja: *Dvostepeni klaster algoritam, algoritam k-sredina*, *algoritam k-modusa*, različiti hijerarhijski algoritmi klasterovanja: *Prosečno povezivanje između grupa*, *Prosečno povezivanje unutar grupa*, *Jednostruko povezivanje*, *Potpuno povezivanje i Ward*-ov metod na celom skupu podataka, a zatim su upoređeni dobijeni rezultati klaster algoritama na celom skupu sa već postojećom klasifikacijom. Kriterijum za određivanje najboljeg klasterskog rešenja na celom skupu je bila ekstremna validnost, to jest najveća tačnost (poklapanje sa postojećom klasifikacijom).

Klaster analiza za kombinove tipove podataka (numeričke i kategorijalne) je vršena primenom algoritma TSCA, korišćenjem podataka iz baze *Istraživanja zdravlja Srbije 2006*. Pretpostavke za primenu ovog algoritma su nezavisnost promenljivih uključenih u analizu i multinonma normalna raspodela za neprekidne numeričke promenljive. Kao mera rastojanja između klastera korišćen je logaritam verodostojnosti. TSCA algoritam je izvršen kroz dve faze: tokom prve faze („preklasterovanje“) skup je podeljen u nekoliko manjih podklastera, a zatim su dobijeni podklasteri grupisani u određen optimalan broj klastera. Određivanje optimalnog broja klastera je izvršeno primenom *Bajesovog informacionog kriterijuma* ($BIC(k)$), baziranog na kombinaciji niže vrednosti $BIC(k)$ (koja ne mora biti najniža) i visoke vrednosti količnika promene u rastojanju između k klastera. Kao mera interne validnosti dobijenog klasterskog rešenja, korišćen je *Silhouette* indeks, pri čemu vrednosti indeksa između 0.5 i 1 [214] govore u prilog kvaliteta klasterskog rešenja.

Dalja analiza velikih skupova podataka je uključila korišćenje prostih slučajnih uzoraka (veličine približno $0.01m$, $0.03m$, $0.05m$, $0.1m$, $0.3m$, gde je m obim polaznog skupa), pri čemu su dobijeni rezultati klasterovanja na celom skupu upoređeni sa rezultatima klasterovanja dobijenim na uzorcima. Primenjeno je ukupno 1750 algoritama (na svakom od 250 uzoraka primenjeno 7 algoritama, za različit broj klastera k) iz baze *Istraživanja zdravlja*, odnosno 140 algoritama (2 algoritma za svaki od 70 uzoraka) iz baze *Mushrooms*. Kao mera eksterne validnosti rezultata klasterovanja na uzorcima (iz baze *Mushrooms*) korišćena je *tačnost* (r), to jest slaganje rezultata klasterovanja sa postojećom klasifikacijom na celom skupu. U deskriptivnom delu rezultata prikazane su srednje vrednosti (prosečna vrednost, medijana) i mere varijabiliteta (opseg vrednosti, standardna devijacija SD). U cilju utvrđivanja optimalne veličine uzorka za bazu

Mushrooms, upoređena je tačnost klasterovanja na uzorcima i celom skupu, primenom *Student*-ovog *t*-testa za razlike između aritmetičkih sredina u uzorku i osnovnom skupu. U cilju utvrđivanja optimalnog broja klastera za uzorke iz baze *Istraživanje zdravlja stanovnika Srbije 2006. godine*, primenjena je *Jednofaktorska analiza varijanse sa ponovljenim merenjima* za svaki skup uzoraka, pri čemu su testirane razlike između dobijenih vrednosti slaganja klasterovanja u odnosu na različit broj klastera (k), a dalje su izvršena međusobna poređenja rezultata za parove klastera za svaku grupu uzoraka (iste veličine).

Zbog vremenske složenosti prilikom rada sa velikim skupovima podataka, predložen je modifikovan postupak klasterovanja, zasnovan na primeni klaster algoritma na prostim slučajnim uzorcima određene kardinalnosti umesto na celom skupu podataka, odabiru najboljeg klasterskog rešenja (na osnovu kirterijuma validnosti) i dodeljivanju preostalih članova skupa najbližim klasterima (postupak opisan u 5.3).

Dvostepeni klaster algoritam (TSCA) je primenjen na podatke iz *Istraživanja zdravlja stanovništva Srbije* (glava 7), u cilju analize klasterovanja bihejvioralnih faktora rizika, pri čemu su kao ulazne promenljive korišćene kategorijalne promenljive: pušenje cigareta, nivo fizičke aktivnosti, unos voća i povrća i štetna upotreba alkohola. Procena validnosti dobijenog klasterskog rešenja (eksterna validnost) je uključila testiranje razlika između klastera na relevantnim eksternim obeležjima, koja nisu korišćena u postupku klasterovanja. Primenom ovog pristupa, izvršeno je testiranje razlika između dobijenih klastera bihejvioralnih faktora rizika u odnosu na sociodemografske karakteristike, primenom *Pearson*-ovog χ^2 testa. Nakom univarijantne analize, primenjena je multivarijantna analiza, to jest *multinomni logistički regresioni model* sa *stepwise* izborom, a kao nezavisne promenljive su posmatrane sociodemografske karakteristike (pol, starost, bračno stanje, nivo obrazovanja, materijalno stanje). Model ocenjuje verovatnoću da ispitanik pripada određenom klasteru u poređenju sa referentnom grupom (klaster „Bez faktora rizika“). Interpretacija modela je uključila prikaz *odnosa šansi* (*odds ratio*), zajedno sa 95% intervalom poverenja (CI). Odnos šansi (*OR)* je modelovan za članove klastera za svaku sociodemografsku karakteristiku posebno, pri čemu su ostale promenljive smatrane konstantnim.

Statistička analiza podataka je izvršena korišćenjem statističkog programa IBM SPSS Statistics 22.0. Programski kod za algoritam *k-modusa* (verzija *Huang* [127]) je napisan u MATLAB (*MATrix LABoratory*) okruženju za numeričke proračune i programski jezik, pri čemu je za početne moduse izabrano prvih k različitih tačaka skupa. Svi testovi su dvostrani sa nivoom značajnosti $p \leq 0.05$.

5. KLASTEROVANJE VELIKIH SKUPOVA PODATAKA. REZULTATI

Kao što smo već naglasili, veliki izazov u klaster analizi predstavlja rad sa velikim skupovima podataka, a naročito rad sa kategorijalnim i kombinovanim tipovima obeležjima. Prvi deo analize klaster algoritama se odnosi na rad sa kategorijalnim podacima, a drugi na rad sa kombinovanim tipovima podataka.

5.1 KLASTEROVANJE KATEGORIJALNIH PODATAKA

U analiziranju klasterovanja kategorijalnih podataka, koristimo veštačku bazu podataka *Mushrooms* [244]. Baza podataka sadrži 8124 objekta i 22 promenljive, od čega je 18 nominalnih (više od dve kategorije) i 4 binarne promenljive. Dalje smo izvršili transformaciju nominalnih promenljivih, tako da je svaka promenljiva sa l kategorija transformisana u l binarnih promenljivih, te konačna baza sadrži 125 binarnih promenljivih.

5.1.1 Klasterovanje na celom skupu podataka

Ovaj deo analize se sastoji u upoređivanju efikasnosti različitih algoritama za kategorijalne podatke (*k – sredina, k – modusa, Twostep klaster algoritam-TSCA*, različiti hijerahijski algoritmi) poređenjem rezultata klasterovanja na celom skupu i stvarne klasifikacije, za različit broj klastera. Od hijerarhijskih algoritama analizirani su *Prosečno povezivanje između grupa* (eng. *average linkage between group*, *ALBG*), *Prosečno povezivanje unutar grupa* (eng. *average linkage withing group, ALWG*), *Jednostruko povezivanje* (eng. *single linkage*, *SL*), *Potpuno povezivanje* (*complete linkage*, *CL*) i *Ward-ov metod*, pri čemu smo koristili mere sličnosti za binarne promenljive: *Jaccard* koeficijent, *Dice*, *Simple Matching*, *Kulczynski*, *Roger&Tanimoto*, *Rasell & Rao*, *Yuley's Y*, pa smo u skladu sa primenjenim hijerarhijskim algoritmom i merom sličnosti i uveli skraćene oznake (Tabela 5.1).

Tabela 5.1. Primena različitih algoritama klasterovanja za $k = 2$ (baza *Mushrooms*)

Algoritam	**Klasifikacija gljiva**				**Tačnost**
	Jestive (N=4208)		Otrovne (N=3916)		
	tačno	pogrešno	tačno	pogrešno	
k – sredine	4208	0	1296	2620	67.7
ALWG (Jaccard)	2960	1248	1952	1964	60.5
ALWG (Dice, Simple Matching)	3872	336	3292	624	88.2
ALWG (Roger& Tanimoto)	2992	1216	1952	1964	60.9
ALWG (Rasell & Rao, Yuley's Y)	3968	240	3100	816	87.0
ALBG (Jaccard, Dice)	4016	192	0	3916	49.4
SL (Jaccard, Dice)	4016	192	0	3916	49.4
CL (Jaccard, Dice)	976	3232	3100	816	50.2
Ward	4208	0	3024	892	89.0
TSCA (kategorijalne p)	4208	0	3024	892	89.0
k – modusi	2738	1470	1856	2060	56.5

** u zagradi je prikazana korišćena mera sličnosti*

U tabeli 5.1 (poslednja kolona) je prikazana tačnost (*accuracy*, definisana u delu 3.4.12) različitih klaster algoritama, za zadat broj klastera ($k = 2$), gde su svi objekti klasifikovani u jednu od dve kategorije/klase. Minimalna tačnost iznosi 49.4% , za ALBG i SL algoritam, a maksimalna tačnost iznosi 89%, za TSCA *i* Ward-ov algoritam.

Analiziranu bazu čine 23 različite vrste gljiva, pa je možemo posmatrati i kao skup koji se sastoji iz 23 odvojena klastera (klase). Za klaster kažemo da je „čist“ ukoliko svi njegovi elementi pripadaju istoj klasi, tj. jednoj od dve vrste gljiva, što je ilustrovano u tabeli 5.2, za primer 3 klastera („čist“ klaster je klaster 3, f_{ij} broj objekata koji se nalazi u i-tom klasteru i j-toj klasi, $f_{ij} \neq 0,\ i = 1,2,3;\ j = 1,2$).

Tabela 5.2 Primer čistih klastera

	Klasa 1	Klasa 2
Klaster 1	f_{11}	f_{12}
Klaster 2	f_{21}	f_{22}
Klaster 3	0	f_{32}
Ukupno	$\sum_{k=1}^{3} f_{k1}$	$\sum_{k=1}^{3} f_{k2}$

Rezultati tabele 5.3 za različit broj klastera i različite klaster algoritme, ukazuju da samo hijerarhijski SL algoritam i ALBG algoritam daju "čiste" klastere i tačan broj ovih klastera (23). Međutim ovde ne možemo da tvrdimo da li je ova klasifkacija tačna, jer u bazi ne postoji promenljiva koja definiše ove 23 vrste gljiva, a intuitivno pretpostavljamo da "čist" klaster tačno klasifikuje objekte.

Tabela 5.3. Broj „čistih" klastera za različite algoritme u odnosu na broj klastera (baza *Mushrooms*)

Algoritam	**Ukupan broj klastera**							
	2	**4**	**6**	**7**	**12**	**17**	**22**	**23**
k – sredine	1	0	4	5	9	14	21	21
ALWG (Jacc)	0	1	1	2	8	12	18	19
ALWG (Dice)	0	1	2	2	7	12	18	19
ALBG (Jacc, Dice)	0	2	3	5	8	13	21	23
SL (Jacc, Dice)	1	3	4	6	10	14	22	23
CL (Jacc, Dice)	0	2	2	4	9	15	20	21
Ward	1	3	4	5	9	15	20	21
TSCA	1	3	4	5	8	14	20	21
k – modusi	0	0	1	2	3	6	9	12

U tabeli 5.4 je prikazana *tačnost*, u smislu tačnog razdvajanja dve klase gljiva (jestive i otrovne). Jasno je da najbolje rezultate za 23 klastera pokazuju navedena dva algoritma, SL algoritam i ALBG (preciznost 100%).

Tabela 5.4. Tačnost različitih algoritama klasterovanja u odnosu na broj klastera (baza *Mushrooms*)

Algoritam	**Ukupan broj klastera**							
	2	**4**	**6**	**7**	**12**	**17**	**22**	**23**
k – sredine	67.7	89.0	89.0	89.0	95.8	97.9	98.8	98.4
ALWG (Jacc)	60.5	88.2	88.2	90.4	97.3	98.2	99.0	99.0
ALWG (Dice)	88.2	88.2	90.4	90.8	97.6	97.9	99.0	99.0
ALBG (Jacc, Dice)	49.4	68.0	89.3	89.5	89.5	94.4	99.6	100.0
SL (Jacc, Dice)	49.5	68.2	89.4	90.0	89.7	91.2	100.0	100.0
CL (Jacc, Dice)	50.2	51.6	82.7	92.1	97.4	98.7	98.7	99.2
Ward	89.0	89.0	89.0	89.0	96.8	98.8	99.0	99.0
TSCA	89.0	89.0	89.0	89.0	93.4	98.2	99.4	99.4
k – modusi	56.5	73.5	89.0	89.0	91.5	96.5	97.0	93.0

Najlošije rezultate, tj tačnost prilikom razdvajanja dva klastera imaju hijerarhijski algoritmi: SL, CL i ALBG. Najbolji rezultat, tj. preciznost prilikom izdvajanja dva klastera postižu TSCA i *Ward*-ov algoritam (preciznost 89%) (tabela 5.1)

Za razliku od podele na dve klase gde nije bilo izdvajanja „čistih“ klastera (Tabela 5.1), prilikom podele skupa na 23 klastera (jer bazu čine 23 vrste gljiva, Tabela 5.3), utvrđene su dve tehnike koje daju sve čiste klastere. To su algoritmi SL i ALBG sa korišćenjem *Jaccard*-ovog koeficijenta. Međutim, ovde možemo samo da govorimo o „čistim klasterima“, ali ne možemo da utvrdimo slaganje sa stvarnom podelom u 23 različite grupe (jer u bazi ne postoji promenljiva koja definiše ovu klasifikaciju).

Vremenska složenost hijerarhijskih algoritama iznosi $O(m^3)$, gde je m broj slučajeva, što predstavlja problem u slučaju velikih baza podataka (za veliko m). Iz tog razloga koristićemo proste slučajne uzorke, a zatim ćemo primeniti Ward-ov algoritam (koji pokazuje najbolje rezultate na celom skupu) na ovim uzorcima, u cilju utvrđivanja da li je dovoljan rad na samim uzorcima određene kardinalnosti, umesto klaster analize na celom skupu.

5.1.2 Korišćenje prostih slučajnih uzoraka

Na slučajan način je iz baze *Mushroom* odabrano ukupno 50 uzoraka, to jest po 10 uzoraka veličine 100, 250, 400, 800 i 2400. Obim uzorka je određen tako da predstavlja redom, približno 1%, 3%, 5%, 10%, odnosno 30% celog skupa. Na svakom od uzoraka su primenjena dva Ward-ova algoritma: za $k = 2$ i $k = 23$. Tačnost je izračunata korišćenjem stvarne klasifikacije (2 klase) za $k = 2$, osnosno za $k = 23$ analiziranjem “čistih” klastera (stvaran broj vrsta je 23). Rezultati su prikazani u tabeli 5.5.

Tabela 5.5. Tačnost Ward-ovog algoritma za k=2, k=23 u odnosu na različite veličine uzoraka (baza *Mushrooms*)

Broj klastera	**Redni broj uzorka**	**Obim uzorka**					**Ceo skup (N=8124)**
		100 (1.2%)	**250 (3.0%)**	**400 (4.9%)**	**800 (9.8%)**	**2400 (29.5%)**	
k=2	1	73.0	88.0	88.5	88.2	88.3	89.0
	2	87.0	90.0	90.8	88.2	89.9	
	3	93.0	88.8	88.0	88.9	89.4	
	4	68.0	88.0	88.0	88.4	88.8	
	5	89.0	90.4	88.5	89.8	89.0	
	6	82.0	85.6	88.2	89.5	88.5	
	7	81.0	91.6	89.0	88.2	88.7	
	8	78.0	88.8	88.8	89.1	89.2	
	9	89.0	89.2	89.8	89.4	89.1	
	10	88.0	88.0	87.5	91.0	90.0	
$\bar{x}$		**82.8**	**88.8**	**88.7**	**89.1**	**89.1**	
k=23	1	96.0	97.6	99.2	98.6	98.8	99.0
	2	96.0	98.4	99.0	98.8	99.0	
	3	99.0	99.6	98.8	98.5	98.8	
	4	100.0	99.2	99.0	98.8	98.7	
	5	98.0	98.4	99.0	98.5	99.2	
	6	98.0	98.8	99.0	98.9	99.0	
	7	97.0	98.8	98.5	99.1	98.8	
	8	99.0	98.0	99.8	99.1	99.2	
	9	100.0	96.8	99.2	99.2	99.2	
	10	99.0	99.2	98.8	98.6	98.8	
$\bar{x}$		98.2	98.5	99.0	98.8	99.0	

U tabeli 5.6 su prikazani deskriptivni parametri za tačnost: srednje vrednosti (aritmetička sredina, medijana) i mere varijabiliteta (opseg, standardna devijacija SD) .Svaka vrednost u tabeli je izračunata korišćenjem 10 uzoraka definisane veličine i za dati broj klastera (k).

Tabela 5.6 Tačnost rezultata klasterovanja za proste slučajne uzorke. Deskriptivni parametri (baza *Mushrooms*)

Obim uzorka	Deskriptivni parametri	Broj klastera		p[1]
		***k*=2**	***k*=23**	
100	$\bar{x}$	82.8	98.2	0.036
	Med	84.5	98.5	
	Min	68.0	96.0	
	Max	93.0	100.0	
	SD	7.9	1.5	
250	$\bar{x}$	88.8	98.5	0.764
	Med	88.8	98.6	
	Min	85.6	96.8	
	Max	91.6	99.6	
	SD	1.6	0.8	
400	$\bar{x}$	88.7	99.0	0.369
	Med	88.5	99.0	
	Min	87.5	98.5	
	Max	90.8	99.8	
	SD	1.0	0.3	
800	$\bar{x}$	89.1	98.8	0.811
	Med	89.0	98.8	
	Min	88.2	98.5	
	Max	91.0	99.2	
	SD	0.9	0.3	
2400	$\bar{x}$	89.1	99.0	0.623
	Med	89.1	98.9	
	Min	88.3	98.7	
	Max	90.0	99.2	
	SD	0.6	0.2	

[1] nivo značajnosti razlike između vrednosti u uzorku i celom skupu, za *k*=2. Korišćen Student-ov t-test za razliku između aritmetičkih sredina u uzorku i osnovni skupu

Kao što se može videti na osnovu tabele 5.6 (poslednja kolona), najveće odstupanje postignute tačnosti (poklapanje sa 2 klase) u odnosu na tačnost na celom skupu (89.0%) je utvrđeno za uzorke najmanjeg obima ($0.01m$; tačnost 82.8%), gde je utvrđena značajna razlika (p=0.036). Nije utvrđena značajna razlika u rezultatima klasterovanja za uzorke približne veličine $0.03m$, $0.05m$, $0.1m$, $0.3m$ (rang 88.7% do 88.9%) u odnosu na postignutu tačnost na celom skupu podataka.

Na osnovu prethodnih koraka, za korišćene uzorke približne veličine $0.01m, 0.03m, 0.05m, 0.1m, 0.3m$ (m obim osnovnog skupa) gde je izvučeno 10 uzoraka iste veličine, najbolje rezultate slaganja klasterovanja na uzorcima i klasifikacije na celom skupu postignuto je za optimalnu veličinu uzorku $m_u = t_u m, \ t_u \in [0.03, 0.30]$, pri čemu možemo smatrati da je dovoljna veličina uzorka $m_u = t_u m, \ t_u \in [0.03, 0.10]$.

5.2 KLASTEROVANJE KOMBINOVANIH TIPOVA PODATAKA

Kao što smo već naglasili, veliki izazov u klaster analizi predstavlja rad sa velikim skupovima podataka i kombinovanim obeležjima. U ovom delu analize koristimo bazu podataka iz *Istraživanja zdravlja stanovnika Republike Srbije 2006. godine.*

5.2.1 Klasterovanje na celom skupu

Primenjujemo TSCA algoritam za kombinovana obeležja (četiri kategorijalne promenljive, četiri kontinuirane promenljive) na celom skupu podataka, za različit broj klastera $(k = 2,3,\ldots,8)$. Rezultujuću particiju skupa primenom ovih algoritama ćemo redom označiti sa $C^{(2)}, C^{(3)}, ..., C^{(k)}$.

Tabela 5.7 Parametri za određivanje optimalnog broja klastera. TSCA algoritam (baza *Istraživanje zdravlja stanovnika Srbije, 2006. godina*)

broj klastera (k)	BIC_k	$dBIC_k = BIC_{k+1} - BIC_k$	$d_k = dBIC_k / dBIC_1$	$R(k) = d_{k-1} / d_k$
1	77446.910			
2	65268.466	-12178.444	1.000	1.417
3	56703.824	-8564.643	0.703	1.245
4	49849.243	-6854.581	0.563	1.059
5	43385.746	-6463.496	0.531	1.810
6	39864.161	-3521.586	0.289	1.318
7	37219.625	-2644.535	0.217	1.042

Korišćenjem kriterijuma baziranog na najboljoj kombinaciji niske vrednosti za *Bajesov informacioni kriterijum* (BIC_k) u prvoj fazi i visoke vrednosti količnika promene $R(k)$ u drugoj fazi klasterovanja dobijena je optimalna vrednost broja klastera $k = 5$ na celom skupu podataka (Tabela 5.7). Maksimalni količnici promene rastojanja između k klastera, dobijeni su za $k = 5$ i $k = 2$:

$R(5) = d_4 / d_5 = 1.810, \quad R(2) = d_1 / d_2 = 1.417.$

Količnik ove dve vrednosti je veći od unapred zadate konstante c_2:

$R(5) / R(2) = 1.277 > c_2$, pa kao optimalno rešenje dobijamo $k = 5$.

Silhouette indeks, kao mera interne validnosti klaster rešenja iznosi 0.5, što govori u prilog kvaliteta dobijenog klasterskog rešenja.

5.2.2 Klasterovanje na prostim slučajnim uzorcima

1. U cilju ispitivanja stabilnosti dobijenih klastera, to jest ponovljivosti rezultata klaster algoritma dobijenih na celom skupu, primenili smo algoritam na prostim slučajnim uzorcima različite veličine. Na slučajan način je odabrano ukupno 250 uzoraka, to jest po 50 uzoraka iste veličine redom za $m_u = 100,\ 300,\ 500,\ 1000,\ 3000$. Ovi uzorci čine približno 1%, 3%, 5%, 10% i 30% veličine polaznog skupa. Na svakom od ovih uzoraka je primenjeno 7 klaster algoritama, za različit broj klastera $(k = 2,3,\ldots,8)$, to jest ukupno 1750 algoritama. Rezultujuće particije uzoraka ($i = 1,2,\ldots,50$) dobijene primenom ovih algoritma ćemo označiti sa $B_{m_u}^{(k)} = \left\{B_{m_u}^1, B_{m_u}^2, \ldots, B_{m_u}^k\right\}$, gde je $k = 2,3,\ldots,8;$ $m_u = 100,\ 300, 500,\ 1000,\ 3000.$

2. Dalje analiziramo slaganje rezultujućih klastera dobijenih na ovim uzorcima i klastera dobijenih na celom skupu

Primenom klaster algoritma sa k klastera na celom skupu dobijamo particiju označenu sa $A^{(k)} = \left\{A^1, A^2, \ldots, A^k\right\}$, a primenom klaster algoritma sa k klastera na uzorku veličine m_u dobijamo particiju označenu sa $B_{m_u}^{(k)} = \left\{B_{m_u}^1, B_{m_u}^2, \ldots, B_{m_u}^k\right\}$, Ukoliko klaster A^j ima većinu zajedničkih članova sa klasterom $B_{m_u}^l$, tada j-tom klasteru u $A^{(k)}$ odgovara l-ti klaster u $B_{m_u}^{(k)}$. Primenjujući ovo pravilo, označimo sa $C_{m_u}^{(k)} = \left\{C_{m_u}^1, C_{m_u}^2, \ldots, C_{m_u}^k\right\}$, particiju dobijenu iz $B_{m_u}^{(k)}$ sa izvršenom permutacijom redosleda klastera, tako da se optimalizuje ukupno slaganje između inicijalne particije $A^{(k)} = \left\{A^1, A^2, \ldots, A^k\right\}$ i particije dobijene na uzorku:

$$\left|A^1 \cap C_{m_u}^1\right| + \left|A^2 \cap C_{m_u}^2\right| + \ldots + \left|A^k \cap C_{m_u}^k\right|.$$

Rezultati slaganja particije dobijene primenom klaster algoritma na uzorku i algoritma primenjenih na uzorcima (za $k = 2,3,\ldots,8$) su prikazani u sledećim tabelama: tabela 5.8 ($n = 100$), tabela 5.9 ($n = 300$), tabela 5.10 ($n = 500$), tabela 5.11 ($n = 1000$) i tabela 5.12 ($n = 3000$).

Tabela 5.8. Slaganje rezultata particije $A^{(k)}$ i $C_{100}^{(k)}$ (baza *Istraživanje zdravlja stanovnika Srbije, 2006. godina*)

R. BR.	K=2	K=3	K=4	K=5	K=6	K=7	K=8
1	100	100	99	98	98	79	81
2	82	60	77	73	85	80	67
3	56	60	75	81	91	72	78
4	68	54	82	71	82	82	81
5	90	86	92	72	72	69	74
6	60	66	72	80	80	83	73
7	91	58	62	60	72	74	85
8	96	95	95	81	75	79	83
9	82	100	100	100	100	85	85
10	66	73	90	90	60	67	85
11	100	100	71	100	100	100	94
12	84	75	91	99	91	100	90
13	50	58	71	77	90	93	77
14	74	57	64	80	91	87	82
15	72	63	79	89	96	83	96
16	72	72	77	78	92	86	79
17	85	66	82	74	87	69	79
18	71	58	76	83	92	91	86
19	56	55	70	82	91	84	85
20	86	67	83	80	90	99	92
21	97	93	93	87	90	90	85
22	89	82	91	81	87	76	69
23	87	66	84	86	75	67	87
24	82	59	75	81	82	85	75
25	55	55	74	80	86	81	74
26	100	100	100	94	94	76	82
27	69	78	82	78	67	67	67
28	94	55	74	80	70	79	82
29	84	85	85	93	82	96	83
30	73	62	79	74	80	82	78
31	57	59	82	80	90	85	83
32	60	61	82	84	95	90	90
33	87	79	82	89	82	98	96
34	63	63	91	89	98	84	93
35	55	57	85	80	80	82	95
36	72	64	79	88	71	76	92
37	84	69	79	87	84	87	83
38	82	73	91	97	84	97	89
39	80	66	90	78	76	76	87
40	58	61	77	74	73	68	65
41	65	62	70	79	79	72	80
42	100	100	70	100	100	93	67
43	85	72	72	86	70	79	90
44	100	90	90	100	100	82	81
45	83	65	78	70	77	82	82
46	53	52	70	66	75	77	81
47	65	57	85	88	75	75	81
48	100	86	89	96	92	86	79
49	60	58	81	90	97	89	84
50	100	94	94	83	69	79	87
$\overline{x}$	**77.6**	**70.9**	**81.6**	**83.7**	**84.3**	**82.4**	**82.4**

Tabela 5.9. Slaganje rezultata particije $A^{(k)}$ i $C_{300}^{(k)}$ (baza *Istraživanje zdravlja stanovnika Srbije, 2006. godina*)

R. BR.	K=2	K=3	K=4	K=5	K=6	K=7	K=8
1	99.7	100.0	100.0	100.0	100.0	100.0	100.0
2	100.0	92.0	92.0	99.3	98.7	97.0	92.3
3	83.0	75.7	93.0	97.3	97.3	91.7	94.3
4	85.0	85.0	96.3	96.3	96.3	78.0	90.7
5	100.0	100.0	100.0	100.0	100.0	79.0	91.7
6	100.0	89.0	89.0	82.7	94.7	79.3	79.3
7	71.0	83.7	89.3	100.0	100.0	94.3	87.0
8	71.3	62.0	74.7	80.7	91.3	81.3	91.0
9	66.7	80.0	81.3	87.0	99.3	82.7	88.3
10	100.0	99.3	99.7	94.7	85.7	78.0	87.7
11	82.0	59.0	86.7	94.0	74.3	78.0	89.0
12	67.0	73.0	90.7	98.7	98.7	96.3	90.7
13	100.0	92.3	93.0	99.7	99.7	95.3	88.0
14	100.0	100.0	100.0	100.0	100.0	93.3	87.3
15	100.0	91.3	91.3	98.3	100.0	93.3	89.0
16	100.0	100.0	72.7	100.0	100.0	96.7	91.3
17	69.3	69.7	100.0	90.7	91.7	96.7	96.0
18	100.0	91.3	91.3	97.0	96.7	93.3	85.3
19	83.0	72.7	83.3	82.3	93.0	71.0	86.3
20	64.7	57.7	84.7	91.7	100.0	82.0	92.3
21	100.0	95.0	95.0	100.0	100.0	79.3	78.3
22	100.0	99.7	99.7	99.7	73.3	94.7	84.0
23	100.0	93.0	93.0	98.0	100.0	85.0	76.7
24	71.0	54.5	76.7	87.7	76.7	78.0	92.7
25	100.0	100.0	98.0	98.0	98.0	99.7	95.3
26	71.7	60.3	83.0	91.0	100.0	98.7	98.0
27	81.7	81.7	100.0	100.0	76.0	100.0	93.7
28	57.3	81.0	85.0	85.0	96.0	86.3	82.3
29	100.0	92.3	92.3	100.0	100.0	93.7	85.7
30	100.0	99.7	99.7	99.7	100.0	96.7	96.7
31	100.0	99.7	99.7	100.0	100.0	82.3	98.0
32	100.0	99.0	99.0	98.0	98.3	84.0	85.7
33	100.0	90.7	90.7	100.0	100.0	82.3	90.0
34	84.3	78.7	94.3	81.7	85.3	77.3	87.7
35	98.7	99.7	97.0	98.7	98.7	87.3	81.3
36	100.0	90.7	90.7	98.3	98.3	77.7	92.7
37	100.0	61.3	95.7	95.7	85.3	82.7	85.7
38	100.0	92.0	92.0	96.0	95.7	78.0	84.7
39	100.0	100.0	100.0	99.0	99.0	84.7	97.3
40	97.7	100.0	97.7	98.3	98.3	95.3	86.7
41	84.3	84.3	92.0	99.7	74.3	98.3	90.7
42	60.3	70.0	80.3	89.7	99.7	81.3	92.3
43	70.7	70.7	88.0	88.0	100.0	92.3	90.0
44	69.7	89.0	89.3	96.3	97.7	85.0	78.7
45	100.0	89.0	72.3	100.0	77.7	92.0	93.0
46	75.0	57.3	65.3	84.3	92.3	92.3	100.0
47	100.0	100.0	100.0	100.0	100.0	90.7	86.3
48	88.7	60.7	95.0	96.0	96.0	89.3	80.7
49	83.7	59.3	94.3	93.7	93.7	89.0	85.7
50	100.0	57.0	99.7	100.0	80.3	86.3	86.3
$\overline{x}$	**88.8**	**83.6**	**91.3**	**95.3**	**94.2**	**87.9**	**89.1**

Tabela 5.10 Slaganje rezultata particije $A^{(k)}$ i $C_{500}^{(k)}$ (baza *Istraživanje zdravlja stanovnika Srbije, 2006. godina*)

R.BR.	K=2	K=3	K=4	K=5	K=6	K=7	K=8
1	100.0	91.8	91.8	96.2	96.2	89.2	90.8
2	100.0	88.8	73.6	100.0	81.0	77.6	82.8
3	100.0	100.0	100.0	100.0	100.0	79.4	94.6
4	100.0	90.8	91.2	100.0	100.0	81.4	90.8
5	100.0	100.0	73.6	100.0	100.0	81.4	82.8
6	81.0	73.4	92.2	97.0	97.0	83.8	94.6
7	100.0	100.0	100.0	100.0	100.0	84.8	98.6
8	100.0	100.0	100.0	100.0	100.0	85.0	85.4
9	100.0	90.6	90.6	100.0	100.0	81.4	89.2
10	100.0	91.4	91.4	98.4	81.2	98.2	91.0
11	100.0	98.8	98.8	99.8	99.0	82.8	74.2
12	51.6	63.4	75.2	84.0	91.0	77.6	80.0
13	99.8	100.0	100.0	100.0	100.0	82.8	96.2
14	100.0	100.0	100.0	100.0	100.0	84.8	97.6
15	68.4	68.4	88.4	88.6	100.0	99.8	95.0
16	100.0	100.0	100.0	99.8	99.8	85.6	97.4
17	84.6	68.8	85.8	95.8	95.4	85.8	95.6
18	66.8	66.8	89.8	89.8	100.0	82.4	82.2
19	100.0	91.6	91.6	100.0	100.0	100.0	96.8
20	99.8	99.8	99.6	99.6	99.6	79.6	81.8
21	100.0	45.8	91.8	99.8	78.4	81.6	85.6
22	66.6	66.4	93.0	92.8	82.6	83.0	74.0
23	65.0	90.4	90.4	98.0	98.0	86.4	82.2
24	85.6	76.2	89.8	100.0	100.0	82.2	87.0
25	100.0	99.6	99.6	100.0	100.0	100.0	90.8
26	100.0	100.0	100.0	100.0	99.6	79.8	84.8
27	84.6	70.2	78.4	88.8	98.0	94.6	80.4
28	100.0	100.0	100.0	100.0	100.0	83.2	82.4
29	65.4	68.6	88.6	89.8	99.6	82.6	89.6
30	100.0	100.0	74.2	99.6	73.0	77.0	82.6
31	67.0	67.0	91.6	91.4	100.0	80.0	90.8
32	68.0	91.6	91.6	99.6	97.6	89.6	87.6
33	100.0	92.6	92.6	100.0	74.8	77.2	89.0
34	100.0	91.2	91.2	99.0	99.0	81.8	96.6
35	100.0	100.0	98.6	95.2	84.4	91.6	85.8
36	100.0	100.0	73.2	100.0	100.0	100.0	94.0
37	100.0	91.6	91.2	99.6	73.6	77.0	85.8
38	100.0	99.6	99.8	100.0	99.6	80.8	84.6
39	59.2	64.0	72.4	84.6	93.4	76.8	82.4
40	88.4	59.8	82.0	90.4	90.4	96.8	85.6
41	100.0	91.0	91.0	98.8	98.8	98.2	90.4
42	83.0	65.6	82.6	89.8	100.0	99.4	98.6
43	52.2	60.0	77.6	85.4	92.2	92.8	80.4
44	72.8	72.8	65.0	89.8	99.8	79.6	88.2
45	100.0	100.0	71.8	100.0	100.0	81.6	89.8
46	89.0	66.8	84.0	95.0	94.6	81.2	91.8
47	82.8	68.2	88.0	88.2	95.8	83.8	90.6
48	100.0	100.0	100.0	99.0	99.0	80.2	88.6
49	85.8	69.2	87.8	98.6	88.4	89.4	86.0
50	97.8	100.0	74.2	99.8	99.8	85.6	74.8
$\overline{x}$	**89.3**	**85.1**	**88.9**	**96.4**	**95.0**	**85.5**	**82.6**

Tabela 5.11 Slaganje rezultata particije $A^{(k)}$ i $C_{1000}^{(k)}$ (baza *Istraživanje zdravlja stanovnika Srbije, 2006. godina*)

R.BR.	K=2	K=3	K=4	K=5	K=6	K=7	K=8
1	83.4	83.2	99.7	99.7	76.5	76.7	92.5
2	84.0	83.8	88.5	99.5	79.0	80.0	80.7
3	82.6	92.6	97.2	97.2	88.3	81.2	91.4
4	99.9	99.9	99.9	100.0	76.7	79.9	84.4
5	96.4	56.2	96.3	96.4	75.4	81.3	91.0
6	85.6	85.6	89.2	100.0	76.5	80.6	92.3
7	82.0	75.3	86.5	84.2	95.3	88.1	87.3
8	57.0	81.7	96.5	96.5	96.5	84.8	88.5
9	83.4	83.4	95.6	95.6	74.7	84.0	99.6
10	100.0	100.0	100.0	100.0	74.3	81.3	91.0
11	82.5	83.4	95.9	95.9	72.8	79.9	92.4
12	84.9	84.9	96.8	96.8	73.9	82.3	79.9
13	99.8	99.8	100.0	100.0	91.8	82.1	83.3
14	96.0	58.6	96.0	96.0	74.8	82.6	94.9
15	84.3	84.3	96.9	96.9	96.9	83.1	90.8
16	100.0	100.0	99.9	100.0	100.0	99.7	91.3
17	100.0	99.9	99.9	100.0	100.0	100.0	91.3
18	83.9	83.9	96.1	96.1	96.1	83.9	90.6
19	99.8	99.8	99.8	99.8	99.9	100.0	90.9
20	81.6	81.4	99.7	99.7	75.8	79.0	88.0
21	83.9	83.8	99.8	99.8	76.4	79.5	76.4
22	100.0	99.9	99.9	99.9	80.6	83.1	81.6
23	83.7	83.6	96.4	96.5	73.9	81.8	87.7
24	60.7	85.3	85.0	95.6	95.6	91.6	82.1
25	84.2	84.2	96.3	96.3	72.9	90.1	90.5
26	86.7	68.4	79.6	93.0	92.9	80.4	90.2
27	100.0	100.0	99.9	99.9	76.9	82.9	90.2
28	99.8	99.8	99.8	99.8	99.8	99.8	97.4
29	85.1	85.0	95.9	95.9	74.7	77.8	88.2
30	83.5	83.4	86.6	96.6	74.2	81.1	84.1
31	95.7	85.6	86.0	85.9	85.9	96.2	84.5
32	83.9	83.8	99.8	99.8	78.6	82.8	89.7
33	82.9	82.9	96.9	96.8	74.9	82.1	96.3
34	83.8	83.6	96.0	96.1	75.7	100.0	89.0
35	85.7	85.7	96.0	96.0	72.8	81.1	98.4
36	85.8	85.8	95.6	95.7	95.7	80.0	91.2
37	83.8	83.8	96.7	96.7	76.1	90.4	82.6
38	82.6	82.5	99.9	99.9	78.0	81.6	86.8
39	84.9	84.9	96.0	96.0	74.7	81.2	82.6
40	82.7	82.6	96.4	96.4	74.2	79.4	91.8
41	99.8	82.5	96.2	96.3	74.3	100.0	88.3
42	96.1	57.6	96.1	96.1	75.0	79.0	92.9
43	84.0	84.0	95.8	95.8	95.8	82.3	82.3
44	58.9	84.5	95.0	95.0	95.0	82.6	86.7
45	88.1	61.1	79.2	69.7	78.4	100.0	87.1
46	85.2	85.1	95.8	95.8	73.7	78.9	87.7
47	84.7	84.7	95.4	95.4	74.1	78.6	72.3
48	83.4	83.4	94.7	94.7	73.0	98.2	91.5
49	96.4	59.3	96.4	96.4	75.9	82.1	92.3
50	84.2	84.1	96.2	96.2	73.7	96.3	88.5
$\overline{x}$	**86.9**	**84.0**	**95.4**	**96.3**	**81.8**	**85.4**	**88.5**

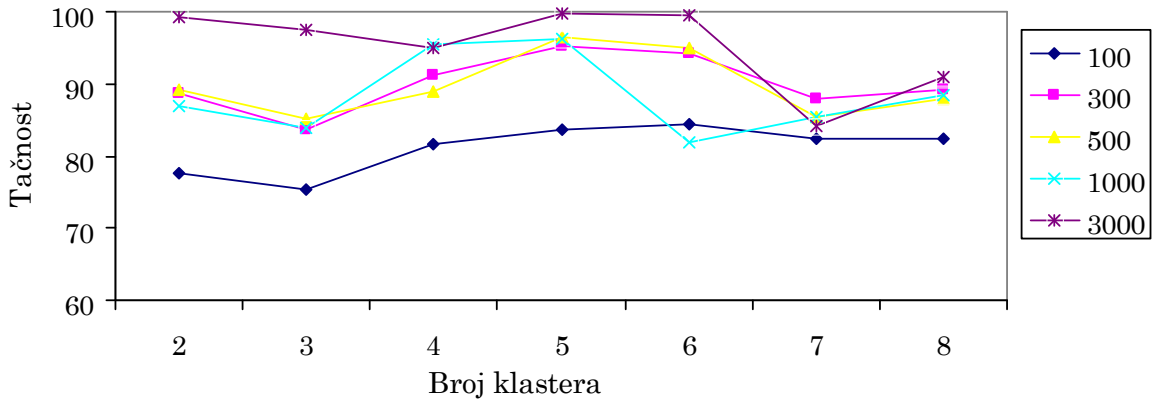
Tabela 5.12 Slaganje rezultata particije $A^{(k)}$ i $C_{3000}^{(k)}$ (baza *Istraživanje zdravlja stanovnika Srbije, 2006. godina*)

R.BR.	K=2	K=3	K=4	K=5	K=6	K=7	K=8
1	100.0	99.9	99.9	100.0	100.0	82.6	90.7
2	100.0	100.0	100.0	100.0	100.0	79.8	96.6
3	100.0	99.7	99.7	99.7	99.8	92.2	91.1
4	100.0	99.9	99.9	99.9	78.0	81.9	98.2
5	100.0	100.0	100.0	100.0	100.0	82.7	98.4
6	100.0	100.0	100.0	100.0	100.0	81.9	90.0
7	100.0	100.0	100.0	100.0	100.0	83.4	98.3
8	100.0	99.9	99.9	99.9	100.0	80.1	96.5
9	100.0	100.0	100.0	100.0	100.0	81.3	90.4
10	100.0	100.0	100.0	100.0	100.0	83.2	98.9
11	100.0	100.0	99.9	99.8	99.8	79.3	81.6
12	100.0	100.0	100.0	100.0	100.0	78.6	88.0
13	100.0	100.0	100.0	100.0	100.0	81.1	97.4
14	100.0	90.2	90.2	100.0	100.0	100.0	91.2
15	100.0	100.0	72.8	100.0	100.0	100.0	89.3
16	100.0	100.0	100.0	100.0	100.0	79.9	92.4
17	100.0	100.0	72.7	99.9	100.0	81.5	89.8
18	100.0	100.0	99.9	100.0	100.0	83.4	92.4
19	100.0	100.0	100.0	100.0	100.0	78.4	94.3
20	99.9	100.0	97.8	100.0	100.0	82.9	88.8
21	100.0	100.0	100.0	100.0	100.0	80.7	91.1
22	100.0	100.0	100.0	100.0	100.0	82.6	90.7
23	100.0	100.0	99.9	100.0	100.0	81.8	96.7
24	82.5	67.8	86.6	86.9	97.0	92.7	86.2
25	99.9	99.9	99.9	100.0	100.0	80.3	91.7
26	100.0	100.0	100.0	100.0	100.0	84.4	87.9
27	100.0	89.6	89.6	100.0	100.0	83.8	88.5
28	100.0	100.0	100.0	100.0	100.0	83.3	98.9
29	100.0	99.9	99.8	99.9	100.0	87.3	88.4
30	100.0	90.2	90.2	100.0	100.0	83.0	97.5
31	100.0	100.0	100.0	99.8	99.8	98.6	86.5
32	100.0	89.7	89.7	100.0	100.0	78.6	93.4
33	83.7	83.7	90.3	100.0	100.0	78.3	86.6
34	100.0	100.0	74.3	100.0	100.0	77.8	86.5
35	100.0	89.8	89.8	100.0	100.0	82.8	81.9
36	100.0	99.9	99.9	100.0	100.0	77.6	86.4
37	100.0	100.0	100.0	100.0	100.0	92.5	80.5
38	100.0	100.0	100.0	100.0	100.0	82.0	90.0
39	100.0	100.0	100.0	100.0	100.0	80.3	90.4
40	100.0	100.0	100.0	100.0	100.0	80.5	83.9
41	100.0	90.2	90.2	100.0	99.9	92.0	88.6
42	100.0	89.6	73.9	100.0	100.0	81.9	90.0
43	100.0	91.7	91.7	97.4	99.7	81.5	82.5
44	100.0	100.0	100.0	100.0	100.0	100.0	99.7
45	100.0	100.0	100.0	100.0	100.0	80.3	96.2
46	100.0	100.0	100.0	100.0	100.0	78.7	94.2
47	100.0	100.0	100.0	100.0	100.0	100.0	98.1
48	100.0	100.0	73.4	99.8	99.8	81.6	88.9
49	100.0	100.0	74.3	100.0	100.0	82.2	81.4
50	100.0	100.0	100.0	100.0	100.0	81.9	89.6
$\overline{x}$	**99.3**	**97.4**	**94.9**	**99.7**	**99.5**	**84.1**	**90.9**

U analiziranju rezultata klasterovanja na slučajnim uzorcima iste veličine, primenili smo sledeće korake:

- Izračunata je tačnost, to jest slaganje rezultujućih klastera na dobijenim uzorcima sa rezultatima dobijenim na celom skupu, za različit broj klastera ($k = 2,3,4,5,6,7,8$). Rezultati su prikazani u tabelama 5.8-5.12.

Prosečno slaganje rezultata klasterovanja za izabrane uzorke iste veličine u zavisnosti od obima uzorka, kao i broja klastera prikazano je na grafikonu 5.1.



Grafikon 5.1 Slaganje rezultata klasterovanja u zavisnosti od broja klastera (k) i obima uzorka (m_u) (baza *Istraživanje zdravlja stanovnika Srbije, 2006. godina*)

U tabeli 5.13 su prikazani sumarni rezultati tabela 5.8-5.12, to jest deskriptivni parametri: srednje vrednosti (aritmetička sredina, medijana) i mere varijabiliteta (opseg vrednosti, standardna devijacija SD). Svaka vrednost u tabeli je izračunata korišćenjem 50 uzoraka definisane veličine i za dati broj klastera (k). Najveće slaganje rezultata klasterovanja je postignuto za $k = 5$, za skoro sve veličine uzorka ($\approx 0.03m$, $0.05m$, $0.1m$, $0.3m$, gde je m obim polaznog skupa), a izuzetak čine uzorci obima $0.01m$, gde je neznatno veće poklapanje postignuto za $k = 6$.

Tabela 5.13 Slaganje rezultata klasterovanja (%) na slučajnim uzorcima i celom skupu. Deskriptivni parametri (baza *Istraživanje zdravlja stanovnika Srbije, 2006. godina*)

Obim uzorka	Deskriptivni parametri	Broj klastera						
		***k*=2**	***k* =3**	***k* =4**	***k* =5**	***k* =6**	***k* =7**	***k* =8**
100	$\bar{x}$	**77.6**	**70.9**	**81.6**	**83.7**	**84.3**	**82.4**	**82.4**
	Med	82.0	66.0	82.0	81.5	84.5	82.0	82.5
	Min	50	52	62	60	60	67	65
	Max	100	100	100	100	100	100	96
	SD	15.4	15.2	9.3	9.4	10.2	9.0	7.7
300	$\bar{x}$	**88.8**	**83.6**	**91.3**	**95.3**	**94.2**	**87.9**	**89.1**
	Med	99.9	89.0	92.7	98.0	98.3	88.2	89.0
	Min	57.3	54.5	65.3	80.7	73.3	71	76.7
	Max	100	100	100	100	100	100	100
	SD	13.9	15.3	8.5	5.9	8.3	7.8	5.6
500	$\bar{x}$	**89.3**	**85.1**	**88.9**	**96.4**	**95.0**	**85.5**	**88.0**
	Med	100.0	91.3	91.2	99.6	99.3	82.9	88.4
	Min	51.6	45.8	65	84	73	76.8	74
	Max	100	100	100	100	100	100	98.6
	SD	15.1	15.6	9.7	5.0	7.9	7.2	6.4
1000	$\bar{x}$	**86.9**	**84.0**	**95.4**	**96.3**	**81.8**	**85.4**	**88.5**
	Med	84.5	84.0	96.3	96.4	76.5	82.2	89.4
	Min	57	56.2	79.2	69.7	72.8	76.7	72.3
	Max	100	100	100	100	100	100	99.6
	SD	9.9	11.0	5.1	4.9	9.8	7.5	5.4
3000	$\bar{x}$	**99.3**	**97.4**	**94.9**	**99.7**	**99.5**	**84.1**	**90.9**
	Med	100.0	100.0	99.9	100.0	100.0	81.9	90.4
	Min	82.5	67.8	72.7	86.9	78	77.6	80.5
	Max	100	100	100	100	100	100	99.7
	SD	3.3	6.0	8.9	1.9	3.1	6.3	5.2

U cilju utvrđivanja da li su dobijene vrednosti slaganja rezultata klasterovanja na uzorcima i celom skupu za $k=5$ statistički značajno veće u odnosu na rezultate dobijene za druge vrednosti broja klastera ($k\neq 5$), primenili smo *Jednofaktorsku analizu varijanse sa ponovljenim merenjima*, posebno za svaki skup uzoraka (obim 100, 300, 500, 1000, 3000). Rezultati su prikazani u tabeli 5.14. Za svaku grupu uzoraka iste veličine važi da se rezultati slaganja značajno razlikuju ($p<0.001$) u zavisnosti od broja klastera (k). Kako nas interesuju rezultati za $k=5$, u tabeli 5.14 su prikazana samo poređenja vrednosti slaganja za $k=5$ u odnosu na dobijene rezultate za ostale vrednosti broja klastera (poslednja kolona).

Tabela 5.14 Poređenje rezultata klasterovanja na uzorcima u odnosu na različite vrednosti broja klastera (baza *Istraživanje zdravlja stanovnika Srbije, 2006. godina*)

Obim uzorka	Razlike vrednosti parova					Test[1]	Parovi klastera[2]	p[3]
	$\overline{x}$	SD	SE	95% CI Donja granica	95% CI Gornja granica			
100	-6.12	14.68	2.08	-10.29	-1.95		k2 - k5	0.005
	-12.80	11.82	1.67	-16.16	-9.44		k3 - k5	<0.001
	-2.08	9.70	1.37	-4.84	0.68	F= 14.304	k4 - k5	0.136
	0.58	9.77	1.38	-2.20	3.36	p<0.001	k6 - k5	0.677
	-1.36	9.75	1.38	-4.13	1.41		k7 - k5	0.329
	-1.34	10.16	1.44	-4.23	1.55		k8 - k5	0.356
300	-6.51	11.42	1.62	-9.75	-3.26		k2 - k5	<0.001
	-11.66	12.99	1.84	-15.35	-7.97		k3 - k5	<0.001
	-3.97	7.29	1.03	-6.04	-1.90	F= 22.598	k4 - k5	<0.001
	-1.10	9.65	1.37	-3.84	1.65	p<0.001	k6 - k5	0.425
	-7.31	7.47	1.06	-9.43	-5.19		k7 - k5	<0.001
	-6.17	8.08	1.14	-8.47	-3.87		k8 - k5	<0.001
500	-7.14	11.41	1.61	-10.38	-3.89		k2 - k5	<0.001
	-11.39	12.37	1.75	-14.90	-7.87		k3 - k5	<0.001
	-7.53	8.72	1.23	-10.01	-5.05	F=39.968,	k4 - k5	<0.001
	-1.43	9.35	1.32	-4.09	1.23	p<0.001	k6 - k5	0.286
	-10.90	9.14	1.29	-13.49	-8.30		k7 - k5	<0.001
	-8.48	7.36	1.04	-10.57	-6.38		k8 - k5	<0.001
1000	-9.34	10.41	1.47	-12.30	-6.38		k2 - k5	<0.001
	-12.31	9.70	1.37	-15.07	-9.56		k3 - k5	<0.001
	-0.89	3.74	0.53	-1.95	0.17	F= 30.344	k4 - k5	0.098
	-14.51	11.01	1.56	-17.64	-11.38	p<0.001	k6 - k5	<0.001
	-10.86	10.05	1.42	-13.72	-8.00		k7 - k5	<0.001
	-7.83	7.08	1.00	-9.84	-5.81		k8 - k5	<0.001
3000	-0.34	2.42	0.34	-1.03	0.35		k2 - k5	0.325
	-2.23	4.80	0.68	-3.59	-0.86		k3 - k5	0.002
	-4.74	8.80	1.24	-7.24	-2.23	F= 61.822	k4 - k5	<0.001
	-0.18	3.46	0.49	-1.17	0.80	p<0.001	k6 - k5	0.708
	-15.60	6.94	0.98	-17.57	-13.62		k7 - k5	<0.001
	-8.72	5.16	0.73	-10.18	-7.25		k8 - k5	<0.001

[1] Jednofaktorska *analiza varijanse sa ponovljenim merenjima;* [2] *Međusobna poređenja za vrednosti k=5 (u oznaci k5) u odnosu na druge vrednosti k;* [3] *nivo značajnosti za međusobna poređenja*

Za uzorke veličine $\geq 0.03m$ važi da je slaganje rezultata za $k=5$ značajno veće u odnosu na slaganje postignuto za ostale vrednosti broja klastera k ($k \in [2,8]$, $k \neq 5$). Izuzetak čine rezultati za $k=6$ (obim uzorka $0.03m$, $0.05m$), $k=4$ (obim uzorka $0.1m$,) odnosno $k=2,6$ (obim uzorka $0.3m$), to jest dobijena razlika za ove vrednosti k nije značajno različita u odnosu na rezultate za $k=5$. Za uzorak obima $0.01m$ je nešto drugačija situacija, to jest značajno su veće vrednosti slaganja rezultata klasterovanja za $k=5$ samo u odnosu na slaganje rezultata klasterovanja za $k=2$, $k=3$. Na osnovu prethodnog, zaključujemo da je najveća tačnost prilikom slaganja rezultata klasterovanja na uzorcima i celom skupu potvrđena upravo za optimalan broj klastera na celom skupu ($k=5$).

Dalji korak u analizi rezultata slaganja klasterovanja (na celom skupu i uzorcima) je međusobno upoređivanje rezultata klasterovanja za dobijeni optimalan broj klastera $k=5$ u odnosu na različite obime uzorka

Najveće odstupanje, to jest najmanja prosečna vrednost slaganja rezultata klasterovanja (83.7%) je za uzorke obima približno $0.01m$, gde je m obim celog skupa, dok se za obim uzorka $\geq 0.03m$, tačnost kreće u intervalu 95.3-99.7%. Nagli skok vrednosti rezultata slaganja postoji za obim $0.03m$ u odnosu na rezultat dobijen za obim uzorka $0.01m$, to jest značajno veće vrednosti rezultata slaganja ($p<0.001$).

Na osnovu prethodnih koraka, za korišćene uzorke približne veličine $0.01m$, $0.03m$, $0.05m$, $0.1m$, $0.3m$, gde je izvučeno 50 uzoraka iste veličine, najbolje rezultate slaganja rezultata na uzorcima i celom skupu postignuto je za optimalnu veličinu uzorku $m_u = t_u m$, gde je $t_u \in [0.03, 0.30]$, a m obim polaznog skupa. Takođe, možemo smatrati da je dovoljna veličina uzorka $m_u = t_u m$, $t_u \in [0.03, 0.10]$.

5.3 MODIFIKOVANI POSTUPAK KLASTEROVANJA ZASNOVAN NA KORIŠĆENJU PROSTIH SLUČAJNIH UZORAKA

Kao što je već naglašeno, vremenska složenost algoritma je funkcija veličine ulaznih podataka, to jest obima skupa koji se posmatra. Jedan od načina da se smanji veličina ulaznih parametara u klaster algoritmu, prilikom rada sa velikim skupovima podataka sa kategorijalnim, odnosno kombinovanim tipovima obeležja je korišćenje prostih slučajnih uzoraka, umesto rada na celom skupu. Za prost slučajan uzorak važi da svaka jedinica osnovnog skupa ima jednaku verovatnoću da bude izabrana u uzorak. Za ovakav uzorak važi reprezentativnost, to jest uzorak na najbolji mogući način opisuje osnovni skup iz koga je izabran. Osnovna ideja predloženog pristupa je da se klaster algoritam primeni na ovako izabranim uzorcima iz datog velikog skupa podataka, umesto na ceo skup. Na ovaj način se smanjuje vreme izvršavanja klaster algoritma, jer kao ulazni podatak za računanje vremenske složenosti algoritma koristimo obim uzoka (m_u), umesto obima osnovnog skupa (m).

Modifikovani postupak klasterovanja velikih skupova podataka predstavlja kombinaciju postupka odabira prostih slučajnih uzoraka određene kardinalnosti i primene odgovarajućeg algoritma klasterovanja (u zavisnosti od tipa obeležja) na ovim uzorcima. Ovaj metod identifikuje kandidate za centre klastera na celom skupu, tako što se od svih uzoraka iste veličine bira onaj uzorak za koji je dobijeno najbolje klastersko rešenje, a korišćenjem kriterijuma validnosti. Modifikovani postupak kalsterovanja primenjuje rešenje dobijeno na ovom uzorku za klasterovanje ostatka skupa, to jest koristi dobijenih k-centara na dobijenom uzorku za klasterovanje ostatka objekata iz celog skupa. Sledi opis predloženog postupka klasterovanja velikih skupova podataka.

Modifikovani postupak klasterovanja:

Korak 1. *Uneti ulazne parametre:*
- *obim osnovnog skupa* m,
- *broj ponavljanja izvlačenja uzorka* $i_{\max}$,
- *koeficijent* t_u *(udeo uzorka u odnosu na osnovni skup),* $t_u \in (0,1)$

Neka je $i = 1$.

Korak 2. *Izabrati na slučajan način uzorak obima* $m_u = t_u m$,

Korak 3. *Primeniti odgovarajući klaster algoritam*

Korak 4. *Izračunati meru interne (ili eksterne) validnosti* v_i

Korak 5. $i \to i+1$. *Ako je* $i = i_{\max}$ *, preći na korak 6; U suprotnom, vratiti se na korak 2*

Korak 6. *Odrediti* $i' = \arg \max_{i=1,\ldots,i_{\max}} v_i$

Korak 7. *Centri klastera za* $i'-ti$ *uzorak predstavljaju centre klastera za ceo skup podataka. Preostale tačke polaznog skupa dodeliti najbližim klasterima.*

Izbor odgovarajućeg algoritma (Korak 3) zavisi od vrste promenljivih uključenih u klaster analizu. Na osnovu izbora odgovarajućeg algoritma sledi i izbor kriterijuma validnosti (Korak 4).

U našem postupku uzorkovanja i primeni klasterovanja na prostim slučajnim uzorcima, koristili smo $i_{\max} = 10$ (baza *Mushrooms*) odnosno $i_{\max} = 50$, (baza *Istraživanje zdravlja Srbije 2006*). Korišćene su veličine uzorka $m_u = t_u m$, gde je $t_u \in \{0.01, 0.03, 0.05, 0.10, 0.30\}$, m je obim celog skupa. Za navedene parametre dobijena je optimalna veličina uzorka (za obe baze podataka) $m_u = t_u m$, $t_u \in [0.03, 0.30]$, pri čemu možemo smatrati da je dovoljna veličina uzorka $m_u = t_u m$, $t_u \in [0.03, 0.10]$.

6. BIHEJVIORALNI FAKTORI RIZIKA

Neko je jednom rekao "Mortalitet (umiranje) nam govori o prošlosti, morbiditet (razboljevanje) o sadašnjosti, a faktori rizika o budućnosti".

6.1 FAKTORI RIZIKA

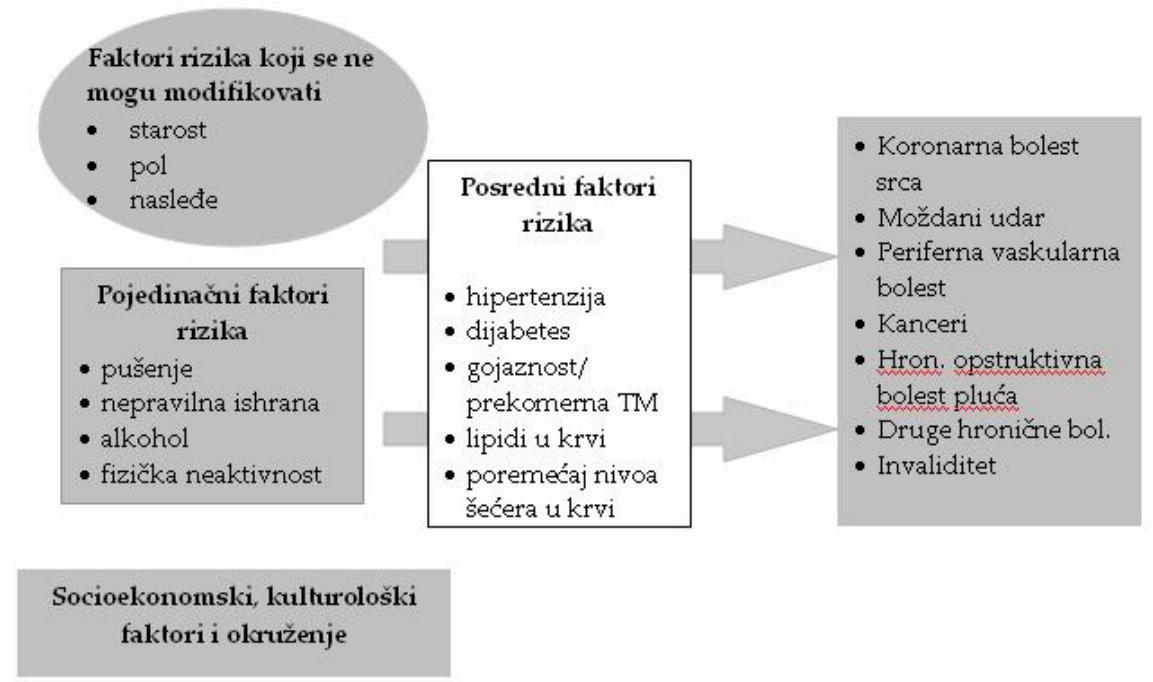
Faktori rizika predstavljaju osobine, zbivanja i navike prisutne kod jedne osobe, grupe ili čitave zajednice koje povećavaju verovatnoću pojavljivanja oboljenja, oštećenja ili smrti.

Po poreklu mogu poticati od ličnih karakteristika i to:
- **Bioloških** (pol, starost, nasleđe, rast i razvoj)
- **Zdravstvenih** (oboljenja u prošlosti ili sadašnjosti)
- **Socioekonomskih i bihejvioralnih** (navike, stil života, obrazovanje)

Prema tome da li su podložni intervenciji, delimo ih na:
- **faktori rizika koji se ne mogu modifikovati** (pol, starost, nasleđe)
- **faktori rizika koji se mogu modifikovati** (pušenje, štetna upotreba alkohola, nedovoljna fizička aktivnost, nepravilna ishrana, gojaznost, povišen nivo masnoće u krvi, hipertenzija i drugo)

Sa kliničkog aspekta najkorisnije su podele faktora rizika na osnovu stepena štetnosti (*glavni* i *ostali*) i prema mogućnosti intervencije (*promenljivi* i *nepromenljivi*).

Bihejvioralni faktori rizika podrazumevaju ona ponašanja koja povećavaju rizik od oboljevanja, kao što su pušenje, štetna upotreba alkohola, loše navike u ishrani, nedovoljna fizička aktivnost. Ovi faktori predstavljaju navike, ponašanja, te ih je moguće menjati u cilju sprečavanja mnogih vrsta hroničnih bolesti kao i prerane smrti.



Slika 6.1. Faktori rizika za kardiovaskularne bolesti, kancer i hronična respiratorna oboljenja
(izvor: [122])

Posebno značajne za nastanak bolesti su četiri rizične navike: pušenje cigareta, štetna upotreba alkohola, nepravilna ishrana i nedovoljna fizička aktivnost, koje su poznate kao "svete četiri" navike (eng. *holy four*). Nasuprot ovim rizičnim navikama, postoje navike koje su povezane sa dobrim zdravljem i dugovečnosti, a u koje spadaju: spavanje sedam do osam sati dnevno, redovan doručak svakog dana i manji obroci između glavnih jela [45].

Faktori rizika koji se mogu modifikovati su uzrok nastanka bolesti koje su vodeći uzrok smrti u razvijenim zemljama [181]. Pušenje cigareta, nepravilna ishrana i drugi bihejvioralni faktori rizika povećavaju rizik za kardiovaskularne bolesti (KVB) i različite vrste karcinoma nezavisno u različitim stopama [251,78]. Osim toga, neke promene u navikama, kao što je povećanje fizičke aktivnosti ili prestanak pušenja su nezavisno povezani sa nižom opštom stopom smrtnosti (mortaliteta) i specifičnog mortaliteta od hroničnog srčanog oboljenja [194].

Procena Svetske zdravstvene organizacije (SZO) je da će 2020. godine 70% svih uzroka smrti biti povezano sa načinom života, gde se kao najvažniji faktori rizika navode: nedovoljna fizička aktivnost, pušenje, štetna upotreba alkohola, neuravnotežena i nepravilna ishrana (prekomerna telesna masa, gojaznost).

Masovne nezarazne bolesti (MNB), koje uključuju kardiovaskularne bolesti, dijabetes, gojaznost, kancer i respiratorna oboljenja su odgovorne za oko dve trećine svih smrtnih slučajeva, pri čemu je polovina usled kardiovaskularnih bolesti. Svetska zdravstvena organizacija je 2013. godine usvojila Globalni akcioni plan za MNB za period 2013-2020.godina [258]. Opšti cilj je smanjenje mortaliteta usled MNB za 25% do 2025. godine, pri čemu su zadati specifični ciljevi: promena navika u ishrani, povećanje nivoa fizičke aktivnosti, smanjenje potrošnje duvana i štetne upotrebe alkohola. Promena stila života ima povoljne efekte na glavne faktore rizika za nastanak KVB (nivo krvnog pristiska, vrednost holesterola u krvi, telesna masa). *Kontis* i kolege su 2014. godine utvrdili da opšti cilj neće u potpunosti biti dostignut do 2025. godine, ali svakako hoće ciljevi koji se odnose na prevenciju KVB [143].

Od svih poremećaja zdravlja, stanovništvo Srbije je najviše opterećeno masovnim nezaraznim bolestima. Vodeći uzroci smrti u našoj zemlji gotovo su identični vodećim uzrocima smrti u razvijenim delovima sveta. To je rezultiralo usvajanjem programa *„Strategija za prevenciju i kontrolu hroničnih nezaraznih bolesti Republike Srbije“* 2009. godine [180], sa ciljem značajnog smanjenja obolevanja, preranog umiranja, opterećenja bolestima, nejednakosti u zdravlju i poboljšanja kvaliteta života stanovnika Srbije.

Prema *Strategiji*, vodeći faktori rizika za nastanak masovnih nezaranih bolesti u Republici Srbiji su: 1. pušenje cigareta, 2. hipertenzija, 3. hiperholesterolemija, 4. štetna upotreba alkohola, 5. gojaznost, 6. nepravilna ishrana, 7. nedovoljna fizička aktivnost (Tabela 6.1).

Tabela 6.1. Prevalencija faktora rizika kod stanovnika Srbije, 2000. i 2006. godina

Prevalencija faktora rizika (%)	**Godina**	
	2000	**2006**
Pušenje	40.5	33.6
Hipertenzija	44.5	46.5
Štetna upotreba alkohola	47.5	40.3
Gojaznost	17.3	18.3
Nedovoljna fizička aktivnost (manje od tri puta nedeljno)	13.7	25.7

6.2 NEDOVOLJNA FIZIČKA AKTIVNOST

Fizička aktivnost se uobičajeno definiše kao kretanje tela koje obavljaju skeletni mišići, a koje dovodi do potrošnje energije. Smatra se da je najefikasnija redovna fizička aktivnost umerenog intenziteta, a podrazumeva onu vrstu aktivnosti koja ubrzava rad srca i stvara osećaj toplote u telu i zadihanost. Sedamdesetih godina prošlog veka, švedski fiziolog *Per-Olof Astrand*, svetski autoritet u oblasti praćenja efekata fizičke aktivnosti na zdravlje, istakao je da ne postoji biološka i psihička funkcija organizma na koju odgovarajuća fizička aktivnost ne deluje povoljno**.** Veliki broj studija u ovoj oblasti pokazao je blagotvorne efekte i potvrdio da fizička ativnost ima neprikosnoveno preventivno, terapijsko i rehabilitaciono dejstvo na sve starosne grupe.

Među brojnim rizičnim ponašanjima, pušenje i nedovoljna fizička aktivnost su od izuzetnog javno zdravstvenog značaja jer se radi o preventabilnim stilovima koji se mogu modifikovati [172]. Nedovoljna fizička aktivnost je povezana sa svim uzrocima smrti [133,190], smanjenjem kvaliteta života [88] i povećanim rizikom za nastanak gojaznosti, dijabetesa, hipertenzije, koronarne srčane bolesti, osteoporoze, fraktura, karcinoma dojke, prostate, depresije [133,190], kao i povećanim rizikom hospitalizacije [110].

Generalno, zdravstvena korist od fizičke aktivnosi se povećava sa povećanjem frekvencije, dužine i intenziteta vežbanja. Svetska zdravstvena organizacija preporučuje za odrasle osobe 30 minuta umerene fizičke aktivnosti svakog dana. Ovo preporučeno vreme može da se podeli u kraće periode, ali ne kraće od 10 minuta. Međutim, i pored ovih preporuka, rezultati istraživanja ukazuju na visoku zastupljenost nedovoljne fizičke aktivnosti u slobodnom vremenu kod odraslog stanovništva. Prevalencija nedovoljne fizičke aktivnosti u svetu se kreće od 31% do 51% sa kojom se povezuje umiranje blizu 2 miliona ljudi. U Evropi je nedovoljna fizička aktivnost povezana sa 600 000 smrtnih slučajeva. Reč je o faktoru rizika zbog koga se godišnje gubi 5 miliona godina zdravog života zbog preranog mortaliteta [239]. Nedovoljna fizička aktivnost je četvrti vodeći rizični faktor umiranja stanovništva.

Procena je da 3.2 miliona smrti i 32.1 milion DALY[7] (učestvuje sa 2.1% globalnog DALY) su povezani sa nedovoljnom fizičkom aktivnosti. Osobe sa nedovoljnom fizičkom aktivnosti imaju 20-30% povećani rizik umiranja od svih uzroka smrti u poredjenju sa fizički aktivnim stanovništvom.

[7] DALY = YLL +YLD; YLL (eng. *Years of Life Lost*)–godine života izgubljene zbog prevremene smrti , YLD (eng. *Years of Life with Disability*) – godine „zdravog“ života izgubljene zbog nesposobnosti određene težine i trajanja

Prema istraživanju *Garretta* i saradnika, skoro 12% depresije i anksioznosti i 31% karcinoma debelog creva, bolesti srca, osteoporoze i moždanog udara je pripisano nedovoljnoj fizičkoj aktivnosti [100]. Redovna fizička aktivnost utiče na psihološko blagostanje, smanjuje stres, anksioznost i osećanja depresije i usamljenosti. Fizička aktivnost smanjuje rizik nastanka ishemične bolesti srca za oko 30%, rizik nastanka dijabetesa za 27% i rizika nastanka karcinoma dojke i kolona za 21-25% [259]. Dodatno, fizička aktivnost snižava i rizik nastanka moždanog udara, hipertenzije i depresije [77,156,250]. Globalno, procena je da je 2008 godine trećina osoba starosti 15 i više godina bilo nedovoljno fizički aktivno, više žene (34.0%) nego muškarci (28.0%)[259].

Pored zdravstvenih efekata, nedovoljna fizička aktivnost ima i ekonomski značaj. Svetska zdravstvena organizacija ističe da pored unapređenja zdravlja i opšteg funkcionalnog stanja, fizička aktivnost može da značajno unapredi funkcionisanje pojedinca na socijalnom, pa i na ekonomskom planu [254], čime se smanjuju troškovi zdravstvene zaštite, povećava produktivnost, povećava efikasnost školskog sistema i smanjuje odsustvovanje sa posla. Takođe se konstatuje da se u mnogim zemljama značajan deo troškova zdravstvene zaštite odnosi na saniranje i tretman stanja povezanih sa nedovoljnom fizičkom aktivnošću i gojaznošću [205]. Rezultati studije koje su sproveli *Garret* i saradnici [100], gde je korišćena analiza troškova bolesti (eng. *cost-of-illness*) koje se pripisuju nedovoljnoj fizičkoj aktivnosti pokazuju da se oko trećina (31%) troškova povezanih sa srčanim oboljenjem, moždanim udarom, karcinomom debelog creva i osteporozom u ispitivanoj populaciji pripisuje nedovoljnoj fizičkoj aktivnosti. Ova studija procenjuje da je fizička neaktivnost, odnosno nedovoljna fizička aktivnost koštala 83.600.000 dolara u 2000. godini za ambulantno i stacionarno lečenje i potraživanje lekova u zdravstvenom sistemu od 1.500.000 osiguranika, dakle 56 dolara po osiguraniku. Kada se vrši procena troškova koji su povezani sa nedovoljnom fizičkom aktivnosti, osim direktnih medicinskih troškova treba uzeti u obzir i indirektne troškove zbog apsentizma, gubitka produktivnosti i izgubljenih godina života. Glavni ekonomski efekat nedovoljne fizičke aktivnosti povezan je sa troškovima u zdravstvu i gubitkom prihoda i produktivnosti koji su povezani sa bolešću i sprečenosti za rad [207]. Prema podacima evropskog zdravstvenog informacionog sistema *EUPHIC (European Union Public Health Information System),* procenjuje se da nedovoljna fizička aktivnost košta državu od 150 do 300 eura po stanovniku godišnje, medicinski troškovi usled nedovoljne fizičke aktivnosti u SAD su procenjeni na 75 milijarde dolara u 2000. godini. U nekim zemljama, direktni troškovi zdravstva koji se odnose na fizičku aktivnost čine više od 2.5% ukupnog budžeta namenjenog zdravstvenoj zaštiti [153].

Prema podacima za 2006. godinu više od dve trećine odraslog stanovništva u Srbiji je nedovoljno fizički aktivno (67.7%). Procenat odraslih stanovnika koji je vežbao više od tri puta nedeljno, tako da se zaduva ili oznoji, iznosio je 25.5%, što je značajno više nego 2000. godine kada je taj procenat bio 13.7%. Dve trećine odraslih stanovnika je slobodno vreme provodilo neaktivno, a skoro trećina zaposlenih se bavi sedentarnim tipom posla [179]. Nedovoljna fizička aktivnost je odgovorna za 8.2% ukupno izgubljenih godina života kod muškaraca i 11.8% kod žena. Opterećenje od raka dojke (15.3%) , karcinoma kolona i rektuma (25.9%), KVB i ishemijske bolesti srca (24.2%), šloga (27.9%), dijabetesa tip 2 (8.3%) se prepisuje nedovoljnoj fizičkoj aktivnosti [16].

6.3 PUŠENJE

Pušenje cigareta, koje je 1964. godine u izveštaju Zdravstvene službe SAD proglašeno faktorom rizika, i danas je jedan od vodećih pojedinačnih faktora rizika za nastanak hroničnih bolesti i jedno od najraširenijih bolesti zavisnosti. Na početku druge decenije 21. veka blizu petine svetske populacije puši cigarete (800 miliona muškaraca i 200 miliona žena). Stope prevalencije pušenja se kreću od 21% u slabo razvijenim zemljama, 30% u visokorazvijenim zemljama i 34% u srednje razvijenim zemljama [245]. Na osnovu izveštaja o globalnoj potrošnji duvana, procenjuje se da će do 2050. godine biti oko 2.2 milijardi ljudi koji će pušiti cigarete. Prema podacima SZO u zemljama Evropske unije puši oko trećine stanovništva, više muškarci (33.0%) nego žene (23%) [259]. Mnoge činjenice ukazuju na razliku među pušačima, pa samim tim i na različite efekte pušenja, kao što su: konstituacija pušača, nasledne osobine, starost, pol, pušački staž, broj dnevno popušenih cigatreta, način pušenja i druge osobine i okolnosti važne za stepen narušenosti zdravlja duvanskim dimom. Duvanska zavisnost je priznata kao bolest prema Međunarodnoj klasifikaciji bolesti SZO (MKB-10) i Dijagnostičkom i statističkom priručniku za duševne poremećaje (DSM-IV), a od strane Američkog Psihijatrijskog udruženja za Dijagnostiku i Statistiku.

Pušenje predstavlja najčešći pojedinačni uzrok prevremenih smrtnih ishoda na koje se može preventivno delovati. Duvan je najznačajniji poznati kancerogen u humanoj populaciji koji ima najveći uticaj na nastanak karcinoma bronha i pluća, karcinoma usta i jezika, grla, ezofagusa, bubrega, pankreasa, ishemijske bolesti srca, srčani udar i hronične opstruktivne bolesti pluća [180]. Procenjuje se da pušenje povećava rizik od moždanog udara, srčanih bolesti i impotencije za 100%. Pušenje povećava rizik od smrti od nedijagnostikovanih srčanih bolesti za 300%. Danas je poznato da je pušenje cigareta najpotentniji faktor rizika za kardiovaskularne bolesti. Pušenje je visoko zastupljeno u svetskoj i našoj populaciji i učestvuje sa 21% u ukupnoj smrtnosti od KVB. U smislu koronarne prevencije, prestanak pušenja dovodi do veće redukcije rizika u poređenju sa bilo kojim drugim izmenjivim faktorom rizika. Povoljnosti od prestanka pušenja su drastične: rizik za KVB kod bivših pušača nakon dve godine od prestanka pušenja se smanjuje toliko da je blizak riziku kod nepušača [44]. Pušenje cigareta povećava rizik za najmanje 50 zdravstvenih stanja koji uključuju demenciju i digestivne probleme [73]. *Ezzati* i saradnici su ocenili da je u 2000. godini, 4.83 (3.94-5.93) miliona slučajeva prerane smrti u svetu pripisano pušenju; 2.41 (1.80-3.15) miliona u zemljama u razvoju i 2.43 (2.13-2.78) miliona u industrijalizovanim zemljama, pri čemu je 3.84 miliona slučajeva kod muškaraca. Vodeći uzroci smrtnosti, povezani sa pušenjem su KVB (1.69

miliona smrtnih slučajeva), hronične opstruktivne bolesti pluća (0.97 miliona smrti), i karcinom pluća (0.85 miliona smrti)[86,87].

Procenjuje se da će broj smrtnih slučajeva usled ishemijskog srčanog oboljenja, moždanog udara i drugih bolesti koji se mogu pripisati duvanu porasti sa 5.4 miliona u 2005. godini na 6.4 miliona u 2013. godini, odnosno 8.3 miliona u 2030. godini [171]. Neke od država članica UNECE[8] (Azerbejdžan, Finska, Island, Norveška, Švedska, SAD) su uspele da smanje prevalencu pušenja, dok druge (Belgija, Irska, Ukrajina) imaju značajan porast broja pušača [266]. Bilano i saradnici su spoveli veliko istraživanje 2014. godine [38], koristeći sveobuhvatne podatke SZO [263,265], sa ciljem ocene trenda prevelence pušenja za period 1990-2010. godine i pravljenje projekcija za 2025. godinu, pri čemu su u analizi uključeni podaci iz 180 država. Zabeleženo je smanjenje prevalence pušenja kod muškaraca u 125 (72%) država, odnosno kod žena u 155 (87%) država, za period 2000-2010. godine. Procenjuje se da će, ukoliko se takav trend nastavi doći do smanjenja prevalence pušenja kod muškaraca samo u 37 (21%) država, odnosno smanjenja prevalence pušenja kod žena u 88 (49%) države, za period 2010-2025. godine. Predviđeno je značajno povećanje prevalence pušenja kod muškaraca u Africi, odnosno muškaraca i žena u istočnom Mediteranu, što ukazuje na potrebu za inteziviranjem aktivnosti usmerenih na kontrolu uporebe duvana na ovim prostorima.

Prevalenca pušenja u Srbiji, zabeležena 2000. godine kod muškaraca (48%) bila je među najvišima u Evropi, a prevalenca pušenja kod žena (33.6%) je bila najviša u Evropi [180]. Nacionalno istraživanje o stilovima života stanovništva Srbije 2014. godine [186] je prvo istraživanje na nacionalnom nivou urađeno u saradnji sa Evropskim monitoring centrom za droge i zavisnosti od droga, agencijom Evropske unije čiji je zadatak da obezbedi uporedive i validne podatke o različitim aspektima u vezi sa zloupotrebom droga. Na osnovu rezultata ovog istraživanja, ukupno 64,5% stanovništva Srbije uzrasta od 18 do 64 godine u toku svog života pušilo je cigarete, dok je njih 40,2% pušilo cigarete u poslednjih 30 dana. Ukupno 36.4% odrasle populacije (40,9% muškaraca i 32% žena) svakodevni su pušači, što odgovara broju od 1 640 000 do 1 762 000 osoba, starosti od 18 do 64 godine. Pušenje je faktor rizika ogovoran za najveće opterećenje mortalitetom stanovništva Srbije: 18% od ukupnih izgubljenih godina života kod muškaraca i 7.9% kod žena. Opterećenje od raka pluća (84.3%), raka grlića materice (9.9%), KVB i ishemijske biolesti srca (18.5%), šloga (17.9%) u Srbiji se prepisuje pušenju [16].

[8] UNECE (engl. *United Nations Economic Commission for Europe*)- Ekonomska komisija za Evropu

6.4 ŠTETNA UPOTREBA ALKOHOLA

Kao i pušenje, upotreba alkohola ima kompleksne fiziološke, bihejvioralne i socijalne veze. Međutim, za razliku od pušenja, konzumiranje alkohola samo po sebi se ne smatra bezuslovno štetnim. Upotreba alkohola treba da se posmatra u kontinuitetu od apstinencije i upotrebe sa malim rizikom (najčešći oblik upotrebe alkohola) preko rizične upotrebe, problema sa unosom alkohola, štetne upotrebe i zloupotrebe alkohola, do ređe, ali mnogo ozbiljnije alkoholne zavisnosti i drugih posledica [207]. Fizički i mentalni zdravstveni efekti upotrebe alkohola se mogu rangirati od korisnog do opasnog/štetnog. Iako je i umerena upotreba alkohola bila sankcionisana u SAD duži vremenski period, samo u poslednjih trideset godina su kvantifikovane njegove objektivne fiziološke zdravstvene koristi.

U mnoštvu dostupne literature danas se opisuju protektivni efekti koje ima ograničen unos alkohola u odnosu na koronarnu srčanu bolest [208]. Štetna upotreba alkohola predstavlja značajan zdravstveni, socijalni i ekonomski problem. Negativne posledice štetne upotrebe alkohola mogu biti:

1. akutne posledice unošenja velikih količina alkohola u kratkom vremenskom peridu, kao što su saobraćajni udesi i trovanje alkoholom
2. hronična oboljenja kao što je ciroza jetre i alkoholna kardiomiopatija
3. primarno hronično oboljenje alkoholne zavisnosti ili alkoholizam [208].

Druge posledice, kao što su razvod braka, ili gubitak posla nisu zdravstveno povezane same po sebi, mada mogu imati negativan zdravtsveni efekat indirektno kroz gubitak prihoda, kao i pristupa sistemu zdravstvene zaštite.

Najnovija klasifikacija SZO daje sledeću definiciju: ”Svako ponovljeno konzumiranje alkoholnih pića i opijanje, koje se nastavlja i pored štetnih neprijatnih posledica, je štetna upotreba alkohola”. Suštinski, alkoholizam je socijalno-medicinski problem, medicinski zato što toksično dejstvo alkohola i metaboličke promene stvaraju oštećenja organizma, a socijalni jer sredina ima značajan uticaj na genezu alkoholizma, ali značajne su i posledice u vidu poremećaja ponašanja alkoholičara u sredini u kojoj živi i radi. Da bi zaista shvatili razmere ove bolesti, svaki broj alkoholičara treba pomnožiti sa tri, jer alkoholizam pojedinaca direktno ugrožava još najmanje tri osobe iz neposrednog okruženja alkoholičara (bračnog partnera, dete/decu, roditelje, prijatelje, kolege sa posla i dr). Štetna upotreba alkohola je treći faktor rizika za opterećenje bolešću u svetu, drugi je u Evropi, a vodeći faktor u zapadnom Pacifiku i Americi. Štetna upotreba alkohola je rangirana kao osmi globalni faktor rizika za smrt, dok je treći vodeći globalni faktor rizika za oboljevanje i invaliditet. Pored brojnih hroničnih i akutnih zdravstvenih efekata, prekomeran unos alkohola je takođe povezan sa rasprostranjenim psiho-socijalnim (društvenim) posledicama, koje uključuju nasilje, zapostavljanje (zanemarivanje) deteta, odsustvovanje sa posla i drugo. Ukupan broj smrtnih

slučajeva prouzrokovanih prekomernim unosom alkohola je procenjen na 2.25 miliona u 2004. godini [261], što predstavlja 4% svih smrtnih slučajeva (6.2% kod muškaraca i 1.1% kod žena), što je gotovo isto kao i od posledica pušenja (4.5%), ili od visokog krvnog pritiska (4.8%). U Evropi godišnje od prekomernog unosa alkohola ili nesreća prouzrokovanih pod njegovim dejstvom, strada 55.000 ljudi. Prekomeran unos alkohola je naročito fatalan u mlađim starosnim kategorijama. Posmatrano u svetskim razmerama, prekomeran unos alkohola je vodeći faktor rizika koji dovodi do smrti kod muškaraca starosti 15-59 godina. Uticaj upotrebe akohola na oboljenja i povrede je povezana sa dve odvojene dimenzije: količinom popijenog alkohola i načinom unosa alkohola.

Štetna upotreba alkohola predstavlja veoma značajan zdravstveni, ali i ekonomski problem. Prema rezultatima istraživanja nacionalnog instituta NIAAA (*National Institute on Alcohol Abuse and Alcoholism*), troškovi povezani sa štetnom upotrebom alkohola u 1998. godini su iznosili 184.6 milijarde dolara (ovo je poslednja godina za koju je takva procena bila na raspolaganju). Od tih troškova, više od 70% su gubici produktivnosti usled preuranjene smrti, povećanog morbiditeta i nesreća prouzrokovanih prekomernim unosom alkohola, dok 10% troškova obuhvata lečenje bolesti zavisnosti od alkohola (NIAAA 2000). Između 15% i 30% pacijenata sa kratkim trajanjem hospitalizacije u opštoj bolnici, imaju probleme sa alkoholom, nezavisno od njihove primarne dijagnoze [207].

Prema rezultatima istraživanja zdravlja stanovništva Srbije u 2006. godini, 40.3% stanovnika je svakodnevno ili povremeno konzumiralo alkohol. Broj stanovnika koji nije konzumirao alkohol u odnosu na 2000. godinu je povećan za 5% [179]. Rezultati Nacionalnog istraživanja o stilovima života stanovništva Srbije, sprovedenog 2014. godine [186], ukazuju da je u prethodnih 12 meseci alkohol konzumiralo ukupno 72.2% odraslih ispitanika (82.1% muškarci, 62% žene). Ekcesivno pijenje alkohola (ovde definisano kao više od 60 grama ili više čistog alkohola u jednoj prilici) jednom nedeljno ili češće, u prethodnih 12 meseci prisutno je kod 3.7% populacije (6.7% muškarci i 0.6% žene). Štetno ili probematično pijenje prisutno je kod 6.2% ukupne populacije (10.6% muškarci i 1.7% žene), pri čemu su u pitanju većinom muškarci i približno trećina populacije uzrasta 18-34 godine starosti. Kada su muškarci u pitanju, količina konzumiranog alkohola se povećava sa godinama, dok se kod žena smanjuje. Iz tog razloga, razlika u količini konzumiranog alkohola u odnosu na pol veća je među starijim stanovništvom. Visokorizično konzumiranje alkohola je zastupljenije kod muškaraca, sa skoro ravnopravnom distribucijom među uzrastima [186]. U Srbiji, opterećenje bolestima raka dojke (8.5%), KVB i ishemijskim bolestima srca (11.4%), šloga (7.1%), povreda (saobraćajne) (35.8%) i samoubistva (6.6%) povezane su sa konzumacijom alkohola. On učestvuje sa 8.5% u ukupnom DALY [16].

6.5 NEPRAVILNA ISHRANA

Hrana je duboko povezana sa našom kulturom, načinom života i emocijama. Pravilna ishrana je jedan od osnovnih preduslova za očuvanje i unapređenje zdravlja ljudi. Osnovni principi pravilne ishrane podrazumevaju redovnost obroka u toku dana, raznovrsnost u izboru namirnica, kao i njihovu odgovarajuću zastupljenost i način pripreme u svakodnevnoj ishrani. Planiranje pravilne, dobro izbalansirane ishrane, ima za cilj postizanje one energetske vrednosti i strukture ishrane pojedinca ili populacije koja može da unapredi zdravlje i prevenira bolest. Svaki prekomeran ili smanjen unos se smatra nepoželjnim za ljudski organizam. Više studija je pokazalo povezanost nepravilne ishrane sa nastankom različitih oboljenja [180].

U savremenim uslovima, posebno u zemljama u tranziciji, navike u ishrani bitno su se promenile (kulturalne promene uzrokovane globalizacijom i razvoj novih tehnika konzervisanja namirnica). Od najranijeg uzrasta konzumira se previše soli, animalnih proteina, monozasićenih masti, a nedovoljna je upotreba namirnica bogatih vlaknima, ugljenim hidratima i polinezasićenim mastima [191]. Sve to stvara uslove za pojavu bolesti koje su povezane sa ishranom, a utiču i na životni vek. Nepravilna ishrana je povezana sa većinom glavnih hroničnih oboljenja, uključujući gojaznost, dijabetes tipa II, kardiovaskularno oboljenje, hipertenziju, loše oralno zdravlje, osteroporozu i različite oblike malignih bolesti. Ishrana je jedan od glavnih potencijalno modifikujućih faktora rizika za hronična oboljenja i može značajno da utiče na globalnu težinu hronične bolesti (eng. *global chronic disease burden*) [207].

Rezultati epidemioloških studija objavljenih krajem prošlog veka, potvrdili su da je u populaciji gde je unos povrća i voća 400 gr ili veći, niža prevalenca KVB i nekih karcinoma. Povećanje unosa voća i povrća je povezano sa nižim rizikom od mnogih hroničnih oboljenja, uključujući maligne bolesti, dijabetes tipa II, kardiovaskularno oboljenje i gojaznost [207,212]. U poređenju sa osobama koje uzimaju manje od tri porcije voća i povrća dnevno, osobe koje često koriste voće i povrće dnevno imaju za 11% niži rizik za nastanak moždanog udara, a one koje imaju više od 5 porcija dnevno imaju niži rizik i do 26% [53]. Ipak, istraživanja koja su ispitivala povezanost između visokog unosa voća i povrća i kontrole telesne mase su nekonzistentna [212,241]. Ovo može bar delimično biti posledica prirodne kompleksnosti i dugoročne prirode istraživanja ishrane [207].

Istraživanja sprovedena u Srbiji pokazuju da su u svim kategorijama stanovništva zastupljene pogrešne navike u ishrani, od dece do starijih osoba, što predstavlja značajan faktor rizika od oboljevanja i umiranja od masovnih

nezaraznih bolesti, od kojih u našoj sredini prednjače bolesti srca i krvnih sudova i maligne bolesti. Korišćenje životinjskih masti za pripremanje obroka je zastupljeno kod 33.8% stanovnika, a najviše u Vojvodini 43.8% i zapadnoj Srbiji 44.5%. U 2006 godini više od polovine stanovnika (57.2%) je koristilo u ishrani beli hleb, a svega 14.8% stanovnika crni, ražani i slične vrste hleba. Manje od jednom nedeljno ribu je jelo 48.7% odraslih stanovnika. Sveže povrće je svakodnevno konzumiralo oko 55%, a voće 44% stanovništva. Svaki peti odrasli stanovnik Srbije pri izboru načina ishrane nikad nije razmišljao o svom zdravlju. Nedovoljno unošenje voća i povrća u ishrani može se smatrati važnim faktorom rizika u našoj populaciji, pošto je odgovorno za 3.3% ukupno izgubljenih godina života kod muškaraca i 2.9% kod žena. Takođe je odgovorno i za 4.3% opterećenja kardiovaskularnim bolestima i ishemijskom bolesti srca, šloga (3.8%) [16].

6.6 KOMBINOVANO DELOVANJE DVA ILI VIŠE BIHEJVIORALNIH FAKTORA RIZIKA

Kombinacija dva ili više bihejvioralnih faktora rizika je obično povezana sa većim rizikom za KVB ili kancer, nego što je to očekivano na osnovu sumiranja njihovih pojedinačnih efekata [94,159]. Globalno, više od 70% mortaliteta od KVB, 40% hroničnih respiratornih oboljenja, 34% mortaliteta usled kancera i oko 50% mortaliteta usled ostalih hroničnih oboljenja se pripisuju malom broju poznatih faktora rizika koji se mogu modifikovati [86,187].

Pušenje cigareta, nepravilna ishrana, štetna upotreba alkohola i nizak nivo fizičke aktivnosti su važne determinante oboljevanja i smrtnosti [256]. Istraživanja ukazuju da je pušenje odgovorno za 4.1% globalnog opterećenja društva bolešću, dok štetna upotreba alkohola, nedovoljna fizička aktivnost i nepravilna ishrana doprinose sa 4%, 1.3% i 1.8% [86]. Pored toga, podaci ukazuju na to da rizici štetni po zdravlje, kao što je nedovoljna fizička aktivnost, gojaznost i pušački status uzrokuju visoke troškove zdravstvene zaštite [70]. Ovo se čini relevantnim za kompanije koje se bave zdravstvenim osiguranjem prilikom razmatranja strateških investicija u prevenciji takvih rizika [200]. Ekonomski efekti ovih rizičnih ponašanja nisu zanemarljivi. Značajan deo sredstava koji se izdvajaju za zdravstvenu zaštitu stanovništva izdvajaju se za zadovoljenje zdravstvenih potreba uzrokovanih bolestima vezanim za dijagnostiku, lečenje, rehabilitaciju kao i troškove zbog apsentizma, onesposobljenosti i preranog mortaliteta.

Loše životne navike koje uključuju nedovoljnu fizičku aktivnost, nepravilnu ishranu i pušenje su strogo povezane sa KVB, dijabetesom, respiratornim i malignim oboljenjima. Ova četiri oboljenja su odgovorna za preko 50% mortaliteta širom sveta [73]. U većini studija se ističe da su nezdrava ishrana i nedovoljna fizička aktivnost ključni faktori rizika za masovne hronične nezarazne bolesti [256]. *Yusuf* i saradnici su sproveli anamnestičku (*case-control*) studiju akutnog infarkta miokarda u 52 zemlje (obuhvaćeni svi kontinenti), koja je uključila 15152 slučaja i 14820 kontrola. Rezultati ukazuju da dnevni unos voća/povrća i redovna fizička aktivnost smanjuje rizik od infarkta miokarda za 40%, a ukoliko se izbegava pušenje, rizik se smanjuje za više od 75% [269].

Hronične bolesti, kao što su karcinom, KVB, moždani udar i dijabetes su odgovorne za većinu smrtnih slučajeva u SAD. Procena je da je približno 70% do 90% tih smrtnih slučajeva nastalo kao posledica loše ishrane, sedentarnog načina života i pušenja. Osim toga, od 23% odraslih pušača, 77% se nezdravo hrani i 78% ima povećan zdravstveni rizik usled nedovoljne fizičke aktivnosti [8]. Prospektivna studija koja je sprovedena na odrasloj

populaciji (starosti 45-79 godina) u Velikoj Britaniji [138], čiji je cilj bio da se kvantifikuje potencijalni kombinovani uticaj četiri zdravstvene navike na mortalitet kod muškaraca i žena, potvrdila je da se rizik za ukupni mortalitet značajno povećava sa povećanjem broja bihejvioralnih faktora rizika. Oni ispitanici koji su imali sva četiri fakora rizika su imali relativni rizik za mortalitet koji je 4.04 puta veći (95% CI:2.95-5.54) u poređenju sa ispitanicima koji su bili bez faktora rizika. Navike povezane sa stilom žvota kao što su pušenje, štetna upotreba alkohola ili nepravilna ishrana, povezane su sa skoro polovinom slučajeva karcinoma u Velikoj Britanji. Prema Istraživanju karcinoma u Velikoj Britaniji (*Cancer Research UK*), ovi karcinomi "koji se mogu sprečiti" čine 45% svih karcinoma kod muškaraca i 40% karcinoma kod žena, tj. više od 100.000 slučajeva godišnje. *Chiuve* i saradnici su u prospektivnoj studiji analizirali zdrave životne navike kod 42 847 muškaraca starosti od 40 do 75 godina [60]. Autori navode da muškarci sa pet poželjnih (nisko-rizičnih) zdravstvenih navika, u koje spadaju nepušači sa BMI<25kg/m^2 koji se bave umerenom do napornom fizičkom aktivnosti, podrazumevajući umeren unos alkohola i nalaze se u grupi prvih 40% zdravih navika u ishrani (na osnovu numeričkog skora) imaju 0.14 puta manji rizik za koronarnu srčanu bolest u odnosu na muškarce koji nemaju nijednu od ovih karakeristika. *Knoops* i saradnici navode da su kod 2339 muškaraca i žena starosti 70-90 godina u 11 evropskih zemalja zastupljenost mediteranskog tipa ishrane, umeren unos alkohola, fizička aktivnost i nepušački status povezani sa stopom mortaliteta koja je tri puta manja u odnosu na stopu mortaliteta kod ispitanika koji nemaju ove zdrave životne navike [142]. Autori navode da je nedostatak zdravih životnih navika povezan sa populacionim atributivnim rizikom za 60% ukupnog mortaliteta, 64% mortaliteta uzrokovanog koronarnom bolesti srca, 61% mortaliteta od KVB, a 60% od tumora.

Faktori povezani sa nepravilnom ishranom, kao što je visok krvni pritisak, nizak holesterol, nizak unos voća i povrća, velika vrednost indeksa telesne mase (BMI) i štetna upotreba alkohola najviše doprinose hroničnim oboljenjima u razvijenim zemljama [86]. Faktori povezani sa nepravilnom ishranom i nedovoljnom fizičkom aktivnosti u SAD su vodeći uzrok prevremene smrti svake godine, a na drugom mestu je pušenje [181]. Dokazano je da poboljšan način života može da redukuje rizik od progresije ka dijabetesu za 58% tokom 4 godine. Druge populacione studije su pokazale da se 80% koronarne bolesti srca i do 90% slučajeva dijabetesa tipa II potencijalno može izbeći promenom nezdravih životnih navika, a oko trećine karcinoma se može izbeći unosom zdrave hrane, održavanjem normalne telesne mase i redovnim vežbanjem tokom života [257]. Rezultati istraživanja sprovedenog 2015. godine [247] ukazuju da se bihejvioralni faktori rizika grupišu zajedno sa mentalnim oboljenjima, ali kako je reč o studiji preseka, ovde se ne može govoriti o kauzalnosti.

Colditz procenjuje da je gojaznost odgovorna za 7% svih direktnih troškova zdravstvene zaštite u SAD i da je nedovoljna fizička aktivnost odgovorna za dodatnih 2.4% svih troškova zdravstvene zaštite [62]. Indirektni troškovi povezani sa gojaznošću i nedovoljnom fizičkom aktivnosti povećavaju račun za još 5% troškova zdravstvene zaštite [86]. *Pronk* i saradnici su procenjivali razliku u troškovima zdravstvene zaštite između pacijenata sa i bez rizičnih faktora za masovne nezarazne bolesti (fizička aktivnost, BMI i pušenje) i otkrili da zdraviji način života koji uključuje bavljenje fizičkom aktivnosti tri puta nedeljno, umeren BMI i nepušački status redukuju troškove zdravstvene zaštite za 49% u poređenju sa nezdravim načinom života [200].

6.7 ZNAČAJ PRIMENE KLASTER ANALIZE U DEFINISANJU KLASTERA BIHEJVIORALNIH FAKTORA RIZIKA

Mada je klaster analiza veoma rasprostranjena u sociologiji, komercijalnom istraživanju marketinga i mnogim drugim oblastima, još uvek nije dovoljno razmatrana njena primena u socijalnoj medicini, epidemiologiji, uopšte u oblasti javnog zdravlja [223].

Bihejvioralni faktori rizika kao što su pušenje, štetna upotreba alkohola, loše navike u ishrani i nedovoljna fizička aktivnost, dugo zauzimaju centralno mesto u istraživanjima iz oblasti javnog zdravlja [235]. S obzirom na to da ovi faktori rizika predstavljaju navike, odnosno ponašanja, moguće ih je menjati u cilju sprečavanja mnogih vrsta hroničnih bolesti kao i prerane smrti. Poslednjih decenija, MNB su vodeći uzrok smrti i nesposobnosti širom sveta. Masovne nezarazne bolesti čine 59% od 56.5 miliona smrtnih slučajeva godišnje i predstavljaju 45.9% globalnog opterećenja društva bolešću (eng. *global burden disease*). Pet od izdvojenih deset vodećih faktora rizika za globalno opterećenje društva bolešću navedenih u izveštaju SZO [264], gojaznost, visok krvni pritisak, visok holesterol, štetna upotreba alkohola i pušenje cigareta su nezavisni faktori i često deluju kombinovano i predstavljaju glavne uzročnike ovih oboljenja [85]. Prema navodima nekih autora, oko 70-80% smrtnih slučajeva u razvijenim zemljama sveta je povezano sa načinom života [223]. Uzroci visoke učestalosti MNB su značajne i brze promene u načinu života savremenih ljudi, a najviše se ogledaju u načinu ishrane, nivou fizičke aktivnosti, povećanoj upotrebi alkohola i duvana (koji predstavljaju bihejvioralne faktore rizika) [65,161,191,260]. Mogućnosti prevencije i kontrole masovnih nezaraznih bolesti postoje. U osnovi ovih bolesti su faktori rizika koji su najvećim delom preventabilni (bihejvioralni faktori rizika), čijim se sprečavanjem ili modifikovanjem doprinosi prevenciji i kontroli vodećih MNB, doprinoseći mentalnom, fizičkom i socijalnom blagostanju, a smanjenju onesposobljenosti, preranog umiranja, bola i patnje. Zbog svega ovoga, bihejvioralni faktori rizika imaju značajnu ulogu u istraživanjima iz oblasti javnog zdravlja [255]. Mnoge preventivne i interventne mere su i dalje fokusirane na modifikovanje pojedinačnih faktora ponašanja. U uslovima planiranja opsežnih preventivnih programa i intervencija, korisno je znati stepen u kome se najvažniji bihejvioralni faktori rizika (nepravilna ishrana, nedovoljna fizička aktivnost, pušenje cigareta i štetna upotreba alkohola) grupišu u određenim populacionim kategorijama i da li se na osnovu toga mogu izdvojiti tipične grupe. Ovo je veoma značajno zbog usmeravanja mera i aktivnosti na bolesti koje povezuju zajednički faktori rizika, socioekonomske determinante i mogućnosti prevencije, što se smatra efektivnim pristupom, imajući u vidu multifaktorsku etiologiju MNB i

čestu udruženost faktora rizika i bolesti (komorbiditet) kod pojedinaca, posebno osetljivih kategorija stanovništva.

Kombinacija dva ili više rizičnih bihejvioralnih faktora je obično povezana sa većim rizikom za kardiovaskularne bolesti i kancer, nego što se može očekivati na osnovu sume pojedinačnih efekata. Ukoliko je veća prevalenca prisustva kombinacije nego što je očekivano na osnovu prevalence odvojenih faktora rizika, govorimo o klasterovanju. Klasterovanje, to jest grupisanje bihejvioralnih faktora rizika kod istog pojedinca može imati multiplicirane efekte na rizik od oboljevanja.

Metod klaster analize omogućava ovu vrstu holističkog pristupa u identifikaciji relevantnih ciljnih grupa za analizu faktora rizika i predviđanje određenih interventnih mera**.** Podela velikog heterogenog skupa podataka na manje homogene podskupove omogućava lakše korišćenje, odvojeno modeliranje i analizu podataka u odnosu na ovako izdvojene podskupove**.** Klaster analiza je u većem broju istraživanja iz oblasti javnog zdravlja ograničena na korelaciju između dva bihejvioralna faktora rizika i ne razmatraju se klasteri formirani na osnovu multidimenzionalnih karakteristika [58,193,198,224]. U literaturi, međusobne veze između bihejvioralnih faktora rizika su ispitivane korišćenjem različitih statističkih tehnika, kao što su: faktorska analiza, diskriminantna analiza, analiza glavnih komponenti, logistička regresija, kao i kombinacije navedenih metoda. Prednost klaster analize u profilisanju bihejvioralnih faktora rizika je da klaster analiza može grupisati pojedince, a ne promenljive, kao što je slučaj kod primene faktorske analize.

7. BIHEJVIORALNI FAKTORI RIZIKA I KLASTER ANALIZA. REZULTATI

Uzorak odraslog stanovništva koji je korišćen u analizi je obuhvatio 11300 ispitanika (tabela 7.1). Više od polovine ispitanika (51.8%) je ženskog pola. Prosečna starost ispitanika iznosi 48 godina (SD=16.4), pri čemu je najstariji ispitanik starosti 95 godina. Više od polovine odraslih ispitanika (53.1%) živi u gradskoj/urbanoj sredini. Više od polovine ispitanika (55.3%) je iz Centralne Srbije, nešto više od četvrtine iz Vojvodine, a 2102 (18.6%) iz Beograda. Veći broj ispitanika živi u bračnoj ili vanbračnoj zajednici (70.0%). Najviše ispitanika je sa završenom srednjom školom (51.9%), nešto više od trećine ispitanika je nižeg obrazovanja (bez škole, nepotpuna/potpuna osnovna škola), a svaki sedmi odrasli ispitanik je sa završenom višom, ili visokom školom (13.9%). Što se tiče materijalnog stanja stanovništva, veći je broj ispitanika sa srednjim ili boljim materijalnim stanjem (59% vs.41%).

Tabela 7.1. Sociodemografske karakteristike odraslog stanovništva Srbije

Sociodemografske karakteristike	N	%
Pol		
muškarci	5449	48.2
žene	5851	51.8
Starost		
20-34	2860	25.3
35-54	4372	38.7
55 i više	4068	36.0
Tip naselja		
urbano	6003	53.1
ostalo	5297	46.9
Region		
Vojvodina	2952	26.1
Beograd	2102	18.6
Centralna Srbija	6246	55.3
Bračni status		
u braku	7905	70.0
žive sami	3356	29.7
Nivo obrazovanja		
niže	3864	34.2
srednja škola	5869	51.9
viša/visoka	1567	13.9
Materijalno stanje		
lošije	4630	41.0
srednje/ bolje	6670	59.0
Ukupno	11300	100.0

7.1 PREVALENCIJA BIHEJVIORALNIH FAKTORA

Više od trećine odraslih ispitanika su pušači (35.9%), svaki deseti ispitanik (10.2%) ima prekomeran unos alkohola. Nezdrave navike u pogledu unosa hrane (nedovoljan unos voća i povrća) ima svaki peti odrasli ispitanik (20.4%), a skoro dve trećine odraslih (61.8%) ima nedovoljan nivo fizičke aktivnosti. Podaci su prikazani u tabeli 7.2.

Pušenje je značajno zastupljenije kod muškaraca u odnosu na žene (42.0% vs. 30.0%, p<0.001), stanovnika urbanog područja (38.4% vs. 33.0%, p<0.001), stanovnika Vojvodine (39,8%, p<0.001), ispitanika sa završenom srednjom školom (42.3%, p<0.001), opada sa starošću (najniža prevalenca 20.3% kod starijih od 55 godina). Nije utvrđena značajna razlika u prevalenci pušenja u odnosu na bračni status odraslih ispitanika.

Štetna upotreba alkohola ima iste karakteristike kao i prevalenca pušenja u odnosu na pol i bračno stanje, pa je tako najzastupljenija kod muškaraca (18.0%), kod ispitanika starosti 20-34 godine (10.9%), stanovnika ruralnog područja (10.9%), ispitanika sa završenom srednjom školom (10.4%), ispitanika lošijeg materijalnog stanja (11.2%). Nije utvrđena značajna razlika u prevalenci štetne upotrebe alkohola u odnosu na bračni status odraslih ispitanika.

Nepravilna ishrana (nedovoljan unos voća/povrća) je najzastupljenija kod muškaraca (22.9%), starosti 20-34 godine (21,6%), odraslih ispitanika koji nisu u braku (23,2%). Prevalenca nepravilne ishrane opada sa višim nivoom obrazovanja i boljim materijalnim stanjem, pa je tako najviša kod ispitanika sa nižim obrazovanjem (24,8%) i ispitanika lošijeg materijalnog stanja (25,0%). Nije utvrđena značajna razlika u prevalenci nepravilne ishrane u odnosu na tip naselja.

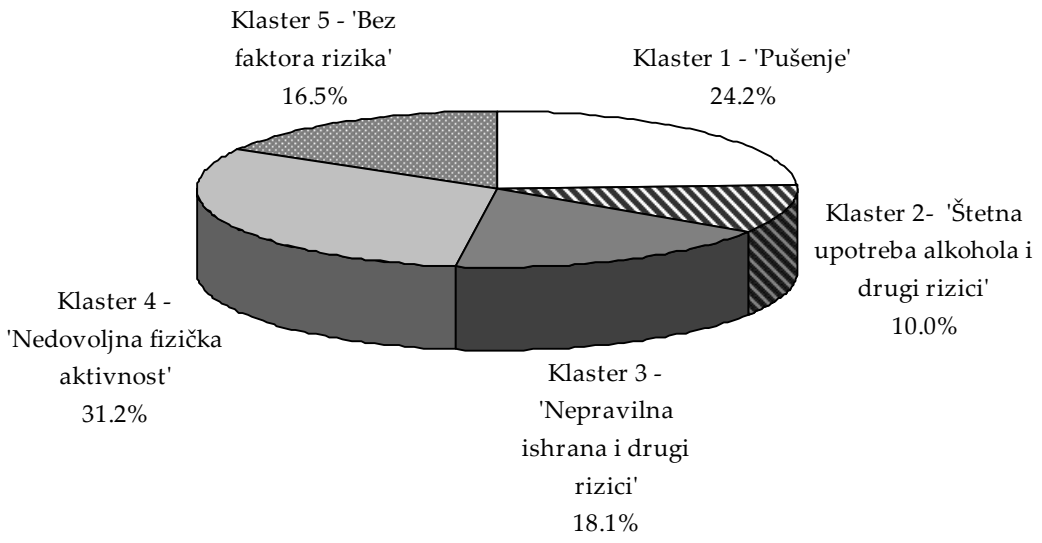
Nedovoljna fizička aktivnost je za razliku od pušenja, štetne upotrebe alkohola i nepravilne ishrane, najzastupljenija kod žena (71.1%), raste sa starošću (najviša u kategoriji 55 i više godina, 73.7%) Ova rizična navika je nazastupljenija kod ispitanika sa nižim obrazovanjem (69.2%) i ispitanika srednjeg/boljeg materijalnog stanje (62,1%). Nije utvrđena značajna razlika u prevalenci nedovoljne fizičke aktivnosti u odnosu na tip naselja i bračni status ispitanika.

Tabela 7.2. Prevalencija bihejvioralnih faktora rizika u odnosu na sociodemografske karakteristike odraslog stanovništva Srbije

Sociodemogr. karakteristike	Ukupan broj ispitanika	Pušenje (%)	Štetna upotreba alkohola i drugi rizici (%)	Nepravilna ishrana i drugi rizici (%)	Nedovoljna fizička aktivnost (%)
Pol					
muškarci	5449	42.0	18.0	22.9	51.8
žene	5851	30.2	2.6	18.0	71.1
Starost					
20-34	2860	43.5	10.9	21.6	53.5
35-54	4372	45.4	10.5	18.8	56.2
55 i više	4068	20.3	8.9	21.3	73.7
Tip naselja					
urbano	6003	38.4	9.2	20.9	61.8
ostalo	5297	33.0	10.9	19.8	61.8
Region					
Vojvodina	2952	39.8	12.0	26.8	58.3
Beograd	2102	37.8	8.6	14.9	63.9
Centralna Srbija	6246	33.4	9.6	19.2	62.7
Bračni status (%)					
u braku	7905	36.0	9.8	19.2	61.9
žive sami	3356	35.7	10.8	23.2	61.4
Nivo obrazovanja					
niže	3864	27.6	9.8	24.8	69.2
srednja škola	5869	42.3	10.7	19.3	57.0
viša/visoka	1567	32.4	8.3	13.7	61.6
Materijalno stanje					
lošije	4630	35.0	11.2	25.0	61.5
srednje/bolje	6670	36.5	9.2	17.2	62.1
Ukupno	11300	**35.9**	**10.0**	**20.4**	**61.8**

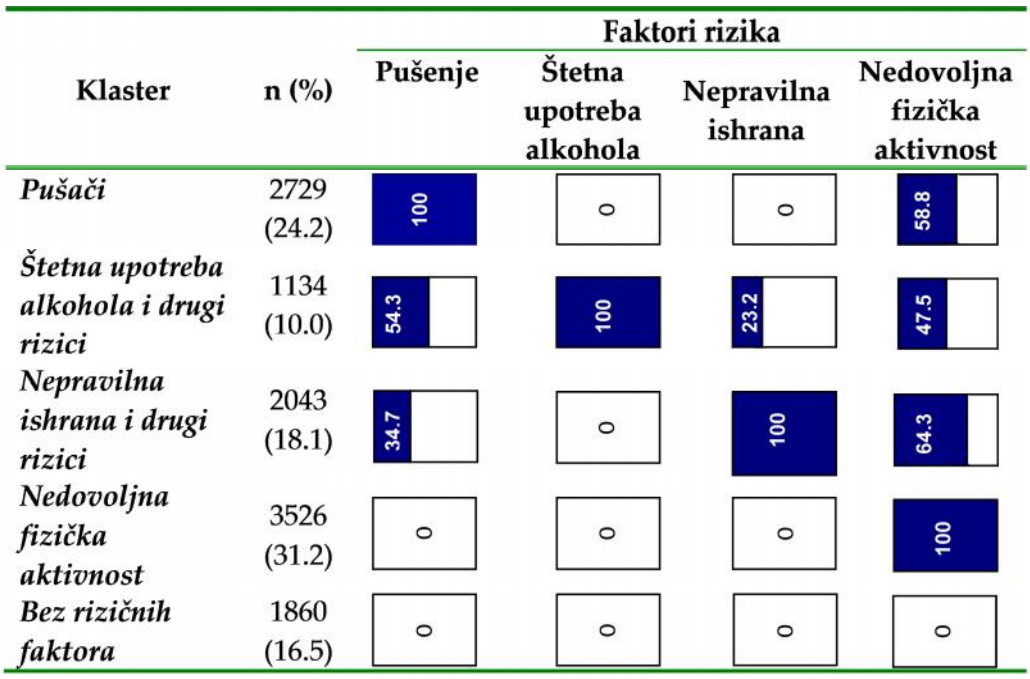
7.2 KLASTEROVANJE BIHEJVIORALNIH FAKTORA RIZIKA

Primenom *TSCA* klaster algoritma na uzorku odraslog stanovništva Srbije, a korišćenjem četiri bihejvioralna faktora rizika (pušenje cigareta, štetna upotreba alkohola, nepravilna ishrana i nedovoljna fizička aktivnost) izdvojeno je pet klastera bihejvioralnih faktora rizika kod odraslog stanovništva Srbije (grafikon 7.1). U određivanju optimalnog broja klastera, korišćen je Bajesov informacioni kriterijum (BIC). Vrednost *Silhouette* indeksa, kao mere interne validnosti iznosi 0.7, čime je potvrđen kvalitet dobijenog klasterskog rešenja.



Grafikon 7.1. Distribucija izdvojenih klastera sa karakterističnim bihejvioralnim faktorima rizika

Skoro trećina ispitanika (31.2%) pripada klasteru *Nedovoljna fizička aktivnost*, nešto manje od četvrtine ipsitanika (24.2%) pripada klasteru *Pušači*, nešto manje od šestine (18.1%) klasteru *Nepravilna ishrana i druge rizične navike*, a svaki deseti ispitanik (10.0%) pripada klasteru *Prekomeran unos alkohola i druge rizične navike*. Oko šestina ispitanika (16.5%) pripada klasteru *Bez rizičnih faktora*.



Slika 7.1. Distribucija bihejvioralnih faktora rizika u izdvojenim klasterima

Klaster 1. 'Pušači'
Ovaj klaster obuhvata skoro četvrtinu svih ispitanika (24.2%). Svi ispitanici u ovom klasteru su pušači, pri čemu se ova navika javlja izolovana (41.2%), ili u kombinaciji sa nedovoljnom fizičkom aktivnosti (58.8%).

Klaster 2. 'Štetna upotreba alkohola i druge rizične navike'
Najmanji klaster (10.0%) čine svi ispitanici koji imaju rizično ponašanje u vezi sa konzumiranjem alkohola. Prekomeran unos alkohola je kombinovan sa drugim rizičnim navikama, pušenjem (54.3%), nepravilnom ishranom (23.2%) i fizičkom neaktivnosti (47.5%).

Klaster 3. 'Nepravilna ishrana i druge rizične navike'
Treći klaster (18.1%) čine ispitanici koji se nepravilno hrane (imaju neredovan unos voća i povrća). Svaki treći ispitanik u ovom klasteru je pušač (34.7%), a skoro dve trećine ispitanika je fizički neaktivno (64.3%).

Klaster 4. 'Nedovoljna fizička aktivnost'
Ovaj klaster sadrži najveći broj ispitanika (N=3526) i predstavlja skoro trećinu svih ispitanika (31.2%). Specifičnost četvrtog klastera čini to što svi ispitanici imaju nizak nivo fizičke aktivnosti i nemaju nijednu drugu rizičnu naviku (nerizičan unos alkohola, zastupljena pravilna ishrana, poželjan nivo fizičke aktivnosti).

Klaster 5. 'Bez rizičnih faktora'

Šestinu ukupnog broja ispitanika koji su uključeni u postupak klaster analize (16.5%) čine ispitanici koji nemaju zdravstveno-rizična ponašanja (ne puše, nemaju prekomeran unos alkohola, svakodnevno unose sveže voće i povrća, imaju poželjan nivo fizičke aktivnosti) (slika 7.1).

Bihejvioralni faktori rizika se retko pojavljuju izolovani, već uglavnom udruženi sa drugim faktorima rizika. Sledeća tabela ilustruje na koji način je ova osobina (udruženost faktora rizika) povezana sa strukturom dobijenih klastera.

Tabela 7.3 Distribucija broja faktora rizika u dobijenim klasterima

Klaster	Zastupljenost dominantnog faktora u klasteru u odnosu na ceo skup [1](%)	Broj faktora rizika u klasteru	Udruženost faktora rizika	%[2]
Pušači	67.3	1	pušenje	42.2
		2	pušenje +fizička neaktivnost	58.8
Štetna upotreba alkohola i drugi rizici	100	1	alkohol	19.5
		2	alkohol+ 1 faktor[3]	42.1
		3	alkohol+2 faktora[3]	32.2
		4	alkohol+3 faktora[3]	6.2
Nepravilna ishrana i drugi rizici	88.6	1	nepravilna ishrana	21.5
		2	neprav. ishrana +1 faktor[4]	57.5
		3	neprav. ishrana +2 faktora[4]	20.5
Nedovoljna fizička aktivnost	50.5	1	nedovoljna fizička aktivnost	100

[1] Ukupan broj ispitanika u klasteru/ukupan broj ispitanika u uzorku sa „dominantnim" faktorom rizika za taj klaster. „Dominantan" rizik za klaster *Pušenje* je pušenje cigareta, itd.
[2] zastupljenost u odnosu na klaster
[3] pušenje, nedovoljna fizička aktivnost, alkohol
[4] pušenje, nedovoljna fizička aktivnost; alkohol nije zastupljen u ovom klasteru

Većina ispitanika (88.6%) koji se nepravilno hrane (nedovoljan unos voća i povrća) pripada klasteru *Nepravilna ishrana i drugi rizici*. Više od dve trećine (67.3%) svih pušača pripada klasteru *Pušenje*, dok se ostali pušači nalaze u jednom od druga dva rizična klastera (*Nepravilna ishrana*, *Štetna upotreba alkohola i drugi rizici*). Polovina svih ispitanika sa nedovoljnim nivoom fizičke aktivnosti se nalazi u klasteru *Nedovoljna fizička aktivnost*

(koji se sastoji isključivo od ispitanika sa ovim faktorom rizika), a drugi deo ispitanika sa ovim faktorom rizika pripada jednom od tri preostala rizična klastera. Razlog ovome je osobina nagomilavanja, to jest udruživanja ovih faktora rizika (pušenje, nedovoljna fizička aktivnost) sa drugim faktorima, o čemu će više biti reči u diskusiji (deo 8.2.2).

7.3 SOCIODEMOGRAFSKE KARAKTERISTIKE KLASTERA

Tabela 7.4. Distribucija izdvojenih klastera (%) u odnosu na socio-demografske karakteristike odraslog stanovništva

Socio-demografske karakterist.	Ukupno (N=11300)	Klaster 1 'Pušači'	Klaster 2 'Štetna upotreba alkohola i drugi rizici'	Klaster 3 'Nepravilna ishrana i drugi rizici	Klaster 4 'Nedovolj na fizička aktivnost'	Klaster 5 'Bez faktora rizika'
Pol**						
muškarci	48.2	48.5	86.6	49.8	32.5	52.5
žene	51.8	51.5	13.4	50.2	67.5	47.5
Starost**						
20-34	25.3	29.4	27.5	26.9	19.1	27.9
35-54	38.7	50.9	40.6	34.9	29.2	41.8
55 i više	36.0	19.8	31.9	38.2	51.6	30.3
Tip naselja**						
urbano	53.1	58.0	48.9	55.1	49.2	53.9
ostalo	46.9	42.0	51.1	44.9	50.8	46.1
Region						
Vojvodina	26.1	25.6	31.3	33.7	21.2	24.8
Beograd	18.6	20.8	16.0	13.1	20.7	19.0
Centralna Srbija	55.3	53.6	52.7	53.2	58.1	56.2
Bračni status **						
oženjen/ udata	70.2	72.3	68.1	66.6	70.6	71.6
živi sam/a	29.8	27.7	31.9	33.4	29.4	28.4
Nivo obrazov**						
niže obrazov.	34.2	23.3	33.2	41.6	42.7	26.7
srednja škola	51.9	62.4	55.3	48.9	42.7	55.2
viša/visoka	13.9	14.3	11.5	9.5	14.6	18.1
Materijalno stanje**						
lošije	41.0	36.6	45.8	49.9	40.2	36.2
srednje/bolje	59.0	63.5	54.3	50.1	59.8	63.7
Ukupno	100.0	24.2	10.0	18.1	31.2	16.5

**p<0.001

Utvrđena je statistički značajna razlika (p<0.001) u distribuciji svakog od posmatranih sociodemografskih obeležja (pol, starost, tip naselja, bračni status, nivo obrazovanja, materijalno stanje) između izdvojenih klastera (tabela 7.4). Prvi klaster, koga čine samo pušači, odnosno pušači koji su i fizički neaktivni, većinom čine stanovnici gradske sredine, iz Beograda, sa završenom srednjom školom, ispitanici koji su u braku i boljeg su

materijalnog stanja. Drugi klaster, koji je karakterističan po prekomernom unosu alkohola i drugim rizičnim navikama, većinu (86.6%) čine muškarci, ispitanici sa završenom srednjom školom, siromašniji i stanovnici Beograda. Za treći klaster, koji čine ispitanici koji se nepravilno hrane i imaju i druge rizične navike, karakteristično je da su iz gradske sredine, nižeg obrazovanja, lošijeg materijalnog stanja i nisu u braku. Četvrti klaster čine ispitanici koji su fizički neaktivni i većinu čine žene (67.5%), veći je broj stanovnika ruralnog područja, ispitanika sa nižim obrazovanjem. Za poslednji klaster 'Bez faktora rizika', karakteristično je da su većim delom zastupljeni muškarci, ispitanici koji su u bračnoj/vanbračnoj zajednici, stanovnici urbanog područja, ispitanici boljeg materijalnog stanja, sa završenom srednjom ili višom, odnosno visokom školom.

Rezultati univarijantne analize (χ^2 test) ne daju potpunu sliku o vezi između navedenih sociodemografskih karakteristika i pripadnosti nekom od izdvojenih klastera, te je sledeći korak u analizi povezanosti ovih obeležja primena multivarijantne analize, koja nam omogućava da kontrolišemo uticaj pridruženih promenljivih. U tabeli 7.5 su prikazani rezultati multinomne logističke regresije, sa zavisnom varijablom pripadnost odgovarajućem klasteru, u odnosu na referentnu kategoriju, klaster ”Bez faktora rizika”. Model ocenjuje verovatnoću da ispitanik pripada određenom klasteru u poređenju sa referentnom grupom, bez faktora rizika. Interpretacija modela je uključila prikaz odnosa šansi (*odds ratio*), zajedno sa 95% intervalom poverenja (CI). Odnos šansi *(OR)* da osoba pripada određenom klasteru (u odnosu na referentnu kategoriju) je izračunat za svaku sociodemografsku karakteristiku posebno, pri čemu su ostale sociodemografske varijable smatrane konstantnim.

Za razliku od ispitanika koji su bez faktora rizika, za pripadnike prvog klastera 'Pušači' važi da je značajno manja šansa da su muškog pola (OR=0.87) i da su starosti 55 i više godina (OR=0.48) u odnosu na 20-34 godine, a značajno je veća šansa da su stanovnici urbanog područja (OR =1.32), lošijeg materijalnog stanja (OR=1.26) i značajno je veća šansa da su sa završenom srednjom školom (OR=1.39), odnosno nižeg obrazovanja (OR=1.28) u odnosu na ispitanike sa završenom višom ili visokom školom.

Za razliku od ispitanika koji su bez faktora rizika, za pripadnike drugog klastera 'Štetna upotreba alkohola i druge rizične navike' važi da je značajno veća šansa da su muškog pola (OR=6.17), žive sami (OR=1.24), lošijeg materijalnog stanja (OR=1.29), nižeg obrazovanja (OR=2.08) u odnosu na ispitanike sa završenom višom, odnosno visokom školom i značajno je veća šansa da su iz Vojvodine (OR=1.45) u odnosu na stanovnike Centralne Srbije.

Tabela 7.5. Multinomna logistička regresija sa zavisnom varijablom-pripadnost određenom klasteru

Socio-demografske karakteristike	*'Pušači'*		*'Prekomeran unos alkohola i drugi rizici'*		*'Nepravilna ishrana i drugi rizici*		*'Nedovoljna fizička aktivnost'*	
	OR (95% CI)	**p**	**OR (95% CI)**	**p**	**OR (95% CI)**	**p**	**OR (95% CI)**	**p**
Pol								
ženski	1.00 (referentna)		1.00 (referentna)		1.00 (referentna)		1.00 (referentna)	
muški	0.86 (0.77-0.97)	0.018	6.15 (5.06-7.48)	<0.001	0.95 (0.84-1.09)	0.478	0.43 (0.38, 0.48)	<0.001
Starost								
20-34	1.00 (referentna)		1.00 (referentna)		1.00 (referentna)		1.00 (referentna)	
35-54	1.17 (1.01-1.36)	0.039	1.04 (0.85-1.27)	0.698	0.88 (0.74-1.04)	0.131	0.99 (0.85-1.16)	0.940
55+	0.64 (0.39-0.58)	<0.001	0.95 (0.77-1.18)	0.662	1.07 (0.90-1.28)	0.452	2.32 (1.97-2.74)	<0.001
Bračno stanje								
bračna/vanbračna zajednica	1.00 (referentna)		1.00 (referentna)		1.00 (referentna)		1.00 (referentna)	
neoženjeni/neudate	0.98 (0.85-1.13)	0.761	1.21 (1.02-1.45)	0.032	1.20 (1.03-1.38)	0.016	0.98 (0.86-1.13)	0.820
Tip naselja								
ostalo	1.00 (referentna)		1.00 (referentna)		1.00 (referentna)		1.00 (referentna)	
urbano	1.32 (1.15-1.52)	<0.001	1.07 (0.90-1.27)	0.456	1.52 (1.31-1.76)	<0.001	0.92 (0.80-1.05)	0.225
Region								
Centralna Srbija	1.00 (referentna)		1.00 (referentna)		1.00 (referentna)		1.00 (referentna)	
Vojvodina	1.08 (0.94-1.25)	0.291	1.45 (1.22-1.73)	<0.001	1.40 (1.21-1.63)	<0.001	0.79 (0.69-0.91)	0.001
Beograd	1.17 (0.99-1.38)	0.061	1.13 (0.90-1.41)	0.300	0.90 (0.74-1.08)	0.256	1.23 (1.05-1.45)	0.011
Nivo obrazovanja								
viša/visoka škola	1.00 (referentna)		1.00 (referentna)		1.00 (referentna)		1.00 (referentna)	
srednja škola	1.39 (1.17-1.65)	<0.001	1.44 (1.14-1.82)	0.002	1.62 (1.32-1.98)	<0.001	1.02 (0.87-1.21)	0.776
bez škole/ osnovna škola	1.26 (1.02-1.55)	0.033	2.05 (1.57-2.69)	<0.001	2.59 (2.06-3.27)	<0.001	1.33 (1.09-1.62)	0.004
Materijalno stanje								
bolje	1.00 (referentna)		1.00 (referentna)		1.00 (referentna)		1.00 (referentna)	
lošije/srednje	1.25 (1.08-1.45)	0.003	1.28 (1.06-1.55)	0.010	1.55 (1.32-1.82)	<0.001	1.23 (1.07-1.43)	0.004

OR= odnos šansi (eng. *odds ratio*), CI = interval poverenja (eng. *confidence interval*)
Referentna kategorija za zavisnu varijablu: Klaster 1 „Bez faktora rizika"

Za pripadnike trećeg klastera, 'Nepravilna ishrana i druge rizične navike' važi da je značajno veća šansa da žive sami (OR=1.20), žive u urbanoj sredini (OR=1.51), lošijeg su materijalnog stanja (OR=1.54), značajno je veća šansa da su sa završenom srednjom školom (OR=1.62), odnosno nižeg obrazovanja (OR=2.57) u odnosu na ispitanike sa završenom višom ili visokom školom, kao i značajno je veća šansa da žive u Vojvodini (OR=1.40) nego u Centralnoj Srbiji.

Za razliku od ispitanika koji su bez faktora rizika (peti klaster), za pripadnike četvrtog klastera 'Nedovoljna fizička aktivnost' važi da je značajno veća šansa da su u njemu niže obrazovani (OR = 1.23), ispitanici koji žive sami (OR = 1.50), stanovnici urbane sredine (OR =1.83), stanovnici Beograda (OR=1.21), lošijeg materijalnog statusa (OR = 1.22), značajno je veća šansa da su starosti 35-54 godine (OR =1.17), odnosno starosti 55 i više godine (OR =2.98) u odnosu na starost 20-34 godine. Značajno je manja šansa da su pripadnici ovog klastera muškarci (OR=0.42) i stanovnici Vojvodine (OR=0.78). Podaci su prikazani u tabeli 7.5.

8. DISKUSIJA

U ovom delu disertacije bavićemo se primenom klaster analize na velike skupove podataka sa kategorijalnim, odnosno kombinovanim obeležjima, posmatrano sa aspekta numeričke optimizacije, kao i sa aspekta oblasti javnog zdravlja. U prvom delu analiziraćemo rezultate primenjenih klaster algoritama na velike skupove podataka sa kategorijalnim, odnosno kombinovanim obeležjima, sa osvrtom na prednosti, odnosno nedostatke u radu sa ovakvim podacima. Analiziraćemo prednosti predloženog pristupa u klasterovanju velikih skupova podataka, gde najvažniji korak postupka predstavlja korišćenje prostih slučajnih uzoraka određene klardinalnosti, umesto celog skupa. U drugom delu analiziraćemo rezultate klasterovanja za podatke iz *Istraživanja zdravlja Srbije 2006*, kako u odnosu na bihejvioralne faktore rizika, tako i u odnosu na sociodemografske karakteristike. Treći deo ovog poglavlja predstavlja osvrt na ograničenja istraživanja.

8.1 KLASTEROVANJE VELIKIH SKUPOVA PODATAKA

Kao što smo već naglasili, veliki problem u klasterovanju predstavljaju veliki skupovi podataka sa kategorijalnim, odnosno kombinovanim obeležjima.

8.1.1 Veliki skupovi podataka sa kategorijalnim obeležjima

Specifičnost korišćene baze podataka *Mushrooms* sa kategorijalnim podacima je da za nju već postoji stvarna klasifikacija (podela na dve klase), tako da smo dobijene rezultate korišćenjem različitih klaster algoritama upoređivali sa već postojećom klasifikacijom. Rezultati naše analize na celom skupu podataka pokazuju da su najefikasniji algoritmi TSCA i Ward–ov algoritam, što je u skladu sa rezultatima drugih istraživanja. *Hands* i *Everitt* su upoređivanjem pet hijerarhijskih metoda klasterovanja (SL, CL, AL, centroid i Ward) za multivarijantne binarne podatke [114], utvrdili da *Ward*-ov metod daje najbolje rezultate. *Milligan* i *Cooper* [174] su koristili četiri aglomerativne hijerarhijske metode (SL, CL, AL i Ward). Rezultati njihovog istraživanja ukazuju da SL metoda manje efikasna dok Ward-ov metod i AL daju najbolje ukupno obnavljanje (eng. *recovery*). Rezultati *Kriksciuniene* i saradnika [146] ukazuju da *Ward*-ov metod ima najbolje rezultate klasterovanja podataka sa binarnim obeležjima, za slučaj dva dobro odvojena klastera. *Milliagan* i saradnici [177] ukazuju na loše rezultate Ward-ovog algoritama u prisustvu autlajera.

Cao i saradnici su korišćenjem eksterne mere validnosti, tj. tačnosti za algoritam *k-modusa* i podelu na dva klastera ($k=2$) dobili rezultate koji u zavisnosti od korišćene mere različitosti iznose: 78.63% (*Huang*), odnosno 79.37% (*Ng*) i 82.43% (za predloženu novu meru) [51]. *He* i saradnici [120] su izvođenjem klasterovanja na serijama slučajnih uzoraka, a korišćenjem prosečne tačnosti utvrdili da je njihov predloženi modifikovani algoritam *k-modusa* (76.44%) efikasniji u odnosu na klasičan algoritam *k-modusa* (73.81%). *Bai* i saradnici [30,32] su pokazali da u zavisnosti od početne selekcije centara, prosečna tačnost iznosi od 71.76%, odnosno 73.18% za klasičan algoritam *k-modusa* (slučajan izbor centara), 87.54% za modifikovani algoritam *k-modusa*, 88.92% za njihov predloženi metod. Predloženi algoritam je efikasan u slučaju malog broja klastera.

Rezultati istraživanja gde su primenjeni drugi klaster algoritmi za iste podatke (baza *Mushroom*) pokazuju slične, ali ne i bolje rezultate u odnosu na TSCA i Ward-ov algoritam. Prema rezultatima *Guha* i saradnika [108], ROCK algoritam deli objekte u 21 čist klaster. Rezultati primenjenih hijerarhijskih algoritama (*Jaccard* mera) u našoj analizi su skoro identični rezultatima primenjenog ROCK algoritma. *Andritsos* i saradnici [13] su primenili *LIMBO* klaster algoritam (za $k=2$) na uzorku veličine 1000 Utvrđeno je da ovaj algoritam ima prednosti i bolje karakteristike (*preciznost* P=0.91, *minimalna greška klasifikacije* E_{min}= 0.11) u odnosu na ROCK algoritam.

Prednosti Ward-ovog algoritma su visoka tačnost u poređenju sa drugim metodama klasterovanja i nisu potrebne a-priori informacije o broju klastera. Nedostaci Ward-ovog algoritma su kao i kod ostalih hijerarhisjkih algoritama vremenska složenost, ukoliko se primenjuje na velikom skupu podataka i osetljivost na prisustvo autlajera.

Finch u svom radu vrši poređenje mera udaljenosti (rastojanja) u klaster analizi sa dihotomnim (binarnim) podacima [89] u smislu korektno grupisanja subjekata. Rezultati ukazuju da se tri od četiri analizirane mere (*Rusell/Rao Index, Jaccard coefficient, Matching coefficient, Dice's coefficient*) slično ponašaju i mogu da daju stope tačnog ponavljanja klastera između 60% i 90%. Ovo je potvrđeno i u našim rezultatima, te smo u daljoj analizi koristili *Jaccard*-ov koeficijent kao meru sličnosti.

Naši rezultati su zasnovani na poređenju mera tačnosti za različite algoritme, pri čemu smo vršili upoređivanje dobijenih klastera sa postojećom klasifikacijom. *Rubinov* i saradnici [215] navode neka od mogućih objašnjenja zbog čega se dobijeni klasteri u potpunosti ne poklapaju sa klasama:

- Izbor modela klasterovanja se ne podudara sa strukturom klasifikacije.

- Skup podataka sadrži visoku proporciju slučajeva šuma i/ili mogućih grešaka, koje se javljaju npr. prilikom prikupljanja podataka
- Neke karakteristike značajnije povezuju objekte nego sama pripadnost klasi

8.1.2 Veliki skupovi podataka sa kombinovanim obeležjima

Prilikom rada sa kombinovanim tipovima obeležja, primenili smo TSCA algoritam. Prednosti ovog algoritma su sledeće:

- Omogućava rad sa kategorijalnim, numeričkim, kao i kombinovanim tipovima obeležja
- Veoma brzo izračunavanje za velike skupove podataka
- Proizvodi klastere različite veličine.
- Postoji kriterijum za određivanje optimalnog broja klastera (BIC ili AIC), za razliku od većine drugih algoritama klasterovanja.
- Za podatke sa kombinovanim tipovima obeležja, eksperimentalni rezultati [59] su potvrdili da TSCA ne samo da stvara bolji kvalitet klastera od tradicionalnog algoritma *k-sredina*, već i pokazuje dobra svojstva merljivosti i sposobnost da tačno identifikuje broj klastera.

Chiu i saradnici (2001) su objavili odlične rezultate za predloženi algoritam za automatsko određivanje broja klastera. Za oko 98% generisanih skupova podataka, TSCA omogućava pronalaženje optimalnog broja klastera. Zbog svega navedenog AIC ili BIC kriterijum se mogu koristiti kao početni korak za dalju klaster analizu [59]

I pored navedenih prednosti, TSCA ima i određene nedostatke:

- Kao i algoritam *k-sredina*, zavisi od redosleda unosa podataka
- Subjekti koji imaju bar jedan nedostajući podatak su isključeni iz analize, što može da značajno smanji veličinu uzorka i utiče na rezultate klasterovanja
- Različito dodeljivanje težina kategorijalnim i neprekidnim promenljivama, pri čemu su kategorijalnim promenljivama dodeljene veće težine.

Rezultati za kombinovane tipove obeležja podržavaju hipotezu da su kategorijalne promenljive dominantne u rezultatima, jer razlike u nominalnim promenljivama daju veće težine nego razlike u neprekidnim promenljivama. Ovaj rezultat može da dovede do previše prilagođene (eng. *overfitting)* razlike između klastera za kategorijalne promenljive i premalo prilagođene (eng. *underfitting*) razlike između klastera za neprekidne

promenljive. U cilju izbegavanja favorizovanja kategorijalnih obeležja prilikom rada sa kombinovanim podacima, poželjno je dodeljivanje težinskih koeficijenata za svaku od kontinuiranih promenljivih u svakom klasteru. Međutim, ove razlike su manje ozbiljne, a glavni problem predstavlja pogrešno određen broj klastera, kao posledica bezuslovnog ponderisanja. Ukoliko je predviđen tačan broj klastera, pristrasnost je mala [18].

8.1.3 Modifikovani postupak klasterovanja zasnovanog na korišćenju prostih slučajnih uzoraka

Modifikovani postupak klasterovanja predstavlja kombinaciju postupka uzorkovanja i odgovarajućeg klaster algoritma. Nedostatak ovog pristupa može predstavljati činjenica da se ne analizira ceo skup podataka, što može dovesti do nedostatka važnih informacija za same klastere skupa, pa samim tim i nedostatka nekih klastera, ili da se pogrešno identifikuju neki klasteri. Zbog svega navedenog važan je način izbora i optimalna veličina uzorka.

Prednost predloženog pristupa klasterovanju prilikom rada sa velikim skupovima podataka sa kategorijalnim, odnosno kombinovanim tipovima obeležja se sastoji u sledećem:

1. Korišćenje prostih slučajnih uzoraka obezbeđuje reprezentativnost podataka.

2. Rezultati dobijeni primenom klaster algoritma na ovako dobijenim uzorcima, možemo uopštiti na posmatrani skup.

3. Vremenska složenost algoritama je funkcija veličine podataka (obima skupa m). Za velike skupove podataka (veliki broj podataka) značajno se smanjuje vreme izvršavanja (vremenska složenost), što naročito važi kod hijerarhijskih algoritama (vremenska složenost $O\left(m^3\right)$).

4. Izbor odgovarajućeg algoritma (u zavisnosti od vrste podataka, to jest skale merenja promenljivih) i izbor odgovarajuće veličine uzorka m_u, gde je $m_u = t_u m,\ t_u \in (0,1)$ daju dobre rezultate (u smilu tačnosti), pa umesto rada na celom skupu, dovoljna je primena klaster algoritma na ovim uzorcima. Na osnovu naših rezultata za klasterovanje kategorijalnih, odnosno kombinovanih tipova podataka, dobijena je optimalna vrednost koeficijenta $t_u \in [0.03, 0.30]$, to jest $t_u \in [0.03, 0.10]$ za dovoljnu veličinu uzorka. Međutim, kako je koeficijent $t_u \in (0,1)$ neprekidna promenljiva, rezultate dobijene za

vrednosti koeficijenta $t_u \in \{0.01, 0.03, 0.05, 0.1, 0.3\}$ ne možemo generalizovati na velike skupove podataka. U analizi je korišćen broj uzoraka $i_{\max} = 10$ (*Mushrooms*), odnosno $i_{\max} = 50$ (baza *Istraživanje zdravlja*), a nisu analizirane druge vrednosti broja izvlačenja uzoraka iste veličine. Neophodna je dalja analiza slučajeva za različit broj promenljivih uključenih u klaster analizu. Osim toga, sam pojam velikog skupa je veoma uopštena, tako da dobijeno rešenje zavisi i od m, to jest koliko je „veliko" m. Optimalna vrednost koeficijenta t_u je svakako predmet daljih istraživanja.

Ovako modifikovani postupak klasterovanja pravi kompromis između dobijene preciznosti (neznatno niža u odnosu na ceo skup) i postignute efikasnosti (manja vremenska složenost).

Kao što je već ranije navedeno, postoje različiti pristupi u klasterovanju velikih skupova podataka sa ciljem smanjenja neophodnog vremena za izvršavanje, a mi smo naveli samo neke od njih. Međutim, efikasnost ovih algoritama važi samo pod određenim pretpostavkama. CLARA algoritam je efikasan u klasterovanju velikog skupa podataka samo ukoliko je mali obim uzorka korišćen u PAM algoritmu. Za velike i kompleksne skupove podataka, mali uzorci ne mogu prezentovati pravu, stvarnu raspodelu podataka, pa CLARA nije idealan algoritam za velike skupove podataka. Neka je sa $m_{u_{\max}}$ označen maksimalan broj objekata koji ovaj algoritam može obraditi za razumno vreme. U slučajevima kada je obim posmatranog skupa m mnogo veći od $m_{u_{\max}}$, rezultujući klasteri dobijeni na ovako malom uzorku mogu sadržati potpuno „promašene" podatke. Sa povećanjem broja klastera, performanse CLARA algoritma brzo slabe, ispod prihvatljivog nivoa. U eksperimentima koji su vršili ovi autori, CLARANS se pokazao efikasnijim nego CLARA algoritam u pogledu vremena izvršavanja. Međutim, ovi eksperimenti su testirali CLARANS algoritam na skupu podataka koji sadrži samo 100 objekata [173]. Utvrđeni su bolji rezultati *Novel algoritma k-sredina* [3] zasnovanog *na postupku razlaganja (divide and conguer)* u odnosu na klasičan algoritam *k-sredina*. Korišćeno je deset baza podataka različite veličine, a testiranje je vršeno korišćenjem mera eksterne validnosti (čistoća, entropija). Međutim, ovaj algoritam je efikasan isključivo u radu sa numeričkim neprekidnim obeležjima i kada je $m < 50000$. Postupak Bagirova (opisan u delu 2.5) je predviđen za redukovanje velikih skupova podataka sa numeričkim neprekidnim obeležjima.

8.1.4 Transformacija kategorijalnih promenljivih u binarne promenljive

Postupak transformisanja kategorijalnih promenljivih u binarne promenljive nosi sa sobom određene nedostatke. Neka je sa m označen ukupan broj objekata, sa $n^{(2)}$ broj kategorijalnih promenljivih, a sa n_j broj kategorija j-te promenljive, gde je $j = 1,\ldots,n^{(2)}$. Ukoliko nominalnu promenljivu sa n_j kategorija transformišemo u n_j binarnih promenljivih, zbog međusobne isključivosti kategorija polazne varijable, dobijamo nove međusobno zavisne promenljive. Nezavisnost promenljivih uključenih u klaster analizu predstavlja jednu od pretpostavki za primenu određenih klaster algoritama.

U slučaju transformisanih binarnih varijabli, mera različitosti za dva objekta sa jednom kategorijalnom promenljivom je 0 ili 1, što ne reprezentuje uvek stvarnu razliku između dva objekta. Ukoliko dva objekta imaju $n^{(2)}$ kategorijalnih promenljivih, maksimalna vrednost razlike između dva objekta tada iznosi $n^{(2)}$, što takođe ne predstavlja uvek meru stvarne razlike dva objekta.

Nedostatak ovakvog pristupa transformisanja promenljivih je da on podrazumeva rad sa većim brojem promenljivih u odnosu na originalni skup podataka. Ulazna matrica podataka za klaster algoritam je reda $m \cdot n^{(2)}$ za originalne podatke, odnosno $m \cdot (n_1 + n_2 + \ldots + n_{n^{(2)}})$, za transformisane promenljive. Kako za promenljive sa nominalnom skalom merenja važi $n_i \geq 2$ (za svako $i = 1,\ldots,n^{(2)}$), pa je dimenzionalnost novog skupa podataka značajno veća od dimenzionalnosti prvobitnog skupa, to jest $m \cdot (n_1 + n_2 + \ldots + n_{n^{(2)}}) \geq 2 \cdot m \cdot n^{(2)}$. Ovo se naročito odnosi na situacije kada imamo bar nešto od navedenog: veliki broj podataka (m), veliki broj kategorijalnih promenljivih $(n^{(2)})$.

8.2 BIHEJVIORALNI FAKTORI RIZIKA I KLASTER ANALIZA

Od svih poremećaja zdravlja stanovništva Srbije, najveće je opterećenje hroničnim bolestima, odgovornim za izgubljene godine života zbog preveremenog mortaliteta i invaliditeta i smanjenja kvaliteta života. Kombinacije bihejvioralnih faktora rizika mogu imati sinergističke efekte na rizik za razvoj karcinoma i drugih negativnih zdravstvenih ishoda, pa je razumevanje obrazaca ovih zdravstveno-rizičnih navika korisno prilikom modeliranja incidence oboljenja. Na osnovu najnovije meta-analize iz 2015. godine [139], većina obuhvaćenih studija je utvrdila efikasnost intervencija koje se odnose na višestruke bihejvioralne faktore rizika kod odraslih. Očekuje se da preduzimanje koraka u pogledu višestrukih zdravstveno-rizičnih ponašanja ima veći uticaj u oblasti javnog zdravlja, nego intervencije koje se odnose na pojedinačne bihejvioralne faktore rizika.

8.2.1 Klasterovanje bihejvioralnih faktora rizika

Kod odraslog stanovništva Srbije, starijeg od 20 godina identifikovano je pet homogenih klastera bihejvioralnih faktora rizika. Jedan klaster čine odrasli sa zdravim stilovima života ('Bez rizičnih faktora'), dva klastera ('Štetna upotreba alkohola i drugi rizici', 'Nezdrava ishrana i drugi rizici') karakteristiše prisustvo višestrukih faktora rizika. Klaster 'Pušenje' predstavlja kombinaciju zdravih i nezdravih stilova života (pušenje, nedovoljna fizička aktivnost), dok odrasli u klasteru 'Fizički neaktivni' nemaju druge faktore rizika osim nedovoljne fizičke aktivnosti. Ovako definisani klasteri bihejbioralnih faktora rizika su veoma slični rezultatima *Schneider* i saradnika [223], koji su sproveli istraživanje u Nemačkoj među odraslim stanovništvom starosti od 50 do 70 godina.

Nacionalna istraživanja rađena na velikim reprezentativnim uzorcima pokazuju da je među odraslima veoma često istovremeno prisustvo dve ili više rizične navike: 68% u Velikoj Britaniji [198], 55% u Holandiji [224], 52% u SAD [65], 59% u Brazilu [230]. Nezdrave životne navike se grupišu u određenim kombinacijama. Dobijeni rezultati ukazuju da je pušenje faktor rizika koji ima najveću verovatnoću grupisanja sa drugim bihejvioralnim faktorima rizika, što je saglasno rezultatima drugih studija [151,193,198]. Rezultati prethodnih studija ukazuju da pušači konzumiraju manje voća i povrća [39,192,226], nepravilno se hrane [68,163,239], piju više alkohola [58,193] i manje su fizički aktivni u odnosu na nepušače [150]. Prisustvo

kombinacije rizičnih navika pušenja cigareta i štetne upotrebe alkohola su deo duboko ukorenjene kulture u nekim zemljama [162]. Rezultati naše studije ukazuju da pušenje i nedovoljna fizička aktivnost imaju najveću verovatnoću zajedničkog grupisanja. Sa njima se klasteruje štetna upotreba alkohola u klasteru 'Štetna upotreba alkohola i drugi rizici', a nedovoljan unos voća i povrća se klasteruje sa ovom kombinacijom faktora rizika (pušenje i nedovoljna fizička aktivnost) u klasteru 'Nezdrava ishrana i drugi rizici'. Dobijeni rezultati su u skladu sa nacionalnim istraživanjem rađenim 2014. godine među odraslim stanovništvom Portugalije [64]. Kako pušači imaju veću verovatnoću prisustva drugih višestrukih bihejvioralnih faktora rizika, u poređenju sa nepušačima, moguće je da ovakvo grupisanje višestrukih faktora u velikoj meri povezano sa pušenjem [58]. Mada postoje jasne veze imeđu dva bihejvioralna faktora rizika (pušenje i štetna upotreba alkohola, pušenje i nedovoljna fizička aktivnost), rezultati koji govore o udruživanju višestrukih faktora rizika su i dalje različiti i zavise od izbora bihejvioralnih faktora rizika i načina na koji su oni definisani [70,205,253]. Suprotno očekivanjima, rezultati nekih studija ukazuju da je veća šansa da su fizički aktivni ljudi pušači i/ili imaju prekomeran unos alkohola. Jedan broj autora [224] objašnjava ovu pojavu time da ljudi posle bavljenja organizovanim vidom sportskih aktivnosti (uglavnom kod kolektivnih sportova) uglavnom nastavljaju zajedničko druženje, pa je i veća verovatnoća da puše i/ili piju. Drugo ili dodatno objašnjenje za ovakav vid grupisanja bi moglo da bude to da su ljudi koji se bave manuelnim poslovima češće pušači i više konzumiraju alkohol [65]. Istraživanja povezanosti unosa alkohola i fizičke aktivnosti, pokazuju kombinovane rezultate [70], a prema istraživanju rađenom među odraslim stanovništvom Holandije [224] svi bihejvioralni faktori rizika se značajno grupišu, osim nedovoljne fizičke aktivnosti i prekomernog unosa alkohola. Prevalenca prekomernog unosa alkohola je podjednako prisutna kod fizički aktivnog i nedovoljno aktivnog odraslog stanovništva. Autori ovo objašnjaju time da nemaju informaciju o mestima na kojima fizički aktivni ispitanici konzumiraju alkohol, pa prema tome ne mogu ni da istražuju da li je tolika zastupljenost „štetne upotrebe alkohola“ kod fizički aktivnih ispitanika usled konzumiranja alkohola, nakon sporta u kantinama klubova. Jača povezanost je dobijena između štetne upotrebe alkohola i pušenja, pri čemu je grupisanje ova dva faktora izraženije kod mlađe populacije, a što opet može biti objašnjeno nedovoljnom samosvešću u ovoj kategoriji stanovništva.

8.2.2 Sociodemografske karakteristike klastera

Pušenje cigareta, štetna upotreba alkohola, nezdrava ishrana i nedovoljna fizička aktivnost grupišu se unutar određenih sociodemografskih grupa stanovništva. Analiziranje klasterovanja ovih faktora rizika u populaciji i njihovog kombinovanog delovanja zahteva sveobuhvatniji pristup, koji podrazumeva i analiziranje uticaja pridruženih faktora kao što su pol, starost ispitanika, sredina u kojoj žive (urbano/ruralno), nivo obrazovanja, bračni status, materijalno stanje.

Nekoliko istraživanja rađenih u različitim populacijama je potvrdilo klasterovanje različitih kombinacija bihejvioralnih faktora rizika i njihove specifične sociodemografske karakteristike [58,96,198,223,224]. Naše istraživanje je sprovedeno na velikom reprezentativnom uzorku odraslog stanovništva Srbije i predstavlja prvu studiju koja se bavi ispitivanjem bihejvioralnih faktora rizika u populaciji, primenom klaster analize. Dobijeni klasteri su karakteristični ne samo u odnosu na bihejvioralne faktore rizika, već imaju i specifične socio-demografske karakteristike. Rezultati naše studije ukazuju da odrasli ispitanici koji žive sami, nižeg su obrazovanja, lošijeg materijalnog stanja i žive u Vojvodini imaju veću šansu da pripadaju klasteru visokog rizika. Dobijeni rezultati su u skladu sa drugim studijama koje ukazuju da su višestruki faktori rizika više zastupljeni među manje obrazovanima [197], ispitanicima koji nisu u braku/vanbračnoj zajednici i lošijeg su materijalnog stanja [198]. Rezultati nekoliko studija ukazuju da je prevalencija višestrukih bihejvioralnih faktora rizika veća kod muškaraca, mlađih odraslih, ekonomski neaktivnog stanovništva, ispitanika koji nisu u braku, manje obrazovanih i ljudi nižeg socio-ekonomskog statusa [58,198,224]. Za razliku od ekonomski aktivnog stanovništva, koga čine zaposlena i nezaposlena lica, ekonomski neaktivno stanovništvo čine lica koja se školuju, penzioneri, lica sa prihodima od imovine, domaćice i oni koji ne spadaju ni u jednu od prethodno navedenih kategorija. Istraživanjem sprovedenim nad odraslim stanovništvom (20-59 godina) Holandije [224] utvrđeno je da su višestruki bihejvioralni faktori rizika češće prisutni među starijima i ispitanicima nižeg nivoa obrazovanja.

Socijalno-ekonomski status (SES) ima veliki uticaj na zdravlje i zdravstveno ponašanje pojedinaca [152]. Zajedničko za sva četiri dobijena klastera rizičnih ponašanja je da je veća verovatnoća da su ispitanici u ovim klasterima nižeg nivoa obrazovanja i lošijeg materijalnog stanja. Prisustvo višestrukih faktora rizika ('Štetna upotreba alkohola i drugi rizici', 'Nepravilna ishrana i drugi rizici') je češće među ljudima sa nižim SES, što je u skladu sa rezultatima drugih studija [65,227,230]. Rezultati naše studije su otkrili da je lošije materijalno stanje povezano sa prekomernim unosom

alkohola, što je potvrđeno i u drugim studijama [81,154]. Nasuprot našim rezultatima, u nekim evropskim zemljama (Grčka, Španija, Portugalija, Poljska, Mađarska), ispitanici sa nižim SES unose više povrća i voća nego ispitanici sa višim SES [213,242]. Ove zemlje takođe imaju najveću stopu potrošnje domaće hrane [213]. Dobijeni rezultati naše analize ukazuju da postoji snažna povezanost između niskog SES i nedovoljne fizičke aktivnosti, što je u skladu sa drugim studijama [48,182,224]. Dizajn korišćene studije preseka onemogućava da utvrdimo da li je usvajanje nezdravih stilova života uzrok lošijeg materijalnog stanja, ili je usvajanje nezdravih stilova života posledica frustracije zbog lošijeg materijalnog stanja [230].

Rezultati ovog istraživanja ukazuju da je nivo obrazovanja negativno povezano sa verovatnoćom za razvoj višestrukih bihejvioralnih faktora rizika kod odraslog stanovništva. Dobijeni rezultati pokazuju da su ljudi sa nižim obrazovanjem imaju veću šansu da su pušači nego što je to slučaj sa osobama višeg obrazovanja, što je potvrđeno i u drugim istraživanjima [57]. Jedno od mogućih objašnjenja je da su odrasli sa nižim nivoom obrazovanja uglavnom lošijeg materijalnog stanja, ili se bave napornim fizičkim poslom i imaju zastupljena zdravstveno rizična ponašanja, to jest češće puše, nezdravo se hrane i manje su fizički aktivni [69,102,222,224]. *Pronk* i saradnici su utvrdili da odrasli sa visokim obrazovanjem imaju za 65% veću šansu da se pridržavaju zdravih životnih navika u odnosu na odrasle koji nemaju visoko obrazovanje [199]. Za pripadnike klastera, 'Nepravilna ishrana i druge rizične navike' važi da je skoro tri puta veća šansa da su nižeg obrazovanja u odnosu na ispitanike sa završenom višom, ili visokom školom. Dobijeni rezultat možemo objasniti i time da su ispitanici sa nižim obrazovanjem uglavnom lošijeg materijalnog stanja, pa je izbor namirnica u ishrani uglavnom uslovljen cenom, a motivisanost zdravljem ima niži prioritet. Nedovoljna fizička aktivnost (u slobodno vreme) kod odraslih u Srbiji je prisutnija kod niže obrazovanih ispitanika, što je u skladu sa rezultatima drugih istraživanja [17,48,182]. Jedno od mogućih objašnjenja je da ispitanici sa višim obrazovanjem imaju razvijeniju svest o značaju fizičke aktivnosti u očuvanju zdravlja, a samim tim i pozitivnih efekata koje ima fizička aktivnost u slobodnom vremenu [164]. Istraživanje *Stephens* i saradnika [234] pokazuje da u Australiji, Kanadi i SAD, odrasli sa visokim obrazovanjem imaju oko 1.5 do 3 puta veću šansu da se bave fizičkim aktivnostima u slobodno vreme u odnosu na odrasle ispitanike sa nižim obrazovanjem. Fizička aktivnost u slobodno vreme može biti potcenjena, posebno kod muškaraca sa nižim nivoom obrazovanja, jer se oni verovatno bave fizički napornijim poslovima u odnosu na ispitanike sa višim nivoom obrazovanja. U populacionom istraživanju koje su sproveli *Laaksonen* i saradnici [151], povezanost nižeg nivoa obrazovanja sa klasterovanjem višestrukih bihejvioralnih faktora rizika je snažnija kod muškaraca nego kod žena. Iako odrasli sa višim nivoom obrazovanja imaju i više znanja koje omogućava pojedincu da donosi zdrave i

kvalitetne izbore i da uključi zdrave navike u koherentan način života, dajući mu osećaj kontrole nad svojim zdravljem [268], ekonomske prepreke mogu ograničiti ponašanje u ovom aspektu [195].

Rezultati našeg istraživanja su pokazala da ispitanici koji nisu u bračnoj/vanbračnoj zajednici imaju veću verovatnoću za razvoj višestrukih faktora rizika ('Štetna upotreba alkohola i drugi rizici', 'Nepravilna ishrana i drugi rizici') nego ispitanci koji su u braku, što je u skladu sa rezultatima drugih istraživanja [224]. Moguće objašnjenje za ovu povezanost može da bude to da bračni status ima protektivni faktor na zdravstveno stanje kroz socijalnu i ekonomsku podršku među bračnim partnerima.

Suprotno očekivanjima, nismo dobili povezanost tipa naselja i klastera 'Fizički neaktivni'. Rezultati drugih studija ukazuju da se stanovnici urbanog područja više bave fizičkim aktivnostima u slobodnom vremenu u odnosu na stanovnike ruralnog područja [128]. Dobijene razlike autori objašnjavaju nivoom obrazovanja i/ili fizičkom aktivnosti na poslu. Stanovnici ruralne sredine uglavnom imaju niži nivo obrazovanja i bave se fizički težim poslovima u odnosu na stanovnike urbane sredine, a oba ova faktora su povezana sa fizičkom aktivnosti u slobodnom vremenu. Rezultati naše studije ukazuju da kombinacija pušenja i fizičke neaktivnosti ('Pušači') ima veću šansu prisustva među odraslima koji žive u gradskom području. Iako se ovi rezultati razlikuju od nekih objavljenih studija [7], u skladu su sa rezultatima *Bauera* i saradnika [35]. Veća je šansa da je kombinacija nepravilne ishrane, pušenja i nedovoljne fizičke aktivnosti ('Nezdrava ishrana i drugi rizici') zastupljena među ispitanicima koji žive u gradu nego ispitanicima koji žive u seoskoj sredini. Moguće objašnjenje za ovaj efekat tipa naselja može ležati u činjenici da ljudi u ruralnom području imaju najvišu stopu potrošnje domaćeg voća i povrća. Drugo moguće objašnjenje je da se za razliku od odraslog seoskog stanovništva, odrasli iz gradske sredine češće bave zahtevnom vrstom posla i angažovani su u više društvenih aktivnosti, kao što su ručkovi, večere sa poslovnim partnerima, ili prijateljima. U takvim situacijama, teško je napraviti najzdraviji izbor hrane. Međutim, moguće je i da su dobijeni rezultati uticaja sredine u kojoj ispitanici žive (urbano/ruralno) posledica definicije pojmova 'urbano' i 'ruralno', tako da razlika između njih postaje manje jasna i svedena je na minimum povećanjem migracija između tih sredina.

Odrasli iz Vojvodine imaju značajno veću šansu za razvoj višestrukih bihejvioralnih rizičnih faktora ('Štetna upotreba alkohola i drugi rizici', 'Nezdrava ishrana i drugi rizici) u odnosu na odraslo stanovništvo Centralne Srbije. Vojvodina je oblast u kojoj ljudi oduvek imaju naviku da jedu dosta hrane, jer se bave poljoprivredom i teškim fizičkim poslovima. Prevalenca gojaznosti je veći u ovoj oblasti u odnosu na druge delove Srbije [106].

Ova studija je pokazala da starost ima značajnu i nezavisnu ulogu u izdvajanju pet klastera bihejvioralnih faktora rizika. Skoro tri puta je veća šansa da su fizički neaktivni sredovečni odrasli ispitanici (35-54) ili stariji odrasli (55 ili više godina) (OR = 2.98). Slični rezultati su dobijeni u drugim studijama [1, 160, 170, 224, 227]. Manja je šansa da su pušači u grupi starijih odraslih, što je u skladu sa drugim studijama [53]. Stariji ljudi, posebno oni koji se suočavaju sa dubokim pogoršanjem zdravlja, imaju viši nivo svesti o sopstvenom zdravstvenom stanju, kao i rizicima pušenja [57]. Studija koju su sproveli *Pronk* i saradnici [199] ukazuje da starost, nivo obrazovanja i prisustvo hroničnih bolesti kod odraslih značajno utiču na usvajanje i negovanje zdravih životnih navika. Kod odraslih osoba, starosti od 50 do 64 godine veća je verovatnoća da će voditi zdrav stil života u poređenju sa osobama starosti od 18 do 49 godina. Odrasli sa visokim obrazovanjem imaju za 65% veću šansu da se pridržavaju zdravih životnih navika u odnosu na odrasle koji nemaju visoko obrazovanje. I na kraju, odsustvo hronične bolesti je povezano sa 90% većom šansom za pridržavanje višestrukih zdravih životnih navika u poređenju sa onim odraslim osobama koje imaju prisutne hronične bolesti. Ova saznanja su potkrepila rezultate *Fine* i saradnika [90] koji su utvrdili sličan uticaj starosti, obrazovanja i prisustva srčanog oboljenja i/ili dijabetesa na prisustvo tri ili četiri faktora rizika.

Obrasci rizičnog ponašanja i njihovo grupisanje se razlikuje u odnosu na pol. Veća je šansa da su žene nedovoljno fizički aktivne u odnosu na muškarce, što je u skladu sa rezultaima drugih studija [48,160,182]. Razlika između muškaraca i žena se povećava sa intenzitetom fizičke aktivnosti i najveća razlika je prisutna kod intenzivne fizičke aktivnosti [137]. Suprotno očekivanjima, veća je šansa da su žene pripadnici klastera 'Pušači'. Postoje, međutim, moguća objašnjenja za ovakav ishod. Odrasli u ovom klasteru su pušači bez drugih rizičnih ponašanja (41.2%), ili pušači koji su nedovoljno fizički fizički aktivni (58.8%). Muškarci koji puše obično imaju udružene i druge bihejvioralne faktore rizika. Veća je šansa da su ispitanici sa više rizičnih navika: pušenje cigareta, prekomeran unos alkohola i nizak nivo fizičke aktivnosti ('Prekomeran unos alkohola i drugi rizici') muškarci, što je u skladu sa rezultatima drugih studija [160]. Razlog ovakvog obrasca ponašanja leži u činjenici da je prekomeran unos alkohola, ili kombinacija prekomernog unosa alkohola i pušenja kod žena manje društveno prihvatljivo ponašanje u odnosu na isto takvo ponašanje kod muškaraca. Drugo objašnjenje bi moglo biti i to da su žene, koje prirodno imaju ulogu 'čuvara' porodice, svesnije negativnog uticaja štetne upotrebe alkohola na zdravlje. Kod žena postoji tendencija prekomernog unosa alkohola usled stresa prouzrokovanog višestrukim odgovornostima, kao što je koordinacija posla i porodice, dok se kod muškaraca ovi razlozi razvijaju iz ličnih izazova vezanih za posao i pritiska da budu uspešni. Rezultati *Bloomfield*-a i saradnika pokazuju da žene visokog obrazovanja i muškarci nižeg obrazovanja imaju

veću šansu teškog opijanja pri čemu te razlike variraju od zemlje do zemlje [41]. Istraživači koji su se bavili polnim razlikama u ovakvom kontekstu istraživanja, ukazuju na činjenicu da i socioekonomski i kulturni faktori mogu da utiču na ovakvo ponašanje [95].

Rezultati koji su dobijeni ovim istraživanjem omogućavaju utvrđivanje visokorizičnih grupa stanovništva čije je zdravlje ugroženo istovremenim prisustvom vodećih faktora rizika za MNB. Na ovaj način se dobijaju potpuniji podaci neophodni za razvoj strategija i programa usmerenih na celo stanovništvo, posebno na visokorizične kategorije stanovništva. Efekti ovakvog programa su dokazani. U Singapuru je sprovođenjem nacionalnog programa zdravih stilova života smanjena prevalenca pušenja kod muškaraca sa 34% na 27% dok je istovremeno povećan procenat fizički aktivnog stanovništva, sa 14% na 17%. U analiziranom periodu (1991-1999. godina), incidencija infarkta miokarda (koja je standardizovana u odnosu na starost) je smanjena sa 98.2 na 83 iskazano na 100 000 stanovnika i mortalitet od koronarne srčane bolesti se smanjio sa 60.8 na 47.2 iskazano na 100 000 stanovnika [252]. Finska je zemlja koja je imala najveće stope mortaliteta od kardiovaskularnih bolesti. Prepoznavajući faktore rizika udruženih sa KVB i sagledavanjem ugroženosti stanovništva tim rizicima omogućilo je razvoj nacionalnog interventnog programa. Primenom ovog integrisanog programa prevencije MNB došlo je do smanjenja prevalencije pušenja, nivoa holesterola i hipertenzije, što se odrazilo na morbiditet i mortalitet od KVB stanovništva Finske koji je smanjen za 75%. Na osnovu sveobuhvatne analize sprovedene 2015. godine [139], a korišćenjem 220 studija, odnosno baza podataka sakupljenih između 1990 i 2013. godine, utvrđeno je da su u većini zemalja, uključujući SAD, Kanadu, Novi Zeland, Japan, Veliku Britaniju, Belgiju, Francusku, Čile, Norvešku i Meksiko, sprovedeni interventni programi usmereni pre svega na navike u ishrani i fizičku aktivnost kod odraslog stanovništva. Fokusiranje na ove stilove života je proisteklo iz činjenice da su oni (energetski unos i potrošnja) glavni pokretači rastućeg problema gojaznosti, koji je dostigao razmere globalne epidemije [262].

8.3 OGRANIČENJA ISTRAŽIVANJA

Prilikom interpretacije rezultata klasterovanja veoma je važno ukazati na postojeća ograničenja istraživanja. Važno ograničenje istraživanja predstavlja korišćenje podataka iz studije preseka, jer se efekti promene stila života (promena navika u ishrani, povećanje fizičke aktivnosti i drugo) mogu utvrditi samo primenom studije praćenja. Korišćenje podataka iz studije preseka onemogućava da se precizno utvrdi kada se grupisanje bihejvioralnih faktora rizika dešava tokom vremena i da li se nečija lična/porodična imovina (materijalno stanje), ili mesto življenja (urbano/ostalo; region) promenilo od tada, bilo zbog vertikalne društvene pokretljivosti (poboljšanje ili pogoršanje imovno stanje) ili migracija unutar zemlje. Za razliku od nivoa obrazovanja pojedinca, koje možemo smatrati fiksnom kategorijom koja se još može i povećati tokom vremena, materijalno stanje je varijabilna kategorija, kako zbog visokog nivoa socijalnog stresa, tako i zbog visoke stope nezaposlenosti tokom poslednjih godina [76].

Svi podaci o bihejvioralnim faktorima rizika su dobijeni na osnovu samoprocene ispitanika, što kao posledicu ima dobro poznate nedostatke. Kao kod većine studija ovog tipa, problem predstavlja to što postoji tendencija da ispitanici daju odgovore primerene društvenim vrednostima (društveno prihvatljive), a ovo se naročito odnosi na odgovore u vezi sa konzumiranjem alkohola.

Navike u ishrani i fizička aktivnost su kompleksna, multi-dimenzionalna obeležja, pa ih je stoga vema teško proceniti. Mnoge objavljene studije u kojima su analizirane navike u ishrani su koristile složenije instrumente za procenu ove navike, kao što je na primer Mediteranski skor ishrane [142]. Deo upitnika koji se odnosi na ishranu nije mogao da obezbedi sveobuhvatnu listu namirnica. Upitnik je obuhvatio pitanja koja se odnose na frekvenciju uzimanja određenih namirnica tokom prethodne nedelje (ponuđeni odgovori: *nijednom*, *1-2 puta nedeljno*, *3-5 puta*, *6-7 puta nedeljno*), ali ne i broj unapred definisanih porcija dnevno. Procena frekvencije unosa određenih namirnica je subjektivna, što dovodi do varijabilnosti u proceni frekvencije određene hrane. Zbog svih navedenih ograničenja, to jest nepotpunih informacija u vezi sa unosom određenih namirnica, nismo bili u mogućnosti da kreiramo neprekidnu promenljivu, to jest numerički skor koji opisuje dobre (unos voća, povrća, salate, ribe i drugo) ili loše navike u ishrani (unos slatkiša, grickalica, „brze hrane“ i drugo). Za opisivanje navika u ishrani kod odraslog stanovništva, koristili smo podatke o svakodnevnom unosu voća i povrća (da/ne). Preporuke o dnevnom unosu pet porcija voća i povrća [162] nismo mogli koristiti, jer ograničenje upitnika predstavlja to što nema podataka o broju unetih porcija voća/povrća u toku dana.

Iz praktičnih razloga, većina epidemioloških studija prvenstveno koristi upitnike, umesto objektivnih merenja za utvrđivanje nivoa fizičke aktivnosti. Međutim, fizička aktivnost je kompleksna i predstavlja bihejvioralno obeležje, pa mogućnost epidemioloških studija da utvrde povezanost između nedovoljne fizičke aktivnosti i hroničnih oboljenja veoma zavisi od validnosti upitnika za samoprocenu fizičke aktivnosti [249]. Često je teško interpretirati i porediti rezultate različitih studija procene fizičke aktivnosti zbog razlike u metodološkom pristupu, razlika u analiziranju podataka i interpretaciji, kao i adaptacije različitih definicija poželjnog nivoa fizičke aktivnosti. Samoprocena nivoa fizičke aktivnosti takođe predstavlja izvor potencijalne pristrasnosti. Postoji više ograničenja koja se odnose na definisanje nivoa fizičke aktivnosti kod odraslog stanovništva. Klasifikacija fizičke aktivnosti koja je korišćena u Istraživanju zdravlja stanovništva Srbije, a predviđena upitnikom *International Physical activity Questionnaire-IPAQ*, ne razmatra nivo fizičke aktivnosti na poslu, već samo one aktivnosti koje se odvijaju u toku slobodnog vremena. Takođe, deo upitnika koji se odnosi na korišćenje slobodnog vremena i fizičku aktivnost je predviđen za mlade i sredovečne odrasle osobe, starosti od 15 do 69 godina i nije prilagođen starijoj populaciji (70 i više godina), to jest pitanja nisu predvidela one aktivnosti koje su poželjne kod starijeg dela stanovništva, kao što su: lagana šetnja, rad u bašti, ples, joga, tai-chi, i slično. *Donahue* i saradnici su adaptirali upitnik o fizičkoj aktivnosti u odnosu na starost i klasifikovali ispitanike u tercile fizičke aktivnosti, posebno za ispitanike starosti od 45 do 64 godine, odnosno ispitanike starosti 65 i više godina [79]. Neke prethodne studije su pokazale da ispitanici bolje pamte detalje vezane za aktivnosti u domaćinstvu (umerenog intenziteta), koje su dobro definisane i rutinski se izvode, kao što su pranje veša, kuvanje, pranje posuđa ili rad u bašti i imaju bolje merne karakteristike u poređenju sa više netipičnim aktivnostima sličnog intenziteta, kao što je hodanje. Iz tog razloga bi, kao što predlažu *Cust* i saradnici [67], bilo poželjno da se unesu manje izmene u EPIC upitnik što može poboljšati njegove karakteristike merenja i bolje razlikovati ljude sa sedentarnim načinom života od onih koji su umereno neaktivni. Predložene mere poboljšanja uključuju razmatranje učestalosti i trajanja aktivnosti na poslu, zatim podelu kućne aktivnosti u dve ili više kategorija (npr. aktivna briga o deci, kuvanje, čišćenje), a sve u cilju preciznije procene intenziteta nivoa fizičke aktivnosti.

Većina objavljenih studija o klasterovanju faktora rizika se uglavnom bavi klasterovanjem bioloških, a ne bihejvioralnih faktora rizika. Osim toga, u većini istraživanja ciljna populacija su deca i adolescenti. Teško je upoređivati rezultate ovih studija, jer su one fokusirane na različite kombinacije bihejvioralnih faktora rizika, koriste različite metodologije za definisanje bihejvioralnih faktora rizika, analiziraju različite populacione grupe i koriste različite analitičke tehnike. Mogući razlozi za varijabilnost

dobijenih rezultata klasterovanja u različitim studijama mogu biti i društveno prihvatljivi odgovori, tip pitanja, kontekst i rang mogućih odgovora i drugo. Dobijeni klasteri se mogu veoma razlikovati u zavisnosti od uključivanja više promenljivih ili različitih promenljivih, načina na koji su bihejvioralni faktori rizika definisani, različitih algoritama klasterovanja. Dok dihotomizacija faktora rizika i izbor *cut-off* vrednosti dozvoljava usklađivanje definicija faktora rizika sa postojećim nacionalnim preporukama, istovremeno može ograničiti uopštavanje rezultata na kontekst i populaciju koje se razlikuju od ovog konteksta. Studije koje se bave ispitivanjem povezanosti višestrukih rizičnih zdravstvenih ponašanja karakteriše korišćenje veoma različitih analitičkih tehnika. Različite metode klasifikacije ponašanja i klasifikacije subjekata u vezi zdravstveno rizičnih ponašanja su usvojene u literaturi, uključujući i faktorsku analizu (obično se koristi Analiza glavnih komponenti), klaster analizu i druge metode. Prednost klaster analize u profilisanju zdravstveno rizičnih navika je da klasifikuje pojedince, a ne bihejvioralne faktore, kao što je slučaj u faktorskoj analizi.

9. ZAKLJUČCI

- Za velike skupove podataka sa kategorijalnim obeležjima korišćena je baza *Mushrooms*, sa već definisanom klasifikacijom. Upoređivanjem različitih algoritama klasterovanja (k-*modusi*, k-*sredine*, TSCA, hijerahijski: AL, SL, CL, Ward) na celom skupu podataka, utvrđeni su najbolji rezultati za TSCA i Ward-ov algoritam klasterovanja. Najbolje klastersko rešenje je određeno na osnovu kriterijuma eksterne validnosti-tačnosti, to jest poklapanja rezultata klasterovanja sa već postojećom klasifikacijom na celom skupu. Nedostatak Ward-ovog algoritma za velike skupove podataka predstavlja njegova vremenska složenost.

- U cilju utvrđivanja da li se smanjenjem ulaznih podataka (obima skupa), a samim tim i vremenske složenosti algoritma, ne narušava struktura dobijenih klastera na celom skupu, Ward-ov algoritam je dalje primenjen na kategorijalne podatke, korišćenjem prostih slučajnih uzoraka (veličine $0.01m, 0.03m, 0.05m, 0.1m, 0.3m$, gde je m obim osnovnog skupa) iz baze *Mushrooms*. Rezultati dobijeni na uzorcima su pokazali visoko slaganje sa postojećom klasifikacijom na celom skupu, to jest nije utvrđena značajna razlika u prosečnoj tačnosti na uzorcima i osnovnom skupu ($p<0.001$). Izuzetak čine uzorci veličine $0.01m$.

- Za velike skupove podataka sa kombinovanim obeležjima korišćena je baza *Istraživanje zdravlja stanovništva Srbije*. Primenjen je TSCA algoritam na celom skupu, pri čemu je korišćenjem Bajesovog informacionog kriterijuma (BIC) utvrđen optimalan broj klastera $k=5$. Kao mera interne validnosti korišćen je Silhouette indeks.

- Daljim korišćenjem postupka višestrukog izvlačenja prostih slučajnih uzoraka (iz baze *Istraživanje zdravlja Srbije*) i primenom TSCA algoritma za kombinovana obeležja na uzorcima, utvrđen je takođe najoptimalniji broj klastera za $k=5$ (što je i optimalan broj klastera dobijen na celom skupu). Najveća tačnost, to jest slaganje rezultata klasterovanja na uzorcima sa dobijenim klasterima na celom skupu je takođe potvrđena za ovaj optimalni broj klastera. Izuzetak čine uzorci veličine $0.01m$.

- Postupak višestrukog izvlačenja ($i_{\max}$ broj izvlačenja) prostih slučajnih uzoraka veličine $t_u m$ (koeficijent $t_u \in (0,1)$, m obim osnovnog skupa) i primena klaster algoritma na ovim uzorcima daje jednako dobre rezultate (nije narušena struktura klastera) kao i algoritam primenjen na celom skupu. Predloženi modifikovani postupak klasterovanja velikih skupova podataka se sastoji iz sledećih faza:
  1. Klasterovanje na prostim slučajnim uzorcima određene kardinalnosti
  2. Primenom odgovarajućeg kriterijuma validnosti dobija se najbolje klastersko rešenje na i'-tom uzorku ($i' \le i_{\max}$)
  3. Dobijeni centri klastera iz i'-tog uzorka služe za klasterovanje ostatka skupa.

- Analiza podataka je sprovedena na velikom reprezentativnom uzorku odraslog stanovništva Srbije i predstavlja prvu studiju koja se bavi ispitivanjem bihejvioralnih faktora rizika u populaciji, primenom klaster analize.

- Primenom dvostepenog klaster algoritma izdvajaju se jasno odvojeni klasteri u populaciji odraslog stanovništva Srbije, sa karaktersističnim kombinacijama bihejvioralnih faktora rizika: *Bez rizičnih faktora*, *Štetna upotreba alkohola i druge rizične navike*, *Nepravilna ishrana i druge rizične navike*, *Nedovoljna fizička aktivnost*, *Pušači.*

- Jedan od načina procene validnosti dobijenog klasterskog rešenja obuhvata testiranje razlika između klastera na nekim relevantnim eksternim promenljivama, koje nisu korišćene u postupku klasterovanja. Testirana je razlika između klastera bihejvioralnih faktora rizika u odnosu na sociodemografske karakteristike (eksterne promenljive). U prvom koraku je primenjena univarijantna analiza, a zatim multinomni logistički regresioni model sa zavisnom promenljivom pripadnost klasteru i sociodemografskim karakteristikama kao nezavisnim promenljivama.

- Primenom multinomnog logističkog regresionog modela, zaključujemo da ispitanici koji nisu u braku, lošijeg su materijalnog stanja, nižeg obrazovanja i žive u Vojvodini imaju veću šansu za prisustvo višestrukih bihejvioralnih faktora rizika.

- Informacije o sociodemografskim karakteristikama dobijenih klastera su korisne u planiranju budućih preventivnih strategija, jer ukazuju na ciljne kategorije stanovništva koje su prioritetne u planiranju i sprovođenju specifičnih preventivnih programa i intervencija u budućnosti.

- Klaster analiza identifikuje fiksirane i stabilne klastere u ispitivanom skupu podataka, omogućavajući davanje preporuka za modifikaciju postojećih navika povezanih sa stilom života kod odraslog stanovništva, kao i dizajniranje budućih anketnih upitnika za prikupljanje podataka

- Navike u ishrani i nivo fizičke aktivnosti su multidimenzionalna obeležja, te je za njihovu precizniju ocenu potrebna modifikacija postojećeg upitnika korišćenog u istraživanju, kao i prilagođavanje starijoj populaciji s obzirom na visoku prevalencu masovnih nezaraznih bolesti u ovoj kategoriji stanovništva.

- Istraživanje daje uvid u udruživanje bihejvioralnih faktora rizika i kovarijate (pridružene faktore) za bihejvioralne faktore rizika kod pojedinaca, a rezultati klaster analize obezbeđuju više sveobuhvatnih informacija o zdravstvenom stanju stanovništva nego pojedinačni faktori rizika.

10. LITERATURA

1 Adabonyan I, Loustalot F, Kruger J, Carlson SA, Fulton JE. Prevalence of highly active adults–Behavioral risk factor surveillance system, 2007. Prev Med 2010; 51(2): 139-143.

2 Aguiar P, Neto D, Lambaz R, Chick J, Ferrinho P. Prognostic factors during outpatient treatment for alcohol dependence: cohort study with 6 months of treatment follow-up. Alcohol Alcoholism 2012; 47: 702-10.

3 Ahirwar R. A Novel K means clustering algorithm for large datasets based on divide and conquer technique. Int J Comput Sci Inf Techn 2014; 5 (1):301-305.

4 Ahmad A, Dey L. A k-means type clustering algorithm for mixed numeric and categorical datasets. Data Knowl Eng 2007; 63 (2): 503-527.

5 Ahmad A, Dey L. A k-means type clustering algorithm for subspace clustering of mixes numeric and categorical datasets. Pattern Recognit Lett 2011;32: 1062-1069.

6 Akaike H. A new look at the statistical model identification. IEEE Trans Autom Control. 1974, 19(6): 716-723.

7 Alam AY, Iqbal A, Mohamud KB, Laporte RE, Ahmed A, Nishtar S. Investigating socio-economic-demographic determinants of tobacco use in Rawalpindi, Pakistan. BMC Public Health 2008; 8(1): 50.

8 Aldana SG, Greenlaw RL, Diehl HA, Salberg A, Merrill RM, Ohmine S, Thomas C. Effects of an intensive diet and physical activity modification program on the health risks of adults. J Amer Diet Ass 2005; 105:371-381.

9 Al-Sultana KS, Khan MM. Computational experience on four algorithms fr the hard clustering problem. Pattern Recognit Lett 1976; 17(3): 295-308.

10 Anderberg MR. Cluster Analysis for Applications. New York: Academic, 1973

11 Andramonov MY, Rubinov AM, Glover BM. Cutting Angle Methods in Global Optimization. Appl Math Lett 1999; 12: 95-100.

12 Andrews NO, Fox EA. Clustering for data reduction: a divide and conquer approach. Technical Report TR-07-36, Computer Science, Virginia Tech, 2007

13 Andritsos P, Tsaparas P, Miller RJ, Sevcik KC. Limbo: scalable clustering of categorical data. In: Bertino E, Christodoulakis S, Plexousakis D, Vassilis C, Koubarakis M, Böhm K, Ferrari E, editors. Advances in Database Technology. 9th International Conference on EDBT; 2004 March 14-18; Heraklion, Crete, Greece. Berlin: Springer Berlin Heidelberg, 2004. p. 123-146.

14 Ankerst M, Breunig MM, Kriegel HP, Sander J (1999). OPTICS: Ordering Points To Identify the Clustering Structure. In: SIGMOD Record - web edition. ACM SIGMOD International Conference on Management of Data; 1999 June 1-3; Philadelphia, Pennsylvania, USA. ACM Press 1999; 28 (2): 49–60.

15 Arbelaitz O, Gurrutxaga I, Muguerza J, Perez JM, Perona I. An extensive comparative study of cluster validity indices. Pattern Recognit 2013; 46:243-256.

16 Atanasković-Marković Z, Bjegović V, Janković S, Kocev N, Laaser U, Marinković J, et al. Opterećenje bolestima i povredama u Srbiji. Beograd: Ministarstvo zdravlja Republike Srbije; 2003.
17 Azevedo MR, Araújo CLP, Reichert FF, Siqueira FV, Da Silva MC, Hallal PC. Gender differences in leisure-time physical activity. Int J Public Health. 2007; 52(1): 8–15.
18 Bacher J, Wenzig K, Vogler M: SPSS TwoStep Cluster-a first evaluation. In Work and discussion paper. Erlangen-Nuremberg, Germany: Department of Sociology, Social Science Institute, Friedrich-Alexander-University; 2004:1-30.
19 Bagirov A, Rubinov AM, Yearwood J. A global optimization approach to classification, Optim Eng. 2002; 3:129-55.
20 Bagirov AM, Karasözen B, Sezer M. Discrete gradient method: a derivative free method for nonsmooth optimization, J Optimiz Theory App. 2008;137: 317-34.
21 Bagirov AM, Rubinov AM, Soukhoroukova N, Yearwood J. Unsupervised and supervised data classification via non-smooth optimization. Socíedad de Estadistica e Invastigación Operativa Top. 2003; 11(1): 1-93.
22 Bagirov AM, Rubinov AM, Zhang J. Local optimization method with global multidimensional search. J Glob Optim. 2005; 32 (2): 161-79.
23 Bagirov AM, Rubinov AM. Cutting angle method and a local search. J Glob Optim. 2003; 27 (2-3): 193-213.
24 Bagirov AM, Rubinov AM. Modified versions of the cutting angle method. In: Hadjisavvas N, Pardalos PM, editors. Advance in Convex Analysis and Global Optimization. Dodrecht: Kluwer Academic Publishers; 2001. p. 245-68.
25 Bagirov AM, Ugon J. An algorithm for minimizing clustering functions. Optimization. 2005; 54 (4-5): 351-68.
26 Bagirov AM, Yearwood J. A new nonsmooth optimization algorithm for minimum sum-of squares clustering problems. Eur J Oper Res. 2006; 170: 578-96.
27 Bagirov AM. Minimization methods for one class of nonsmooth functions and calculation of semi-equilibrium prices: In Eberhart A, Hill R, Ralph D, Glover BM, editors. Progress in Optimization. Dordrecht: Kluwer Academic Publishers; 1999. p.147-75.
28 Bagirov AM. Modified global k-means algorithm for minimum sum of squares clustering problems. Pattern Recognit. 2008; 41:3192-99.
29 Bagirov AM. Numerical methods for minimizing quasidifferentiable functions: a survey and comparison. In: Demyanov VF, Rubinov AM, editors. Quasidifferentiability and Related Topics. Dordrecht: Kluwer Academic Publishers; 2000. p.33-71.
30 Bai L, Liang J, Dang C, Cao F. A cluster centers initialization method for clustering categorical data. Expert Syst Appl. 2012; 39: 8022-29.
31 Bai L, Liang J, Dang C, Cao F. A novel attribute weighting algorithm for clustering high-dimensional categorical data. Pattern Recognit. 2011; 44:2843-61.

32 Bai L, Liang J, Dang C. An initialization method to simultaneously find initial centers and the number of clusters for clustering categorical data. Knowl-Based Syst. 2011;24:785-95.
33 Baker FB, Hubert LJ. Measuring the power of hierarhical cluster analysis. J Am Stat Assoc. 1975;70:31-38.
34 Ball GH, Ball DJ. Clustering technique for summarizing multivariate data. Behav Sci. 1967;12:153-55.
35 Bauer T, Göhlmann S, Sinning M. Gender differences in smoking behavior. Health Econ. 2007;16:895-909.
36 Bezdek JC. A converegence theorem for the fuzzy ISODATA clustering algorithms. IEEE Trans Pattern Anal Mach Intell. 1980;2:1-8.
37 Bezdek, J. C., Ehrlich, R., & Full, W. FCM: The fuzzy c-means clustering algorithm. Comput Geosci. 1984;10(2):191-203.
38 Bilano V, Gilmour S, Moffiet T, d'Espaignet ET, Stevens GA, Commar A, Tuyl F, Shibuya K. Global trends and projections for tobacco use, 1990–2025: an analysis of smoking indicators from the WHO Comprehensive Information Systems for Tobacco Control. Lancet. 2015; 385(9972):966-976.
39 Billson H, Pryer JA, Nichols R. Variation in fruit and vegetable consumption among adults in Britain. An analysis from the dietary and nutritional survey of British adults. Eur J Clin Nutr. 1999;53(12):946-952.
40 Blashfield RK.The Growth of cluster analysis: Tryon, Ward, And Johnson. Multivar Behav Res. 1980; 15(4): 439-458
41 Bloomfield K, Gmel G, Wilsnack S. Introduction to special issue 'Gender, Culture and Alcohol Problems: A Multi-National Study'. Alcohol Alcoholism. 2006;41 (Suppl 1):i3-7.
Bock HH. Automatische Klassifikation. Göttingen: Vandenhoeck & Ruprecht; 1974.
42 Bonner RE. Cluster analysis. Ann NY Acad Sci. 1966; 28:972–983.
44 Borzanović M, Stožinić S, Borzanović B, Maravić V. Noviji pogledi na prirodu arteroskleroze i značaj. Medicinska revija. 2010; 2(1):35-43.
45 Breslow L, Enstrom JE. Persistence of health habits and their relationship to mortality. Prev Med. 1980;9:469-83.
46 Buhmiler S, Krejić N. A new smoothing quasi-Newton method for nonlinear complementarity problems. J Comput Appl Math. 2008;211:141-155.
47 Burke JV, Lewis AS, Overton ML. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. Siam J Optimiz. 2005;15(3):751-779.
48 Burton, N. W., & Turrell, G. Occupation, hours worked, and leisure-time physical activity. Prev Med. 2000; 31(6):673-681.
49 Calinski T, Harabasz J. A dendrite method for cluster analysis. Commun Stat Theory Methods. 1974;3:1-27.
50 Cao F, Liang J, Bai L. A new initialization method for categorical data clustering. Expert Syst Appl. 2009;36:10223-10228.

51 Cao F, Liang J, Li D, Bai L, Dang C. A dissimilarity measure for the k-modes clustering algorithm. Knowl-Based Syst. 2012; 26:120-12.
52 Cariou V, Verdun S, Diaz E, Qannari EM, Vigneau E. Comparasion of three hypothesis testing approaches for the selection of the appropriate number of cluster of variables. Adv Data Anal Classif. 2009; 3:227-241.
53 Cavelaars AE, Kunst AE, Geurts JJ, Crialesi R, Grötvedt L, Helmert U et al. Educational differences in smoking: international comparison. BMJ. 2000; 320(7242):1102-1107.
54 Centers for Disease Control and Prevention. Alcohol and Public Health. Atlanta, GA: Department of Health and Human Services 2013
55 Chan EZ, Ching WK, Ng MK, Huang JZ. An optimization algorithm for clustering using weighted dissimilarity measures. Pattern Recognit. 2004; 37: 943-952.
56 Chaturvedi A, Foods A, Green PE, Carrol JD. K-modes Clustering. J Classif. 2001; 18:35-55.
57 Cheah YK, Naidu BM. Exploring factors influencing smoking behaviour in Malaysia. Asian Pac J Cancer Prev. 2012;13(4):1125–30.
58 Chiolero A, Wietlisbach V, Ruffieux C, Paccaud F, Cornuz J. Clustering of risk behaviors with cigarette consumption: A population-based survey. Prev Med. 2006; 42: 348–353.
59 ChiuT, Fang D, Chen J, Wang Y, Jeris C. A robust and scalable clustering algorithm for mixed type attributes in large database environment. In: KDD-2001. Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2001 August 26-29; San Francisco, California. New York: ACM 2001. p.263–8.
60 Chiuve SE, McCullough ML, Sacks FM, Rimm EB. Healthy lifestyle factors in the primary prevention of coronary heart disease among men: benefits among users and nonusers of lipid-lowering and antihypertensive medications. Circulation. 2006; 114(2):160-7.
61 Clarke F. Optimization and Nonsmooth Analysis. New York: Wiley; 1983.
62 Colditz GA. Economic costs of obesity and inactivity. Med Sci Sports Exerc. 1999;31:S663-7.
63 Cook SA. The complexity of theorem proving procedures. In: STOC '71. Proceedings of the third annual ACM symposium on Theory of computing; 1971; New York: ACM; 1971. p.151-8.
64 Costa E, Dias CM, Oliveira L, Gonçalves L. Clustering of behavioral risk factors in the Portuguese population: Data from National Health Interview Survey. J Behav Health. 2014; 3(4):205-11.
65 Coups EJ, Gaba A, Orleans CT. Physician Screening for Multiple Behavioral Health Risk Factors. Am J Prev Med 2004; 27(2):34–41.
66 Cover T, Thomas J. Elements of Information Theory. New York: Wiley Intersence; 1991.

67 Cust AE, Smith BJ, Chau J, Hidde P van der Ploeg, Friedenreich CM, Armstrong BK, Bauman A. Validity and repeatability of the EPIC physical activity questionnaire: a validation study using accelerometers as an objective measure. Int J Behav Nutr Phys Act. 2008; 5(1):33.
68 Dallongeville J, Marecaux N, Fruchart JC, Amouyel P. Cigarette smoking is associated with unhealthy patterns of nutrient intake: a meta-analysis. J Nutr. 1998;128:1450–7.
69 Darmon N, Drenowski A. Does social class predict diet quality? Am J Clin Nutr. 2008; 87: 1107-17.
70 De Vries H, Van T Riet J, Spigt M, Metsemakers J, Van Den Akker M, Vermunt JK, Kremers S. Clusters of lifestyle behaviors: results from the Dutch SMILE study. Prev Med. 2008; 46(3): 203–8.
71 Demyanov VF, Rubinov A. Quasidifferential Calculus. Optimization Software. New-York: Inc. Publications Division; 1986.
72 Demyanov VF, Rubinov AM. Constructive Nonsmooth Analysis. Approximation and Optimization 7, Peter Lang, Frankfurt am Main; 1995.
73 Derman EW, Patel DN, Nossel CJ, Schwellnus MP. Healthy lifestyle interventions in general practice. Part 1: An introduction to lifestyle and diseases of lifestyle. S Afr Fam Pract. 2008;50(4):6-12.
74 Dice, LR. Measures of the amount of ecologic association between species. Ecology. 1945; 26(3):297-302.
75 Dimitriadou E, Dolničar S, Weingessel A. An Examination of Indexes for Determining the Number of Clusters in Binary Data Sets. Psychometrika. 2002; 67(1):137-60.
76 Djikanovic B, Marinkovic J, Jankovic J, Vujanac V, Simic S. Gender differences in smoking experience and cessation: do wealth and education matter equally for women and men in Serbia? J Public Health (Oxf) 2011;33(1):31–8.
77 Do Lee C, Folsom AR, Blair SN. Physical activity and stroke risk: a meta-analysis. Stroke. 2003; 34(10):2475–81.
78 Doll R, Peto R, Boreham J, Sutherland I. Mortality in relation to smoking: 50 years' observations on male British doctors. BMJ. 2004; 328(7455): 1519.
79 Donahue PR, Abott RD, Reed DM, Yano K. Physical Activity and Coronary Heart Disease in Middle-Aged and Elderly Men: The Honolulu Heart Program. Am J Public Health. 1988;78(6):683-5.
80 Driver HE, Kroeber AL. Quantitative expression of cultural relationships. University of California Publications in American Archeology and Ethnology. 1932;31(4):211-56.
81 Droomers M, Schrijvers CTM, Stronks K, van de Mheen D, Mackenbach JP. Educational differences in excessive alcohol consumption: the role of psychosocial and material stressors. Prev Med. 1999;29(1):1-10.
82 Dubes R, Jain AK. Clustering techniques: The user's dilemma. Pattern Recognit. 1976; 8(4):247-60.

83 Emmons H, Rai S. Computational Complexity Theory. In: Floudas CA, and Pardalos PM, editors. Encyclopedia of Optimization. Dordrecht: Kluwer Academic Publishers; 2001;1: 310-315.
84 Ester M, Kriegel HP, Sander J, Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise In: Evangelos Simoudis E, Han J, Fayyad U, editors. The Second International Conference on Knowledge Discovery and Data Mining (KDD-96); 1996 August 2–4; Portland, Oregon. Menlo Park, California: The AAAI Press; 1996, p.226-31.
85 Ezzati M, Lopez AD, Rodgers A, Vander Hoorn S, Murray CJ. Selected major risk and global and regional burden of disease. Lancet. 2002;360(9343):1347-60.
86 Ezzati M, Hoorn SV, Rodgers A, Lopez AD, Mathers CD, Murray CJ. Estimates of global and regional potential health gains from reducing multiple major risk factors. Lancet. 2003;362(9380):271-280.
87 Ezzati M, Lopez AD. Estimates of global mortality attributable to smoking in 2000. Lancet 2003;362(9387):847-52.
88 Ferrucci L, Izmirlian G, Leveille S, Phillips CL, Corti MC, Brock DB, Guralnik JM. Smoking, physical activity, and active life expectancy. Am J Epidemiol. 1999;149:645-53.
89 Finch H. Comparasion of Distance Measures in Cluster Analysis with Dichotomous Data. J Data Sci. 2005;3:85-100.
90 Fine LJ, Philogene GS, GramlingR, Coups EJ, Sinha S. Prevalence of multiple chronic disease risk factors. 2001 National Health Interview Survey. Am J Prev Med. 2004;27(2 Suppl):18–24.
91 Fletcher R. Practical methods of Optimization, second ed. Chichester, UK: John Wiley and Sons; 1987.
92 Floudas CA, Counaris CE. A review of recent advances in global optimization. J Glob Optim. 2009;45:3-38.
93 Floudas CA, Pardalos PM. Recent advances in global optimization. Princeton University Press; 2014.
94 Ford ES, Bergmann MM, Boeing H, Li C, Capewell S. Healthy lifestyle behaviors and all-cause mortality among adults in the United States. Prev Med 2012, 55(1): 23-7.
95 Fornari C, Donfrancesco C, Riva MA, Palmieri L, Panico S, Vanuzzo D, Cesana G. Social status and cardiovascular disease: a Mediterranean case. Results from the Italian Progetto CUORE cohort study. BMC Public Health. 2010; 10(1):574.
96 French S, Rosenberg M, Knuiman M. The clustering of health risk behaviors in a Western Australian adult population. Health Promot J Austr. 2008;19(3):203-9.
97 Gan G, Ma C, Wu J. Data Clustering: Theory, Algorithms, and Applications. Philadelphia: SIAM; 2007.
98 Gan G, Yang Z, Wu J. A Genetic k-modes algorithm for clustering for categorical data. In: Xue L, Shuliang W, Zhao Yang D, editors. Advanced Data Mining and Applications. First International Conference, ADMA; 2005 July 22-

24; Wuhan, China. Berlin: Springer Heidelberg; 2005. p.195-202.

99 Garey MR, Johnson DS. Computers and intractability. New York:Freeman; 1979.
100 Garrett NA, Brasure M, Schmitz KH, Schultz MM, Huber MR. Physical Inactivity. Direct Cost to a Health Plan. Am J Prev Med. 2004; 27(4):304-9
101 Ghosh R, Rubinov A, Zhang J. Optimization approach for clustering datasets with weights. Optim Methods Softw. 2005; 20(2-3): 335-51.
102 Gidlow C, Johnston LH, Crone D, Ellis N, James D. A systematic review of the relationship between socio-economic position and physical activity. Health Educ J. 2006;65:338-67
103 Gower JC, Legendre P. Metric and Euclidean properties of dissimilarity coefficients. J Classif. 1986;3(1):5-48.
104 Gower JC. A general coefficient of similarity and some of its properties. Biometrics. 1971;27:857–72.
105 Grambeier J, Rudolph A. Techniques of cluster algorithms in data mining. J Data Min Knowl Discov. 2002; 6: 303-60.
106 Grujić V, Dragnić N, Harhaji S, Radić I, Šušnjević S. Overweight and obesity among adults in Serbia: Results from the National Health Survey, Eat Weight Disord. 2010; 15 (1-2):E34-E42.
107 Guha S, Rastogi R, Shim K. CURE: An efficient clustering algorithm for large databases. Inform Syst. 2001; 26(1):35-58.
108 Guha S, Rastogi R, Shim K. ROCK: A Robust Clustering Algorithm for Categorical Attributes. In: Kitsuregawa M, Maciaszek L, Papazoglou M, editors. Proceedings of the15th International Conference on Data Engineering; 1999. March 23-26; Sydney, Australia. IEE Computer Society; 1999. p.512-21.
109 Gurrutxaga I, Muguerza J, Arbelaitz O, Perez JM, Martin JI. Towards a standard methodology to evaluate internal cluster validity indices. Pattern Recognit Lett. 2011;32:505-15.
110 Haapanen-Niemi N, Miilunpalo S, Pasanen M, Oja P. The impact of smoking, alcohol consumption, and physical activity on the use of hospital services. Am J Public Health. 1999;89:691-98.
111 Hair JF, Black WC, Babin BJ, Anderson RE, Tatham RL. Multivariate data analysis. New Jersey: Prentice Hall; 2006.
112 Halkidi M, Bastiakis Y, Vazirgiannis M. On clustering validation techniques. J Intell Inf Syst. 2001;17(2-3):107-45.
113 Halkidi M, Vazirgiannis M, Batsistakis I. Quality scheme assessment in the clustering process. In: Zighed DA, Komorowski J, Zytkov J. Principles of Data Mining and Knowledge Discovery. Proceedings of the 4th European Conference, 2000; September 13–16 Lyon, France. Berlin: Springer-Verlag; 2000. p.265-76.
114 Hands S, Everitt B. A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical clustering techniques. Multivar Behav Res. 1987; 22:235-43.

115 Hansen P, Jaumard B. Cluster analysis and mathematical programming. Math Program. 1997; 79(1-3):191-215.
116 Hardy A. On the number of the clusters. Comput Stat Data Anal. 1996;23(1):83-96.
117 Hardy GH, Wright EM. An introduction to the theory of numbers. 6 th ed. Oxford: Clarendon Press;1979.
118 Hartigan JA. Clustering Algorithms . New York: John Wiley & Sons Inc; 1975.

119 He Z, Deng S, Xu X. Improving K-modes algorithm considering frequencies of attribute values in mode. In: Hao Y, Liu J, Wang Y, Cheung YM, editors. Lecture Notes in Computer Science 3801. Computational Intelligence and Security Part I. International Conference, CIS; 2005 December 15-19; Xi´an, China. Berlin: Springer Berlin Heidelberg; 2005. p.157-62.
120 He Z, Xu X, Deng S. Attribute value weighting in K-modes clustering. Expert Syst Appl. 2011;38(12):15365-9.
121 Hedar AR, Fukushima M. Hybrid simulated annealing and direct search method for nonlinear unconstrained global optimization. Optim Methods Softw. 2002;17:891-912
122 Hillo D, Nishida C, James WPT. A life course approach to diet, nutrition and prevention of chronic diseases. Public Health Nutr. 2004;7(1A):101-21.
123 Hiriart-Urruty JB, Lemarechal C. Convex Analysis and Minimization Algorithms 1&2. Berlin: Springer-Verlag; 1993.
124 Huang Z, Ng MK. A fuzzy k-modes algorithm for clustering categorical data, IEEE Trans Fuzzy Syst.1999; 7(4):446-452.
125 Huang Z. A fast clustering algorithm to cluster very large categorical data sets in data mining. In: Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery. Dept. of Computer Science. The University of British Columbia, Canada;1997. p. 1-8.
126 Huang Z. Clustering large data sets with mixed numeric and categorical values. In: Proceedings of the first Pacific Conference of Knowledge Discovery and Data Mining (PAKDD). Singapore: World Scientific; 1997; 21-34.
127 Huang Z. Extensions to the K-means algorithm for clustering large data sets with categorical values. J Data Min Knowl Discov. 1998;2:283-304.
128 Iwai N, Yoshiike N, Saitoh S, Nose T, Kushiro T, Tanaka H. Leisure-time physical activity and related lifestyle characteristics among middle-aged Japanese. J Epidemiol. 2000;10(4):226–33.
129 Jaccard P. The distribution of the flora in the Alpine zone. New Phytol 1912; 11:37-50.
130 Jain AK, Dubes RC. Algorithms for Clustering Data. New Jersey: Prentice-Hall; 1988.
131 Jain AK, Murty MN, Flynn PJ. Data clustering: A review. ACM Comput Surv. 1999; 31(3): 264-323.
132 Jain AK. Data Clustering: 50 years beyond k-means. Pattern Recognit Lett. 2010; 31:651-66.

133 Kaplan GA, Strawbridge WJ, Cohen R, Hungerford LN. Natural history of leisure time physical activity and its correlates: associations with mortality from all causes and cardiovascular disease over 28 years. Am J Epidemiol. 1996;144:793-97.
134 Karp RM. Reducibility among combinatorial problems. In: Miller RE, Thatcher JW, editors. Complexity of Computer Computations. Proceedings of a symposium on the Complexity of Computer Computations; 1972 March 20-22, New York. New York: Plenum Press; 1972.p.85-103.
135 Karypis G, Han EH, Kumar V. Chameleon: hierarchical clustering using dynamic modeling. Computer. 1999;32(8):68-75.
136 Kaufman L, Rousseeuw PJ. Finding Groups in Data. New York: John Wiley & Sons; 1990.
137 Khaw KT, Jakes R, Bingham S, Welch A, Luben R, Day N, Wareham N. Work and leisure time physical activity assessed using a simple, pragmatic, validated questionnaire and incident cardiovascular disease and all-cause mortality in men and women: The European Prospective Investigation into Cancer in Norfolk prospective population study. Int J Epidemiol. 2006;35(4):1034-43.
138 Khaw KT, Wareham N, Bingham S, Welch A, Luben R, Day N. Combined Impact of Health Behaviours and Mortality in Men and Women: The EPIC-Norfolk Prospective Population Study. Obstet Gynecol Surv. 2008; 63(6):376-7.
139 King K, Meader N, Wright K., Graham H, Power C, Petticrew M et al. Characteristics of interventions targeting multiple lifestyle risk behaviours in adult populations: a systematic scoping review. PloS one. 2015;10(1): e0117015.
140 Kiwiel KC, Krzysztof C. Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization. Siam J Optimiz. 2007;18(2):379-88
141 Kiwiel KC. Methods of descent for nondifferentiable optimization. Lecture Notes in Mathematics 1133. Berlin: Springer-Verlag; 1985
142 Knoops KB, de Groot LM, Kromhout D, Perrin AE, Moreiras-Varela O, Menotti A, Van Staveren WA. Mediterranean diet, lifestyle factors, and 10-year mortality in elderly European men and women: the HALE project. Jama. 2004; 292(12):1433-1439.
143 Kontis V, Mathers CD, Rehm J, Stevens GA, Shield KD, Bonita R, et al. Contribution of six risk factors to achieving the 25×25 noncommunicable disease mortality reduction target: a modelling study. Lancet. 2014;384(9941): 427-437.
144 Kovačić ZJ. Multivarijaciona analiza. Beograd. Ekonomski fakultet;1994.
145 Krejić N, Lužanin Z, Rapajić S. Jacobian smoothing Brown's method for NCP. Nonlinear Anal Theory Methods Appl. 2009; 70: 642-657
146 Kriksciuniene D, Sakalauskas V, Tamasauskas D. Evaluation framework of hierarchical clustering methods for binary data. In: Hybrid Intelligent Systems. 12th International Conference on HIS; 2012 4-7 December; Pune, India. IEE; 2012. p421-26.

147 Kryszczuk K, Hurley P. Estimation of the number of clusters using multiple clustering validity indices. In: Gayar NE, Kittler J, Rolli F, editors. Multiple Classifier Systems. Berlin: Springer-Verlag Heidelberg; 2010. p 114-23.
148 Krzanowski WJ, Lai YT. A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering. Biometrics. 1988;44:23-34.
149 Kulczynski S. Die Pflanzenassoziationen der Pieninen. Bull Int Acad Tcheque Sci. 1927;3:57-203.
150 Kvaavik E, Meyer HE, Tverdal A, 2004. Food habits, physical activity and body mass index in relation to smoking status in 40-42 year old Norwegian women and men. Prev Med. 2004;38:1-5.
151 Laaksonen M, Prättälä R, Karisto A. Patterns of unhealthy behaviour in Finland. Eur J Public Health. 2001;11(3):294–300.
152 Lahelma E, Martikainen P, Laaksonen M, Aittomaki A. Pathways between socioeconomic determinants of health. J Epidemiol Community Health. 2004; 58(4):327-32.
153 Lambert EV, Kolbe-Alexandre T. Physical activity and chronic diseases of lifestyle in South Africa. In: Steyn K, Fourie J, Temple N, editors. Chronic Diseases of Lifestyle in South Africa: 1995-2005. Technical Report. CapeTown: Medical Research Council; 2006.
154 Lantz PM, House JS, Lepkowski JM, Williams DR, Mero RP, Chen J. Socioeconomic factors, health behaviors, and mortality: results from a nationally representative prospective sudy of US adults. Jama. 1998;279(21): 1703-8.
155 Li C, Biswas G. Unsupervised learning with mixed numeric and nominal data. IEEE Trans Knowl Data Eng. 2002;14(4):673-90.
156 Li J, Siegrist J. Physical activity and risk of cardiovascular disease- a meta-analysis of prospective cohort studies. Int J Environ Res Public Health. 2012; 9(2):391-407.
157 Li T, Ma S, Ogihara M. Entropy-based criterion in categorical clustering. In: Greiner R, Schuurmans D, editors. Proceedings of the 21 st International Conference on Machine Learning (ICLM-04); ACM Press; 2004. p.536-43
158 Likas A, Vlassis N, Verbeek JJ. The global k-means clustering algorithm. Pattern Recognit. 2003;36:451-61.
159 Lim SS, Vos T, Flaxman AD, Danaei G, Shibuya K, Adair-Rohani H, Davis A. (2013). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet. 2010;380(9859):2224-60.
160 Linardakis M, Smpokos E, Papadaki A, Komninos ID, Tzanakis N, Philalithis A. Prevalence of multiple behavioral risk factors for chronic diseases in adults aged 50+, from eleven European countries-the SHARE study (2004). Prev Med. 2013;57(3):168-72.

161 Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJL. Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. Lancet 2006;367:1747-57.
162 Lv J, Liu Q, Ren Y, Gong T, Wang S, Li L. Socio-demographic association of multiple modifiable lifestyle risk factors and their clustering in a representative urban population of adults: a cross-sectional study in Hanghzhou, China. Int J Behav Nutr Phys Act. 2011;8(40):1-13.
163 Maatoug J, Harrabi I, Hmad S, Belkacem M, Al'absi M, Lando H, Ghannem H. Clustering of risk factors with smoking habits among adults, Sousse, Tunisia. Prev Chronic Dis. 2013;10:E211.
164 Macias R, Garrido-Munoz M, Tejero-Gonzalez CM, Lucia A, Lopez-Adan E, Rodriguez-Romo G. Prevalence of leisure-time sedentary behaviour and sociodemographic correlates: a cross-sectional study in Spanish adults. BMC Public Health, 2014;14:972.
165 MacQueen JB. Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Berkeley. University of California Press, 1967; 1: 281-97.
166 Mäkelä MM, Neittanmäki P. Nonsmooth Optimization: Analysis and Algorithms with Applications to Optimal Control. Singapore: World Scientific Publishing Co; 1992.
167 Mäkelä MM. Survey of bundle methods for nonsmooth optimization. Optim Methods Softw. 2002;17(1):1-29.
168 Mäkelä MM, Karmitsa N, Bagirov A. Subgradient and bundle methods for nonsmooth optimization. In: Repin S, Tiihonen T, Tuovinen T, editors. Numerical Methods for Differential Equations, Optimization, and Technological Problems. Dordrecht: Springer Netherlands; 2013. p. 275-304.
169 Mangasarian OL. Mathematical programming in data mining. J Data Min Knowl Discov. 1997;1(2):183-201.
170 Martínez-González M, Varo J, Santos J, De Irala J, Gibney M, Kearney J, Martínez J. Prevalence of physical activity during leisure time in the European Union. Med Sci Sports Exerc. 2001;33(7):1142-1146.
171 Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med. 2006; 3:e442.
172 McGinnis JM, Foege WH. Actual causes of death in the United States. Jama 1993; 270(18):2207-2212.
173 Mercer DP. Clustering large datasets, Linacre College; 2003. http://www.stats.ox.ac.uk/~mercer/documents/Transfer.pdf
174 Milligan GW, Cooper MC. A study of standardization of variables in cluster analysis. J Classif. 1988; 5(2):181-204.
175 Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. Psychometrika. 1985; 50(2):159-179.
176 Milligan GW. A monte carlo study of thirty internal criterion measures for cluster analysis. Psychometrika. 1981;46(2):187-199.

177 Milligan GW. An examination of the effect of six types of error petrurbation on fifteen clustering algorithms. Psychometrika. 1980;45(3):325-342.
178 Ming-Tso Chiang M, Mirkin B. Intelligent choice of the number of clusters in K-means clustering: an experimental study with different cluster spreads. J Classif. 2010;27(1):3-40.
179 Ministarstvo zdravlja Republike Srbije. Istraživanje zdravlja stanovnika Republike Srbije, 2006.godina. Osnovni rezultati. Beograd: Ministarstvo zdravlja Republike Srbije, 2007.
180 Ministartsvo zdravlja Republike Srbije. Strategija za prevenciju i kontrolu hroničnih nezaraznih bolesti Repub
like Srbije; 2009
181 Mokdad AH, Marks JS, Stroup DF, Gerberding JL. Actual causes of death in the United States, 2000. Jama. 2004; 291(10):1238–1245.
182 Monteiro CA, Conde WL, Matsudo SM, Matsudo VR, Bonsenor IM, Lotufo PA. A descriptive epidemiology of leisure-time physical activity in Brazil, 1996–1997. Rev Panam Salud Publica. 2003;14(4):246–254.
183 Mordukhovich BS. Variational Analysis & Generalized Differentiation I. Basic Theory, Berlin: Springer-Verlag Grundlehren; 2006.
184 Mordukhovich BS. Variational Analysis & Generalized Differentiation II. Applications, Springer-Verlag Grundlehren; 2006.
185 Murty MN, Krishna G. A computationally efficient technique for data clustering. Pattern Recognit. 1980;12:153-158.
186 Nacionalno istraživanje o stilovima života stanovništva Srbije 2014.godine korišćenje. Osnovni rezultati o korišćenju psihoaktivnih supstanci i igre na sreću. Beograd, Institut za javno zdravlje Srbije „Dr Milan Jovanović Batut“, 2015.
187 Negin J, Cumming R, de Ramirez SS, Abimbola S, Sachs SE. Risk factors for non-communicable diseases among older adults in rural Afrika. Trop Med Int Health. 2011;16(5):640-6
188 Ng MK, Junjie Li M, Zhexue Huang J, He Z. On the impact dissimilarity measure in K-Modes clustering algorithm. IEEE Trans Pattern Anal Mach Intell 2007; 29(3):503-7.
189 Ng RT, Han J. Efficient and effective clustering methods for spatial data mining. In: Proceedings of the 20th International Conference on Very Large Data Bases. Santiago, Chile. San Francisco: Morgan Kaufmann Publishers Inc; 1994. p144-55.
190 Norman A, Bellocco R, Vaida F, Wolk A. Total physiscal activity in relation to age, body mass, health and other factors in a cohort of Swedish men. Int J Obes. 2002;26:670-5.
191 Novaković B, Božić D. Učestalost šećerne bolesti, gojaznosti i visokog krvnog pritiska u populaciji AP Vojvodine. Edicija monografije 62. Novi Sad: Medicinski fakultet; 2004.

192 Osler M, Tjønneland A, Suntum M, Thomsen BL, Stripp C, Grønbaek M et al. Does the association between smoking status and selected healthy foods depend on gender? A population-based study of 54 417 middle-aged Danes. Eur J Clin Nutr. 2002;56(1):57–63.
193 Padrão P, Lunet N, Santos AC, Barros H. Smoking, alcohol, and dietary choices: evidence from the Portuguese National Health Survey. BMC Public Health. 2007;7(1):138.
194 Paffenbarger RS, HydeRT, Wing AL, Lee IM, Jung DL, Kampert JB. The association of changes in physical-activity level and other lifestyle characteristics with mortality among men. N Engl J Med. 1993;328(8):538–45.
195 Panagiotakos DB, Pitsavos C, Chrysohoou C, Risvas G, Kontogianni MD, Zampelas A, Stefanadis C. Epidemiology of overweight and obesity in a Greek adult population: the ATTICA Study. Obes Res. 2004; 12(12):1914-20.
196 Pardalos PM, Boginski VL, Prokopyev OA, Suharitdamrong W, Carney PR, Chaovalitwongse W et al. Optimization Techniques in Medicine. In: Audet C, Hansen P, Savard G. Essays and Surveys in Global Optimization. New York: Springer; 2005. p. 211-232.
197 Pomerleau J, Pederson LL, Østbye T, Speechley M, Speechley KN. Health behaviors and socio-economic status in Ontario, Canada. Eur J Epidemiol. 1997;13:613 – 622.
198 Poortinga W. The prevalence and clustering of four major lifestyle risk factors in an English adult population. Prev Med. 2007;44(2):124–128.
199 Pronk NP, Anderson LH, Crain AL, Martinson, O'Connor PJ, Sherwood NE, Whitebird RR. Meeting recommendations for multiple healthy lifestyle factors. Prevalence, clustering, and predictors among adolescent, adult, and senior health plan members. Am J Prev Med. 2004;27(2):25-33.
200 Pronk P, Goodman M, O'Connor P, Martinson B. Relationship between modifiable health risks and short-term health care charges. Jama. 1999; 282(23): 2235–9.
201 Punj G, Stewart DW. Cluster analysis in marketing research: review and suggestions for application. J Mark Res. 1983;20(2):134-48.
202 Rademacher H. Uber paftielle und totale Differenzierbarkeit. Math Ann. 1919; 89:340-59.
203 Ralambondrainy H. A conceptual version of the K-Means algorithm. Pattern Recognit Lett. 1995;16:1147-57.
204 Rama B, Jayashree P, Jiwani S. A survey on clustering. Current status and challenging issues. International Journal on Computer Science and Engineering. 2010; 2(9):2976-80.
205 Reedy J, Haines PS, Campbell MK. The influence of health behavior clusters on dietary change. Prev Med. 2005;41(1):268–75.
206 Reeves CR. Modern heuristic techniques for combinatorial problems. London: Blackwell; 1993.
207 Remington PL, Brownson RC, Wegner MV. Chronic Disease Epidemiology and Control. 3 rev.ed. Washington: APHA Press; 2010.

208 Rimm EB, Moats C. Alcohol and coronary heart disease: drinking patterns and mediators of effect. Ann Epidemiol. 2007;17(5):S3-S7.
209 Rockfaller RT. Convex Analysis. Princeton University Press; 1970.
210 Rogers DJ, Tanimoto TT. A Computer program for classifying plants. Science. 1960;132:1115-8.
211 Rokach L, Maimon O. Clustering methods. In: Maimon O, Rokach L, editors. Data mining and knowledge discovery handbook. New York: Springer Verlag; 2005. p.321-52.
212 Rolls BJ, Ello-Martin JA, Tohill BC. What can intervention studies tell us about the relationship between fruit and vegetable consumption and weight management? Nutr Rev. 2004; 62(1):1-17.
213 Roos G, Johansson L, Kasmel A, Klumbiené J, Prättälä R. Disparities in vegetable and fruit consumption: European cases from the north to the south. Public Health Nutr. 2001;4(1):35-43.
214 Rousseeuw P. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53-65.
215 Rubinov AM, Soukhoroukova NV, Ugon J. Classess and clusters in data analysis. Eur J Oper Res. 2006;173 849-65.
216 Rubinov AM, Soukhoroukova NV, Ugon J. Minimization of the sum of minima of convex functions and its application to clustering. In: Jeyakumar V, Rubinov A, editors. Continuous Optimization. Current Trends and Modern Applications. New York: Springer; 2005. p.409-34.
217 Rubinov AM. Abstract Convexity and Global Optimization. Dordrecht: Kluwer Academic Publishers; 2000.
218 Rusell PF, Rao TR. On habitat and association of species of anopheline larvae in south-eastern Madras. J Malar Inst India. 1940;3(1):153-178.
219 Saitta S, Raphael B, Smith IFC. A comprenhesive validity index for clustering. IntelL Data Anal 2008;12:529-48.
220 Saltin B, Grimsby G. Physiological analysis of middle-aged and old former athletes. Comparison with still active athletes of the same ages. Circulation. 1968;38:1104-15.
221 San OM, Huynh VN, Nakamori Y. An alternative extension of the k-means algorithm for clustering categorical data. Int J Appl Math Comp Sci. 2004;14(2):241-247 .
222 Schaap MM, Van Agt HME, Kunst AE. Identification of socioeconomic groups at increased risk for smoking in European countries: looking beyond educational level. Nicotine Tob Res. 2008; 10(2):359-69.
223 Schneider S, Huy C, Schuessler M, Diehl K, Schwar S. Optimising lifestyle interventions: identification of health behaviour patterns by cluster analysis in a German 50+ survey. Eur J Public Health. 2009; 19(3):271-7.
224 Schuit AJ, Loon AJ, Tijhuis M, Ocke M. Clustering of lifestyle risk factors in a general adult population. Prev Med. 2002; 35:219–24.
225 Schwarz G. Estimating the dimension of a model. Ann Stat.1978, 6(2):461-4

226 Serdula MK, Byers T, Simoes AH, Mendlein E, Coates JM. The association between fruit and vegetable intake and chronic disease risk actors. Epidemiology. 1996;7:161–165.
227 Shankar A, McMunn A, SteptoeA. Health-related behaviors in older adults relationships with socioeconomic status. Am J Prev Med. 2010;38(1):39–46.
228 Sherali HD, Desai J. A global optimization RLT-based approach for solving the hard clustering problems. J Glob Optim. 2005;32(2):281-306.
229 Shor NZ. Minimization methods for non-differentiable functions. Berlin: Springer-Verlag; 1985.
230 Silva DAS, Peres KG, Boing AF, González-Chica DA, Peres MA. Clustering of risk behaviors for chronic noncommunicable diseases: a population-based study in southern Brazil. Prev Med. 2013; 56(1):20–24.
231 Sokal RR, Michener CD. A statistical method for evaluating systematic relationships. Univ Kans Sci Bull.1958; 28:1409-1438.
232 Sokal RR, Sneath PH. Principles of Numerical Taxonomy. San Francisco: W.H. Freeman and Company; 1963.
233 Spath H. Cluster Analysis Algorithms. Chichester: Ellis Horwood Ltd; 1980.
234 Stephens T, Caspersen CJ. The demography of physical activity. In: Bouchard C, Shepard RJ. Stephens T, editors. Physical activity, fitness and health: international proceedings and consensus statement. Champaign: Human Kinetics Pub; 1994. p.204-213
235 Steptoe A, Wardle J. What the expert think: a European survey of expert opinion about the influence of lifestyle on health. Eur J Epidemiol. 1994; 10:195-203.
236 Sugar CA, James GM. Finding the number of clusters in a dataset: An Information-Theoretic approach. J Am Stat Assoc. 2003;98(463):750-763.
237 Sun Y, Zhu Q, Chen Z. An iterative initial-points refinement algorithm for categorical data clustering. Pattern Recognit Lett. 2002;23:875-884.
238 Theodoridis S, Koutroumbas K. Pattern Recognition. San Diego: Academic Press; 1998.
239 Thompson RL, Margetts BM, Wood DA, Jackson AA. Cigarette smoking and food and nutrient intakes in relation to coronary heart disease. Nutr Res. 1992; 5(01):131–152.
240 Timm NH. Applied Multivariate Analysis. New York: Springer; 2002.
241 Tohill BC, Seymour J, Serdula M, Kettel-Khan L, Rolls BJ. What epidemiologic studies tell us about the relationship between fruit and vegetable consumption and body weight. Nutr Rev. 2004; 62(10):365-374.
242 Trichopoulou A, Naska A, Costacou T. Disparities in food habits across Europe. Proc Nutr Soc 2002;61(4):553-558.
243 Tyron RC, Bailey DE. Cluster Analysis. New York: McGraw-Hill; 1970.
244 UCI machine learning respository. URL <http://www.ics.uci.edu/mlearn/MLRRespository.html>
245 Ukropina S. Prevalencija pušenja i uticaj na ishod trudnoće [doktorska disertacija]. Medicinski fakultet Novi Sad; 2012.

246 Vendramin L, Campello, RJ, Hruschka ER. Relative clustering validity criteria: A comparative overview. Stat Anal Data Min. 2010;3(4):209-35.
247 Vermeulen-Smit E, Ten Have M, Van Laar M, De Graaf R. Clustering of health risk behaviours and the relationship with mental disorders. J Affect Disord. 2015;171:111-9.
248 Wang W, Yang J, Muntz R. STING: A Statistical Grid Appraoch to Spatial Data Mining. In: Jarke M, Carey MJ, Dittrich KR, Lochovsky FH, Loucopoulos P, Jeusfeld MA, editors. Proceedings of the 23rd International Conference on Very Large Data Bases. San Francisco: Morgan Kaufmann Publishers Inc; 1997. p.186-95.
249 Wareham NJ, Jakes RW, Rennie KL, Schuit J, Mitchell J, Hennings S, Day N. Validity and repeatability of a simple index derived from the short physical activity questionnaire used in the European Prospective Investigation into Cancer and Nutrition (EPIC) study. Public Health Nutr. 2003; 6(4):407-413.
250 Wendel-Vos GC, Schuit AJ, Feskens EJ, Boshuizen HC, Verschuren WM, Saris WH, Kromhout D. Physical activity and stroke. A meta-analysis of observational data. Int J Epidemiol. 2004;33(4):787-98.
251 Willet WC, Dietz WH, Colditz GA. Guidelines for healthy weight. N Engl J Med. 1999;341(6):427-434.
252 Willet WC, Koplan JP, Nugent R, Dusenbury C, Puska P, Gayiano TA. Prevention of Chronic Disease by Means of Diet and Lifestyle Changes. In: Jamison DT, Breman J, Measham AR, Alleyne G, Claeson M, Evans D et al. Disease Control Priorities in Developing Countries. World Bank Publications; 2006.
253 Wirfalt E, Mattisson I, Gullberg B, Berglund G. Food patterns defined by cluster analysis and their utility as dietary exposure variables: a report from the Malmo Diet and Cancer Study. Public Health Nutr. 2000;3;159-173.
254 World Health Organisation. Global Strategy on Diet, Physical Activity and Health. Geneva: WHO; 2004.
255 World Health Organization. Bridging the gaps. The World Health Report. Geneva: WHO; 1995.
256 World Health Organization. Diet and physical activity: a public health priority. www.who.int/dietphysicalactivity/en/
257 World Health Organization. Diet, Nutrition and the Prevention of Chronic Diseases. Report of a Joint WHO/FAO Expert Consultation. WHO Technical Report Series 916. Geneva: WHO; 2003.
258 World Health Organization. Global Action Plan for the Prevention and Control of Non-communicable diseases 2013-2020. Geneva: WHO; 2013.
259 World Health Organization. Global Health Observatory Database: Prevalence of insufficient physical activity. http://www.who.int/gho/ncd/risk_factors/physical_activity_text/en/index.html.
260 World Health Organization. Global Health Risks: Mortality and Burden Disease Attributable to Selected Major Risks. Geneva: WHO; 2009.

261 World Health Organization. Global status report on alcohol and health. Geneva: WHO; 2011.
262 World Health Organization. Obesity and overweight. 2003.
263 World Health Organization. Parties to the WHO framework convention on tobacco control. http://www.who.int/fctc/signatories_parties/en/ (accessed Jan 22, 2015).
264 World Health Organization. Reducing risk, promoting healthy life. World Health Report. Geneva:WHO; 2002.
265 World Health Organization. WHO framework convention on tobacco control. Geneva: WHO;2003.
266 World Health Organization. World Health Report 2004: Changing history. Geneva: WHO; 2004.
267 Xu R, Wunsch D. Survey of clustering algorithms. IEEE Trans Neural Netw. 2005;16(3):645-78.
268 Yoon YS, Oh SW, Park HS. Socioeconomic status in relation to obesity and abdominal obesity in Korean adults: a focus on sex differences. Obesity. 2006; 14(5): 909-19.
269 Yusuf S, Hawken S, Ounpuu S, Dans T, Avezum A, Lanas F et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. Lancet. 2004; 364(9438):937-52.
270 Zhang T, Ramakakrishnan R, Livny M. BIRCH: A new data clustering agorithm and its applications. J Data Min Knowl Discov. 1997;1(2):141-82.
271 Zhang T, Ramakrishnan R, Livny M. BIRCH: An Efficient Data Clustering Method For Very Large Databases. In: Windom J, editor. Proceedings of the 1996 ACM SIGMOD international conference on Management of data ACM SIGMOD Record. New York: ACM; 1996; 25(2):103-14.
272 Zubin JA. A technique for measuring like-mindedness. J Abnorm Soc Psych. 1938;33(4):508-16.

PRILOZI

Prilog 1. Upitnik za domaćinstvo

Prilog 2. Upitnik za odrasle osobe starosti 20 i više godina

Prilog 3. Upitnik za samopopunjavanje za odrasle

Dobro jutro/dan/veče, moje ime je _____ . Mi smo iz Istraživačke agencije Strateški Marketing. Radimo na projektu Ministarstva zdravlja Republike Srbije koji se bavi istraživanjem zdravlja stanovništva. Bili bismo Vam veoma zahvalni ukoliko uzmete učešće u našoj anketi. Sve dobijene informacije će se tretirati kao strogo poverljive i nikada se neće otkriti njihov izvor. Niste obavezni da odgovorite na pitanje na koje ne želite i u svakom trenutku možete prekinuti razgovor.

UPITNIK ZA DOMAĆINSTVO

DEO DM - INFORMACIONI PANEL ZA DOMAĆINSTVO

DM1. Redni broj popisnog kruga u uzorku: ☐☐☐

DM2. Redni broj domaćinstva u popisnom krugu: ☐☐

DM3. Ime i prezime anketara: _____

Šifra anketara: ☐☐☐☐☐

DM4. Ime i prezime kontrolora: _____

Šifra kontrolora: ☐☐☐☐☐

DM5. Dan / mesec / godina anketiranja: ☐☐ / ☐☐ / 2006

DM6. Adresa domaćinstva: _____

DM7. Telefon domaćinstva: _____

[ANK] Nije obavezno upisati broj telelefona domaćinstva

Pošto su popunjeni svi upitnici za ovo domaćinstvo, uneti sledeće podatke:

DM8. Rezultat popunjavanja Upitnika za domaćinstvo:

Upitnik za domaćinstvo je popunjen.....1
Niko nije kod kuće..... 2
Odbili da sarađuju..... 3
Domaćinstvo nije pronađeno..... 4

Drugo (*navesti*) _____ 95

DM9. Ime glavnog ispitanika:

Ime:_____

[ANK] Nije obavezno upisati ime glavnog ispitanika

DM10. Ukupan broj članova domaćinstva: ☐☐

DM11. Broj odraslih osoba - 20 godina i više: ☐☐

DM12. Broj popunjenih Upitnika za odrasle osobe 20+ : ☐☐

DM13. Broj urađenih Upitnika za samopopunjavanje za odrasle osobe 20+ : ☐☐

DM14. Broj dece od 7 do 19 godina: ☐☐

DM15. Broj popunjenih Upitnika za decu od 7 do 19 g. : ☐☐

DM16. Broj dece od 12 do 19 godina: ☐☐

DM17. Broj urađenih Upitnika za samopopunjavanje za decu od 12 do 19 godina: ☐☐

Napomena za anketara / kontrolora: Upisati napomene u vezi sa anketiranjem članova domaćinstva, kao što su broj ponovljenih poseta, nepotpuni pojedinačni formulari, broj pokušaja ponovljene posete i slično.

DM18. Šifra lica koje vrši unos podataka: ☐☐

DEO SD – SPISAK ČLANOVA DOMAĆINSTVA

U red 01 upisana je šifra 1. Glavni ispitanik. Dalje redom upisivati šifre srodstva sa glavnim ispitanikom svih članova domaćinstva (kolona SD2).

Zatim pitati: **Da li je još neko član domaćinstva, iako trenutno ne živi ovde?** *Ukoliko je odgovor potvrdan, dopuniti spisak ostalim članovima domaćinstva. Zatim preći na deo KD – Karakteristike domaćinstva.*

Broj reda člana domaćinstva:	SD1. Ime:	SD2. Srodstvo sa glavnim ispitanikom: 2.Supruga/Suprug 3. Ćerka/Sin 4. Majka/Otac 5. Sestra/Brat 6. Unuka/Unuk 7. Baba/Deda 8. Drugi rođaci 9. Nisu u srodstvu, ali žive u istom domaćinstvu	SD3. Pol: Ženski	SD3. Pol: Muški	SD4. Datum rođenja: Dan/Mesec/Godina	SD5. Navršene godine života:	SD6. Važi za upitnik: Za odraslu osobu 20+ *Zaokružiti broj reda za osobu 20+*	SD6. Važi za upitnik: Za dete od 7 do 19 *Zaokružiti broj reda za dete od 7 do 19*	SD6. Važi za upitnik: Za dete od 12 do 19 *Zaokružiti broj reda za dete od 12 do 19*
01		1. Glavni ispitanik	1	2	__ __ / __ __ / __ __ __ __		01	01	01
02			1	2	__ __ / __ __ / __ __ __ __		02	02	02
03			1	2	__ __ / __ __ / __ __ __ __		03	03	03
04			1	2	__ __ / __ __ / __ __ __ __		04	04	04
05			1	2	__ __ / __ __ / __ __ __ __		05	05	05
06			1	2	__ __ / __ __ / __ __ __ __		06	06	06
07			1	2	__ __ / __ __ / __ __ __ __		07	07	07
08			1	2	__ __ / __ __ / __ __ __ __		08	08	08
09			1	2	__ __ / __ __ / __ __ __ __		09	09	09
10			1	2	__ __ / __ __ / __ __ __ __		10	10	10
11			1	2	__ __ / __ __ / __ __ __ __		11	11	11
12			1	2	__ __ / __ __ / __ __ __ __		12	12	12
13			1	2	__ __ / __ __ / __ __ __ __		13	13	13
14			1	2	__ __ / __ __ / __ __ __ __		14	14	14
15			1	2	__ __ / __ __ / __ __ __ __		15	15	15
							Odrasla osoba	**Dete 7-19**	**Dete 12-19**
Ukupno							__ __	__ __	__ __

[ANK] Sada bi trebalo pripremiti posebne upitnike za svaku odraslu osobu starosti 20 i više godina koja živi u ovom domaćinstvu: Upitnik za odraslu osobu 20+ i Upitnik za samopopunjavanje za odraslu osobu 20+.

Za svako dete uzrasta 7 do 19 godina pripremiti Upitnik za dete od 7 do 19 godina.

Za svako dete uzrasta 12 do 19 godina pripremiti Upitnik za samopopunjavanje za dete od 12 do 19 godina.

NASTAVITI SA POPUNJAVANJEM UPITNIKA ZA DOMAĆINSTVO.

DEO KD – KARAKTERISTIKE DOMAĆINSTVA

KD1	**Koje je vrste stambeni objekat u kome živite?**	1. Kuća 2. Stan u kući 3. Stan u zgradi sa manje od 15 stanova 4. Stan u zgradi sa više od 15 stanova 95. Drugo, (navesti)_____	KD2
KD2	**Ko je vlasnik kuće/stana?**	1. Jedan od članova domaćinstva 2. Država/Preduzeće 3. Roditelj 4. Stanodavac 95. Drugo, (navesti)_____ *NZ (Ne zna)*	KD3
KD3	**Kolika je površina Vaše kuće/stana?**	_____ m^2	KD4
KD4	**Koliko imate soba u kući/stanu?** *[ANK] Računaju se sve sobe uključujući i dnevnu sobu.*	_____ soba	KD4A
KD4A	**Koliko se prostorija u kući/stanu domaćinstva koristi za spavanje?**	_____ prostorija za spavanje	KD5
KD5	**Od kog <u>osnovnog</u> materijala je napravljen pod u kući/stanu?** *ANK] Jedan odgovor. Zaokružiti preovlađujući materijal.*	1. Parket/laminat/brodski pod/keramičke pločice 2. Patos/beton 3. Zemlja 95. Drugo, (navesti)_____	KD5A
KD5A	**Od kog <u>osnovnog</u> materijala je napravljen krov kuće/zgrade?** *ANK] Jedan odgovor. Zaokružiti preovlađujući materijal.*	1. Slama 2. Trska 3. Drvene daske 4. Lim 5. Crep 6. Cementna/betonska ploča 7. Krovna šindra 95. Drugo, (navesti)_____	KD5B
KD5B	**Od kog <u>osnovnog</u> materijala su sagrađeni zidovi u kući/stanu domaćinstva?** *[ANK] Jedan odgovor. Zaokružiti preovlađujući materijal.*	1. Naboj (trska, slama, blato) 2. Kamen i blato 3. Nepečena cigla 4. Šperploča 5. Karton 6. Polovna građa 7. Beton 8. Kamen sa cementom 9. Cigla 10. Cementni blok 11. Drvene daske/šindra 95. Drugo, (navesti)_____	KD6
KD6	**Da li u Vašoj kući/stanu imate električnu energiju (struju)?**	1. Ne 2. Da	KD7
KD7	**Koji izvor energije <u>najčešće</u> koristite za grejanje u Vašoj kući/stanu?** *[ANK] Jedan odgovor*	1. Električna energija 2. Gas 3. Drvo 4. Ugalj 5. Nafta/Lož ulje/Mazut 95. Drugo, (navesti)_____	KD8
KD8	**Kako procenjujete uslove Vašeg stanovanja?** *[ANK] Jedan odgovor*	1. Vrlo loši 2. Loši 3. Prosečni 4. Dobri 5. Vrlo dobri *NZ (Ne zna)*	DEO VO

DEO VO - SNABDEVANJE PIJAĆOM VODOM I UKLANJANJE OTPADNIH MATERIJA

VO1	**Da li imate izvor vode/priključak za vodu u kući/stanu?**	1. Ne	VO2
		2. Da	VO4
VO2	**Koliko je udaljen izvor vode za piće od Vašeg domaćinstva?** *[ANK] Jedan odgovor*	1. U dvorištu 2. Manje od 100 m 3. Od 100 do 199 m 4. Od 200 do 499 m 5. Od 500 m do 1 km 6. Više od 1 km *NZ (Ne zna)*	VO3
VO3	**Koliko je vremena potrebno da biste došli do vode za piće?** *[ANK] Računa se samo u jednom pravcu*	_____ minuta	VO4
VO4	**Koji je <u>glavni</u> izvor pijaće vode za članove Vašeg domaćinstva**? *[ANK] Jedan odgovor*	1. Gradski vodovod 2. Seoski (lokalni) vodovod 3. Javna česma 4. Bušeni bunar 5. Pokriven kopani bunar ili uređen izvor 6. Nepokriven kopani bunar ili neuređen izvor 7. Jezero, reka, potok 8. Kišnica	VO5
		9. Flaširana voda	VO4A
		10. Cisterna 95. Drugo, (navesti)_____	VO5
VO4A	**Koji je glavni izvor vode koji Vaše domaćinstvo koristi za ostale potrebe, kao što je kuvanje ili pranje ruku?** *[ANK] Jedan odgovor*	1. Gradski vodovod 2. Seoski (lokalni) vodovod 3. Javna česma 4. Bušeni bunar 5. Pokriven kopani bunar ili uređen izvor 6. Nepokriven kopani bunar ili neuređen izvor 7. Jezero, reka, potok 8. Kišnica 9. Cisterna 95. Drugo, (navesti)_____	VO5
VO5	**Da li postoje prekidi u snabdevanju vodom?**	1. Ne 2. Da, povremeno 3. Da, svakodnevno 4. Da, u toku leta	VO6
VO6	**Da li imate nužnik (WC) u kući/stanu?**	1. Ne	VO7
		2. Da	VO8
VO7	**Koliko je udaljen nužnik (WC) od Vašeg domaćinstva?**	1. Manje od 50 m 2. Više od 50 m *NZ (Ne zna)*	VO8
VO8	**Kakvu vrstu nužnika (WC-a) koristi Vaše domaćinstvo?** *[ANK] Jedan odgovor*	1. Nužnik na ispiranje sa priključkom na kanalizaciju 2. Nužnik na ispiranje sa priključkom na septičku jamu 3. Nužnik bez ispiranja sa vodonepropusnom jamom 4. Poljski nužnik 5. Nema nužnik	VO9
VO9	**Da li pored Vašeg domaćinstva, nužnik (WC) koristi i neko drugo domaćinstvo?**	1. Ne 2. Da, 1—2 domaćinstva 3. Da, 3—5 domaćinstava 4. Da, više od 5 domaćinstava	VO10
VO10	**Kako uklanjate otpadne materije (đubre) iz Vašeg domaćinstva?** *[ANK] Jedan odgovor*	1. Odnosi se organizovano na nivou opštine/naselja 2. Odlažete ih na mesto predviđeno za to u naselju gde živite 3. Bacate ih na "divlje" deponije 4. Spaljujete ih 5. Zakopavate ih 6. Stavljate ih na gomilu u blizini kuće 7. Bacate ih u reku 8. Izbacujete u neposrednu okolinu 95. Drugo, (navesti)_____	DEO SE

DEO SE - SOCIOEKONOMSKO STANJE DOMAĆINSTVA

SE1 **Koliko ukupno izvora prihoda ima Vaše domaćinstvo?** _____ SE2

SE2 **Koji je <u>glavni</u> izvor novčanih prihoda u Vašem domaćinstvu?**
[ANK] Jedan odgovor.

1. Plata u državnoj službi
2. Plata kod privatnika
3. Penzija
4. Sopstveni posao
5. Poljoprivreda
6. Izdavanje nekretnina
7. Socijalna pomoć
8. Nema novčanih primanja
95. Drugo (navesti) _____

BO (Odbija da odgovori) SE3

SE3 **Na koji način obezbeđujete hranu za domaćinstvo?**

	Ne	Da
1. Kupovinom	1	2
2. Sopstvenom proizvodnjom	1	2
3. Dobijanjem pomoći od rođaka/prijatelja/ komšija	1	2
4. Hranim/o se u narodnoj kuhinji	1	2
95. Drugo (navesti) _____	1	2

SE4

SE4 **Procenite rashode Vašeg domaćinstva za troškove ishrane u toku prethodnog meseca:**

1. Manje od 30%
2. Od 30 do 50%
3. Od 51 do 70%
4. Preko 70%

NZ (Ne zna) SE5

SE5 **Da li su prihodi Vašeg domaćinstva u toku prethodnog meseca bili dovoljni za troškove:**

[ANK] Šifru 3 – Ne koristi moguće je zaokružiti isključivo za kategorije odgovora 6, 7 i 8.

	Ne	Da	Ne koristi
1. Ishrane	1	2	
2. Lične higijene	1	2	
3. Higijene domaćinstva	1	2	
4. Odeću, obuću	1	2	
5. Režijske troškove	1	2	
6. Zdravstvenu zaštitu (preglede, lekove)	1	2	3
7. Rekreaciju	1	2	3
8. Izlaske (u pozorište, bioskop, kafanu…)	1	2	3

SE6

SE6 **Da li je neko od članova Vašeg domaćinstva u toku prethodnih 12 meseci bio na letovanju/zimovanju?**

1. Ne
2. Da

SE7

SE7 **Da li Vaše domaćinstvo ima:**

	Ne	Da
1. Zemlju	1	2
2. Automobil	1	2
3. Traktor	1	2
4. Frižider	1	2
5. Bojler	1	2
6. Mašinu za pranje veša	1	2
7. Mašinu za pranje sudova	1	2
8. Televizor u boji	1	2
9. Telefon	1	2
10. Mobilni telefon	1	2
11. Personalni računar	1	2
12. Pristup internetu	1	2
13. Kupatilo	1	2
14. Centralno grejanje	1	2
15. Klima uređaj	1	2
16. Uštedevinu	1	2

SE8

SE8 **Kako procenjujete materijalno stanje Vašeg domaćinstva?**
[ANK] Jedan odgovor

1. Vrlo loše
2. Loše
3. Prosečno
4. Dobro
5. Vrlo dobro

NZ (Ne zna) SE9

SE9	**Koliki je ukupan prihod Vašeg DOMAĆINSTVA u prethodnom mesecu?** **Suma u dinarima.**	_____ dinara			SE10
SE10	**Koliko novca je, po Vašem mišljenju, mesečno potrebno Vašem domaćinstvu da bi moglo normalno da živi?** **Suma u dinarima.** *[ANK] Suma u pitanju SE 10 ne sme biti manja od sume iz pitanja SE 11.*	_____ dinara			SE11
SE11	**Koji je po Vašem mišljenju apsolutno minimalan iznos mesečno potreban da bi Vaše domaćinstvo bilo u stanju da pokrije najosnovnije životne potrebe?** **Suma u dinarima.** *[ANK] Suma u pitanju SE11 ne sme biti veća od sume date u pitanju SE10.* *Pod najosnovnijim životnim potrebama se podrazumevaju opcije od 1 do 6 iz pitanja SE5.*	_____ dinara			SE12
SE12	**Koliko je od Vaše kuće/stana udaljena najbliža:** *[ANK] Kada je u pitanju udaljenost u izražena u minutima, podrazumeva se vreme koje je potrebno da se stigne do određene zdravstvene ustanove prevoznim sredstvom koje se <u>najčešće</u> koristi .*	1. Ambulanta	a_____min.	b_____km.	DEO IZ
		2. Dom zdravlja	a_____min.	b_____km.	
		3. Bolnica	a_____min.	b_____km.	
		4. Apoteka	a_____min.	b_____km.	

DEO IZ - IZDACI ZA ZDRAVSTVENU ZAŠTITU PLAĆENI IZ SOPSTVENOG DŽEPA

IZ1	**Da li ste Vi ili neko od članova Vašeg domaćinstva u proteklih 12 meseci imali neke troškove vezane za zdravlje – zdravstveno stanje?** *[ANK] Ukoliko je zaokružen odgovor 2. „Da“ nastaviti sa pitanjima IZ2 do IZ4 za svakog člana domaćinstva.* *Ukoliko je na IZ1 zaokružen odgovor 1. „Ne“, preći sa postavljenjem pitanja za sledećeg člana domaćinstva.*		**1. Ne** **2. Da**	**1. Ne** **2. Da**	**1. Ne** **2. Da**	**1. Ne** **2. Da**	**1. Ne** **2. Da**	**1. Ne** **2. Da**	**1. Ne** **2. Da**	**1. Ne** **2. Da**	**1. Ne** **2. Da**	**1. Ne** **2. Da**	IZ2
			Broj reda člana domaćinstva iz Dela SD – Spisak članova domaćinstva									
IZ2	**Koliko su u toku prethodne 4 nedelje članovi Vašeg domaćinstva potrošili na:** *[ANK] Upisati u dinarima.*		**01**	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	
	1. Vanbolničko lečenje	A državna ustanova											
		B privatna praksa											
	2. Stomatološke usluge	A državna ustanova											
		B privatna praksa											
	3. Laboratorijske testove, rentgen snimke, ultrazvuk, skener, magnetnu rezonancu	A državna ustanova											
		B privatna praksa											
	4. Lekove												
	5. Pomoćna lekovita sredstva (vitamini, minerali i sl.)												
	6. Alternativnu medicinu (akupunktura, homeopatska medicina, hiropraktika, kostolomac i sl.)												
	7. Drugo (špricevi, igle, flasteri, gaza, vata, jod i sl.)												IZ3
IZ3	**Koliko su u toku prethodnih 12 meseci članovi Vašeg domaćinstva potrošili na:** *[ANK] Upisati u dinarima.*												
	1. Bolničko lečenje	A državna ustanova											
		B privatna praksa											
	2. Rehabilitaciju (stacionarnog tipa)												
	3. Pomagala (naočare, kontaktna sočiva, slušni aparat, proteze, ortopedske cipele, invalidska kolica...)												IZ4
IZ4	**Ako je neko od članova Vašeg domaćinstva u toku prethodnih 12 meseci direktno dao zdravstvenom osoblju određeni novčani iznos ili dao bilo kakav poklon, kolika je to bila vrednost?** *[ANK] Upisati u dinarima. UKOLIKO JE ODGOVOR „NISU ZAHTEVALI“ ZAOKRUŽITI ŠIFRU 2.*												
	Koliko ste platili zdravstvenom osoblju (natura ili gotovina) **na njihov zahtev**? *Ukoliko je odgovor „Nisu zahtevali“ zaokružiti šifru 2.*		2. Nisu zahtevali	2. Nisu zahtevali	2. Nisu zahtevali	2. Nisu zahtevali	2. Nisu zahtevali	2. Nisu zahtevali	2. Nisu zahtevali	2. Nisu zahtevali	2. Nisu zahtevali	2. Nisu zahtevali	
	Kolika je vrednost poklona (natura ili gotovina) koje ste **samoinicijativno** dali zdravstvenom osoblju?												
	Ne želim da odgovorim (BO) Ne čitati.		BO	BO	BO	BO	BO	BO	BO	BO	BO	BO	Sledeći upitnik

Mi smo iz Istraživačke agencije Strateški Marketing. Radimo na projektu Ministarstva zdravlja Republike Srbije koji se bavi istraživanjem zdravlja stanovništva. Želeo(la) bih da o tome porazgovaram sa Vama. Ovaj razgovor će trajati oko 30 minuta. Sve dobijene informacije će se tretirati kao strogo poverljive i nikada se neće otkriti njihov izvor. Niste obavezni da odgovorite na pitanje na koje ne želite i u svakom trenutku možete prekinuti razgovor.

UPITNIK ZA ODRASLE OSOBE STARE 20 GODINA I VIŠE

DEO DO- INFORMACIONI PANEL UPITNIK ZA ODRASLE OSOBE STARE 20 GODINA I VIŠE

Potrebno je popuniti poseban upitnik za svakog člana domaćinstva koji ispunjava uslov, a koji živi u tom domaćinstvu.
Upisati redni broj popisnog kruga u uzorku i redni broj domaćinstva u popisnom krugu, kao i ime i broj reda ispitanika.
Upisati ime i šifru anketara i datum anketiranja.

DO1. Redni broj popisnog kruga u uzorku: ___ ___ ___	**DO2. Redni broj domaćinstva u popisnom krugu:** ___ ___
DO3. Ime člana domaćinstva: _____	**DO4. Broj reda člana domaćinstva:** ___ ___
DO5. Ime i prezime anketara: _____ **Šifra anketara:** ___ ___ ___ ___ ___	**DO6. Dan / mesec / godina anketiranja:** ___ ___ / ___ ___ / **2 0 0 6**
DO7. Rezultat ankete rađene za odrasle osobe stare 20 godina i više: *[ANK] Šifre se odnose na člana domaćinstva koji bi trebalo da odgovara na upitnik, tj. na situaciju da li je taj član domaćinstva pristao/la na anketiranje.*	1. Upitnik je popunjen 2. Ispitanik nije kod kuće 3. Ispitanik odbija razgovor 4. Upitnik je delimično popunjen 95. Drugo, navesti: _____

Ponoviti uvodni pozdrav ukoliko to već nije učinjeno:

Poštovani,

Ovom prilikom želimo da Vam se najsrdačnije zahvalimo u ime istraživačke agencije Strategic Marketing Research što ste izdvojili svoje vreme i učestvovali u ovoj anketi.

Strategic Marketing Research garantuje i štiti vašu anonimnost. Podaci prikupljeni na ovaj način posmatraju se samo grupno i koristiće se jedino u svrhu ovog istraživanja. Ne postoji način da se bilo koji Vaš odgovor iz ove ankete poveže sa podacima o Vašem identitetu.

U slučaju pitanja molimo Vas kontaktirajte nas na broj 011 328 49 87, Natalija Biliškov.

Hvala Vam na saradnji!

Možemo li da počnemo?

Po dobijanju pristanka, početi sa razgovorom. Ako ispitanik ne želi da nastavi, treba mu / joj se zahvaliti, kod pitanja DO7 zaokružiti odgovarajuću šifru i preći na sledeći upitnik. Konsultujte se sa kontrolorom o ishodu i sledećoj poseti..

DEO DK – DEMOGRAFSKE KARAKTERISTIKE I SOCIOEKONOMSKI STATUS

DK1	**Koji je najviši stepen obrazovanja koji ste stekli do sada?** *[ANK] Pokazati karticu DK1.Jedan odgovor*	1. Bez škole 2. Nepotpuna osnovna škola 3. Osnovna škola 4. Srednja škola (3 ili 4 godine) 5. Viša škola 6. Visoka škola	DK2
DK2	**Koje je Vaše bračno stanje?**	1. Oženjen/udata 2. Živim u vanbračnoj zajednici 3. Neoženjen/neudata 4. Razveden/a, razdvojen/a 5. Udovac/udovica	DK3
DK3	**Koliko dece imate?**	_____	DK4
DK4	**Da li Vaše domaćinstvo ima više od jednog člana?**	1. Da, ____ članova	DK6
		2. Ne, živim sam/a	DK5
DK5	**Koliko dugo živite sami?**	_____godina	DK6
DK6	**Kakav je Vaš radni status?** *[ANK] Pokazati karticu DK6.Jedan odgovor*	1. Zaposlen/a 2. Samostalan/samozaposlen/a	DK7
		3. Penzionisan/a 4. Domaćica 5. Student, učenik 6. Nezaposlen/a 7. Nesposoban/na za rad	DEO HN
DK7	**Kojoj kategoriji zanimanja pripadate?** *[ANK] Pokazati karticu DK7.Jedan odgovor*	1. Zakonodavci, funkcioneri i rukovodioci 2. Stručnjaci 3. Stručni saradnici i tehničari 4. Službenici 5. Uslužni radnici i trgovci 6. Radnici u poljoprivredi, ribarstvu i šumarstvu 7. Zanatlije i srodni radnici 8. Rukovaoci mašinama i uređajima 9. Osnovna - jednostavna zanimanja 10. Vojna lica	DEO HN

DEO HN – HIGIJENSKE NAVIKE

HN1 **Da li perete ruke:**

	Skoro nikad	Kako – kad	Uvek
1. Po ulasku u kuću	1	2	3
2. Pre jela	1	2	3
3. Posle upotrebe WC-a (nužnika)	1	2	3

HN2

HN2	**Koliko često perete zube?** *[ANK] Pokazati karticu HN2. Jedan odgovor*	1. Nikad 2. Povremeno 3. Jednom dnevno 4. Više od jednom dnevno 5. Nemam svoje zube ni protezu	HN3
HN3	**Koliko puta ste se tokom prošle nedelje kupali ili tuširali?** *[ANK] Odnosi se na broj dana u nedelji. Pokazati karticu HN3. Jedan odgovor*	1. Nijednom 2. Jednom 3. 2 do 3 puta 4. 4 do 6 puta 5. Svaki dan	DEO IS

DEO IS - ISHRANA

IS1 **Koliko puta nedeljno:**

	Nikad	Ponekad	Svaki dan
1. Doručkujete	1	2	3
2. Užinate pre podne	1	2	3
3. Ručate	1	2	3
4. Užinate posle podne	1	2	3
5. Večerate	1	2	3

IS2

IS2	**Koliko čaša vode u proseku popijete u toku jednog dana?** *[ANK] Upišite tačan broj, ne pišite intervale. Misli se na obične čaše od 2dl obične, mineralne, gazirane ili negazirane vode.*	_____ čaša	IS3
IS3	**Da li pijete mleko, jogurt, kiselo mleko, belu kafu ili kakao?** *[ANK] Pokazati karticu IS3. Jedan odgovor.*	1. Nikad 2. Ponekad	IS5
		3. Svaki dan po jednu šolju 4. Svaki dan po 2 ili više šolja	IS4
IS4	**Koliki procenat masnoće ima mleko koje obično konzumirate?** *[ANK] Pokazati karticu IS4. Jedan odgovor.*	1. Ne obraćam pažnju na sadržaj masti 2. Manje od 0.5% masti (obrano) 3. 0.5% do 3.2% masti (delimično obrano) 4. Više od 3.2% masti (punomasno) 5. Ne pijem mleko	IS5

IS5 **Koliko često ste tokom prošle nedelje jeli ili pili***: [ANK] Odnosi se na broj dana u nedelji. Pokazati karticu IS5.*

	Nijednom	1 do 2 puta	3 do 5 puta	6 do 7 puta
1. Kuvan krompir	1	2	3	4
2. Pržen krompir	1	2	3	4
3. Pirinač/testenine	1	2	3	4
4. Žitarice (kuvano žito,mekinje,ovsene, kukuruzne i druge pahuljice,palenta/kačamak)	1	2	3	4
5. Sir	1	2	3	4
6. Ribu	1	2	3	4
7. Piletinu i ostala živinska mesa	1	2	3	4
8. Meso (juneće, svinjsko, jagnjeće)	1	2	3	4
9. Mesne prerađevine	1	2	3	4
10. Jaja	1	2	3	4
11. Pasulj, grašak, sočivo i slično	1	2	3	4
12. Sveže povrće, salatu	1	2	3	4
13. Drugo povrće (jela od povrća, smrznuto, konzervirano)	1	2	3	4
14. Sveže voće	1	2	3	4
15. Drugo voće (smrznuto, konzervirano)	1	2	3	4
16. Kolače, keks	1	2	3	4
17. Slatkiše (bombone, čokolade)	1	2	3	4
18. Slatka bezalkoholna pića (gazirane/negazirane sokove, toplu čokoladu)	1	2	3	4
19. Sendvič	1	2	3	4
20. Čips i druge grickalice	1	2	3	4
21. Hranu kupljenu u pekari (paštete, pogačice, burek, pica i sl.), kiosku, restoranu brze hrane	1	2	3	4

IS6

IS6	**Koju vrstu hleba najčešće koristite u ishrani?** *[ANK] Jedan odgovor*	1. Beli 2. Polubeli 3. Crni, ražani i slične vrste 4. Kombinovano	IS7
		5. Ne jedem hleb	IS8
IS7	**Koju vrstu masnih namaza najčešće mažete na hleb?** *[ANK] Jedan odgovor*	1. Margarin 2. Puter 3. Majonez 4. Mast 5. Kajmak/krem sir/pavlaku i slično 6. Paštetu 7. Ne koristim nikakav namaz	IS8
IS8	**Koja vrsta masnoće se <u>NAJČEŠĆE</u> koristi za pripremanje hrane (kuvanje, pečenje, priprema kolača i dr.) u Vašem domaćinstvu?** *[ANK] Jedan odgovor*	1. Svinjska mast, puter 2. Biljna mast, margarin 3. Ulje 4. Ne koristim masnoću *NZ (Ne zna)*	IS9
IS9	**Da li dosoljavate hranu koju jedete?** *[ANK] Jedan odgovor*	1. Nikad 2. Kada hrana nije dovoljno slana 3. Skoro uvek pre nego što probam hranu	IS10

IS10	**Da li pri izboru načina ishrane razmišljate o svom zdravlju?** *[ANK] Jedan odgovor*	1. Nikad 2. Ponekad 3. Često 4. Uvek	DEO SV

DEO SV - SLOBODNO VREME, FIZIČKA AKTIVNOST I SPORT

SV1 **Kako provodite slobodno vreme?**

	Nikad ili skoro nikad	Ponekad	Često
1. Gledam televiziju, DVD, video-kasete	1	2	3
2. Provodim vreme za kompjuterom	1	2	3
3. Provodim vreme sa decom	1	2	3
4. Provodim vreme sa prijateljima	1	2	3
5. Brinem o kućnom ljubimcu	1	2	3
6. Radim u kući i oko nje	1	2	3
7. Radim u polju, na njivi	1	2	3
8. Idem u bioskop, pozorište, na koncerte	1	2	3
9. Bavim se individualnim sportom (teretana, trčanje, tenis...)	1	2	3
10. Bavim se timskim sportom (fudbal, košarka, odbojka...)	1	2	3
11. Čitam knjigu	1	2	3
12. Čitam novine, časopise, rešavam ukrštenice	1	2	3
13. Igram društvene igre (šah, domine, karte, jamb...)	1	2	3
14. Imam aktivnosti vezane za dodatnu zaradu	1	2	3
95. Drugo, navesti: _____	1	2	3

SV2

SV2	**Koliko ste fizički aktivni u Vašem slobodnom vremenu?** *[ANK] Jedan odgovor. Pokazati karticu SV2. Ukoliko ispitanik navodi više od jednog odgovora, treba zaokružiti onaj koji se odnosi na najintenzivniju aktivnost, odnosno na aktivnost koja zahteva najveći fizički napor.*	1. Čitam, gledam televiziju, sedim, leškarim ili slično 2. Uglavnom hodam, vozim bicikl ili slično (šetanje, pecanje, lov) najmanje 4 sata nedeljno 3. Bavim se fizičkim aktivnostima radi održavanja fizičke kondicije (trčanjem, plivanjem, skijanjem, igrama loptom, težim radom u bašti i sl.) najmanje 4 sata nedeljno 4. Treniram redovno, nekoliko puta nedeljno	SV3
SV3	**Koliko često se, u slobodno vreme, bavite fizičkim aktivnostima bar 30 minuta tako da se bar malo zaduvate ili oznojite?** *[ANK] Jedan odgovor. Ukoliko je ispitanik starija osoba i odgovori da ne može da vežba jer je isuviše stara zaokružiti šifru 7.*	1. Svaki dan 2. 4 - 6 puta nedeljno 3. 2 - 3 puta nedeljno 4. Jednom nedeljno 5. 2 - 3 puta mesečno 6. Nekoliko puta godišnje/nikad 7. Ne mogu da vežbam zbog bolesti/invalidnosti	SV4
SV4	**Koliko je fizički naporan posao kojim se bavite?** *[ANK] Odgovaraju samo zaposlena lica i lica koja samostalno obavljaju neku delatnost (uključuje i zemljoradnike).*	1. Uglavnom sedim 2. Uglavnom stojim/hodam, ali ne nosim težak teret 3. Puno hodam, penjem se uz stepenice i/ili podižem teret 4. Teško fizički radim, nosim/podižem težak teret	SV5

*Razmislite o aktivnostima koje zahtevaju **veliki fizički napor**, a kojima ste se bavili u poslednjih nedelju dana**. Naporne fizičke aktivnosti** su one pri kojima dišete znatno teže nego obično i uključuju podizanje tereta, kopanje, aerobik ili brzu vožnju bicikla. Uzmite u obzir samo one fizičke aktivnosti koje su trajale najmanje 10 minuta u kontinuitetu.*

SV5	**SV5a. Koliko ste se u poslednjih nedelju dana bavili naprornim fizičkim aktivnostima?** *[ANK] Ukoliko je odgovor "Nijedan" upisati 0 dana.*	1. Upisati broj dana: _____dana *BO (Odbija da odgovori)* *NZ (Ne zna)*	SV5b SV6a
	SV5b. Koliko ste obično vremena u toku jednog dana proveli baveći se napornim fizičkim aktivnostima? *[ANK] Ukoliko je odgovor "Nijedan" upisati 0 sati i 0 minuta.*	1. Upisati broj sati _____ i _____minuta	SV6a

*Razmislite o aktivnostima koje zahtevaju **umeren fizički napor**, a kojima ste se bavili u poslednjih nedelju dana. **Umerene fizičke aktivnosti** su one pri kojima se malo zaduvate i uključuju nošenje lakših tereta, vožnju bicikla umerenom brzinom i sl. Nemojte uključivati hodanje.Uzmite u obzir samo one fizičke aktivnosti koje su trajale najmanje 10 minuta u kontinuitetu.*

SV6	**SV6a. Koliko ste se u poslednjih nedelju dana bavili umerenim fizičkim aktivnostima?** *[ANK] Ukoliko je odgovor "Nijedan" upisati 0 dana.*	1. Upisati broj dana: _____dana *BO (Odbija da odgovori)* NZ (Ne zna)	SV6b SV7a
	SV6b. Koliko ste obično vremena u toku jednog dana proveli baveći se umerenim fizičkim aktivnostima? *[ANK] Ukoliko je odgovor "Nijedan" upisati 0 sati i 0 minuta.*	1. Upisati broj sati _____ i _____minuta	SV7a

*Sada razmislite o tome koliko ste vremena u poslednjih nedelju dana proveli **hodajući**. Ovo uključuje hodanje do posla i nazad, hodanje od jednog do drugog mesta, kao i ono koje ste preduzeli samo zbog rekreacije, sporta, vežbanja ili razonode.*

SV7	**SV7a. U poslednjih nedelju dana, tokom koliko dana ste hodali najmanje 10 minuta u kontinuitetu?** *[ANK] Ukoliko je odgovor "Nijedan" upisati 0 dana.*	1. Upisati broj dana: _____dana *BO (Odbija da odgovori)* NZ (Ne zna)	SV7b SV8
	SV7b. Koliko ste obično vremena u toku jednog dana proveli hodajući? *[ANK] Ukoliko je odgovor "Nijedan" upisati 0 sati i 0 minuta.*	1. Upisati broj sati _____ i _____minuta	SV8

*Sada razmislite o tome koliko ste vremena proveli **sedeći** tokom radnih dana u poslednoj nedelji, uključujući vreme koji ste proveli na poslu/fakultetu i kod kuće. Ovo podrazumeva sedenje za stolom, sedenje i ležanje prilikom gledanja televizije/čitanja, posete prijateljima, putovanje kolima/autobusom.*

SV8	**U poslednjih nedelju dana, koliko ste ukupno vremena obično proveli sedeći u toku jednog RADNOG DANA?**	1. Upisati broj sati _____ i _____minuta	DEO PS

DEO PS - PONAŠANJE U SAOBRAĆAJU

PS1 **Ako vozite da li prilikom vožnje:**

	Ne	Ponekad	Da	Ne vozim
1. rolera/skejtborda koristite štitnike i kacigu	1	2	3	4
2. bicikla nosite zaštitnu kacigu	1	2	3	4
3. bicikla noću koristite prednje i zadnje svetlo	1	2	3	4
4. traktora noću koristite prednja i zadnja svetla	1	2	3	4
5. motora nosite zaštitnu kacigu	1	2	3	4
6. automobila koristite sigurnosni pojas	1	2	3	4
7. automobila vozite pod uticajem alkohola	1	2	3	4
8. automobila prekoračujete dozvoljenu brzinu	1	2	3	4
9. automobila koristite mobilni telefon	1	2	3	4

PS2

PS2	**Da li prelazite ulicu van pešačkog prelaza ili na crveno svetlo semafora za pešake?**	1. Ne 2. Da, ponekad 3. Da, često	PS3

PS3 **Da li kao suvozač koristite:**

	Ne	Ponekad	Da	Ne vozim se
1. u automobilu sigurnosni pojas	1	2	3	4
2. na motoru zaštitnu kacigu	1	2	3	4

DEO PO

DEO PO - POVREDE

PO1	**Da li ste se u toku prethodnih 12 meseci povređivali?**	1. Ne 2. Da	DEO OR PO2
PO2	**Gde se dogodilo poslednje povređivanje?**	1. U saobraćaju 2. U kući 3. Na poslu 4. U školi 5. Na ulici 6. U polju/na njivi 7. Na sportskom terenu 95. Drugo, navesti: _____	PO3

PO3	**Gde Vam je tom prilikom prvo pružena pomoć?**	1. Na licu mesta - hitna pomoć 2. U bolnici 3. U domu zdravlja/ambulanti 4. Kod privatnog lekara 5. Kod narodnog iscelitelja 95. Drugo, navesti: _____ 6. Nisam se obratio/la za pomoć	DEO OR

DEO OR – OSTALI RIZICI I ZNANJA O ZDRAVLJU

OR1 **Da li smatrate da u Vašem mestu postoje rizici po zdravlje:**

	Ne	Da	Ne znam	
1. Buka	1	2	NZ	
2. Zagađenje vazduha	1	2	NZ	
3. Zagađenje vode	1	2	NZ	
4. Otpadne materije	1	2	NZ	
5. Radioaktivno zračenje	1	2	NZ	
6. Ultraljubičasto (UV) / sunčevo zračenje	1	2	NZ	
7. Nasilje, kriminal	1	2	NZ	
95. Drugo, navesti: _____	1	2	NZ	OR2

OR2 **Da li smatrate da svojim ponašanjem rizikujete da obolite od neke od navedenih bolesti?**

	Ne	Da	Već oboleo/la	Ne znam	
1. Gojaznosti	1	2	3	NZ	
2. Povišenog krvnog pritiska	1	2	3	NZ	
3. Šećerne bolesti	1	2	3	NZ	
4. Bolesti srca i krvnih sudova (infarkt, šlog, angina pektoris)	1	2	3	NZ	
5. Plućnih bolesti (hronični bronhitis)	1	2	3	NZ	
6. Raka	1	2	3	NZ	
7. Ciroze jetre	1	2	3	NZ	
8. Polnih bolesti, side (AIDS)	1	2	3	NZ	
9. Bolesti kao posledice povrede	1	2	3	NZ	
95. Drugo, navesti: _____	1	2	3	NZ	OR3

OR3 **Da li Vam je, od strane lekara, otkriveno neko od sledećih stanja/oboljenja?**

	Otkriveno u toku prethodnih 12 meseci			Otkriveno ranije			Dani odsustva sa posla (bolovanja) u prethodnih 12 meseci	
Samo lica kod kojh stanje/oboljenje nije otkriveno u prethodnih 12 meseci odgovaraju i na opciju "otkriveno ranije".	Ne	Da, lečeno	Da, nije lečeno	Ne	Da, lečeno	Da, nije lečeno	*Odgovaraju samo zaposlena lica*	
1. Tuberkuloza	1	2	3	4	5	6	_____dana	
2. Infarkt miokarda (srčani udar)	1	2	3	4	5	6	_____dana	
3. Moždani udar (šlog)	1	2	3	4	5	6	_____dana	
4. Povišen krvni pritisak	1	2	3	4	5	6	_____dana	
5. Hronični bronhitis, emfizem	1	2	3	4	5	6	_____dana	
6. Astma	1	2	3	4	5	6	_____dana	
7. Maligno oboljenje (rak)	1	2	3	4	5	6	_____dana	
8. Šećerna bolest	1	2	3	4	5	6	_____dana	
9. Povišene masnoće u krvi	1	2	3	4	5	6	_____dana	
10. Migrena	1	2	3	4	5	6	_____dana	
11. Hronična anksioznost ili depresija	1	2	3	4	5	6	_____dana	
12. Oboljenje bubrega	1	2	3	4	5	6	_____dana	
13. Čir dvanaestopalačnog creva,želuca	1	2	3	4	5	6	_____dana	
14. Oboljenje žučne kese	1	2	3	4	5	6	_____dana	
15. Reumatska oboljenja zglobova	1	2	3	4	5	6	_____dana	
16. Osteoporoza	1	2	3	4	5	6	_____dana	
17. Alergija (bez astme)	1	2	3	4	5	6	_____dana	
18. Katarakta	1	2	3	4	5	6	_____dana	
19. Anemija	1	2	3	4	5	6	_____dana	OR4

OR4 **Da li ste u toku prethodne 4 nedelje imali sledeće simptome/probleme?**

[ANK] Čitajte ispitaniku simptom po simptom.

		Ne	Da	
1.	Bol u grudima tokom naprezanja	1	2	
2.	Bol u zglobovima	1	2	
3.	Bol u leđima	1	2	
4.	Bol u vratu/ramenima	1	2	
5.	Oticanje stopala	1	2	
6.	Proširene vene	1	2	
7.	Ekcem	1	2	
8.	Zatvor/Hemoroidi	1	2	
9.	Glavobolju	1	2	
10.	Nesanicu	1	2	
11.	Potištenost	1	2	
12.	Zubobolju	1	2	
13.	Bolove u celom telu	1	2	
14.	Nesvesticu	1	2	
15.	Učestalo mokrenje	1	2	OR5

OR5 **Kada Vam je poslednji put u nekoj od službi doma zdravlja pružena neka od navedenih usluga:**

[ANK] Čitajte ispitaniku uslugu po uslugu. Pokazati karticu OR5.

	Nikad	Pre više od 5 godina	Pre 1 - 5 godina	Tokom prethodnih 12 meseci	
1. Merenje krvnog pritiska	1	2	3	4	
2. Određivanje šećera u krvi	1	2	3	4	
3. Određivanje masnoća u krvi	1	2	3	4	
4. Određivanje hemoglobina u krvi	1	2	3	4	
5. Analiza mokraće	1	2	3	4	
6. Merenje telesne mase	1	2	3	4	
7. Kontrola vida	1	2	3	4	
8. Kontrola sluha	1	2	3	4	OR6

OR6 **Da li Vam je lekar rekao da imate povišen krvni pritisak?**

1. Ne — OR10
2. Da — OR7

OR7 **Da li lečite povišeni krvni pritisak?**

1. Da, samo dijetom — OR10
2. Da, samo lekovima
3. Da, na oba navedena načina — OR9
4. Ne, ne lečim se — OR8

OR8 **Navedite razlog:**

1. Nema potrebe
2. Nema lekova
3. Nemam novca
95. Drugo, navesti: _____ — OR10

OR9 **Da li ste u toku prethodne 4 nedelje uzimali lekove za lečenje visokog krvnog pritiska?**

1. Ne
2. Da, ponekad
3. Da, redovno

OR10

OR10 **Da li Vam je tokom prethodnih 12 meseci neka od navedenih osoba savetovala da:**

	Lekar, drugi zdravstveni radnik	Član porodice	Niko me nije savetovao	Nije bilo potrebno	
1. Manje jedete masno	1	2	3	4	
2. Manje koristite so	1	2	3	4	
3. Uzimate manje šećera	1	2	3	4	
4. Jedete više voća i povrća	1	2	3	4	
5. Smanjite težinu (oslabite)	1	2	3	4	
6. Povećate fizičku aktivnost	1	2	3	4	
7. Prestanete da pušite	1	2	3	4	
8. Pijete manje alkoholnih pića	1	2	3	4	OR11

OR11 **Da li pratite teme o zdravlju putem sledećih sredstava javnog informisanja?**

	Ne	Povremeno	Da
1. TV	1	2	3
2. Radio	1	2	3
3. Štampa	1	2	3
4. Internet	1	2	3

OR12

OR12 **Kakav uticaj na zdravlje po Vašem mišljenju imaju:**

	Veliki	Umeren	Mali	Ne znam/ nemam mišljenje
1. Ishrana	1	2	3	NZ
2. Fizička aktivnost	1	2	3	NZ
3. Pušenje	1	2	3	NZ
4. Konzumiranje alkohola	1	2	3	NZ
5. Društvene aktivnosti (druženje sa prijateljima, izlasci, izleti...)	1	2	3	NZ

OR13

OR13 **Da li ste u prethodnih 12 meseci:**

	Ne	Da	Nije bilo potrebno
1. Smanjili unos masnoća	1	2	3
2. Promenili vrstu masnoća u ishrani	1	2	3
3. Smanjili unos soli	1	2	3
4. Smanjili unos šećera	1	2	3
5. Povećali konzumiranje voća i povrća	1	2	3
6. Smanjili težinu (oslabili)	1	2	3
7. Povećali fizičku aktivnost	1	2	3
8. Prestali da pušite	1	2	3
9. Smanjili konzumiranje alkoholnih pića	1	2	3

OR14

OR14 **Ako ste promenili nešto u svom ponašanju u toku prethodnih 12 meseci, koji je bio najvažniji razlog?**

1. Zdravstveni (zbog bolesti)
2. Zbog lepote/izgleda
3. Zbog zdravijeg načina života
95. Drugo, navesti: _____
4. Nisam promenio/la ništa

OR15

OR15 **Šta su po Vašem mišljenju tri najvažnija razloga obolevanja stanovništva u našoj zemlji?**
[ANK] Pokazati karticu OR15 i dozvoljeno tri odgovora

-1- Pogrešna ishrana
-2- Stres
-3- Teški uslovi života
-4- Naporan rad
-5- Pušenje
-6- Nedovoljno bavljenje fizičkim aktivnostima
-7- Nedovoljno uzimanje vitamina, minerala
-8- Gojaznost
-9- Genetski (nasledni) faktori
-10- Alkohol
-11- Nedovoljna zdravstvena zaštita
-95- Drugo, navesti: _____

DEO ZZ

DEO ZZ - PROCENA ZDRAVLJA I ZADOVOLJSTVO ŽIVOTOM

ZZ1 **Kako biste ocenili svoje zdravlje u celini?**
[ANK] Jedan odgovor

1. Vrlo loše
2. Loše
3. Prosečno
4. Dobro
5. Vrlo dobro

ZZ2

ZZ2 **Kakvo je, prema Vašoj proceni, Vaše sadašnje zdravlje u odnosu na ono pre 12 meseci?**
[ANK] Jedan odgovor

1. Mnogo lošije
2. Nešto lošije
3. Uglavnom isto
4. Nešto bolje
5. Mnogo bolje

ZZ3

ZZ3 **Kako procenjujete svoju težinu?**
[ANK] Jedan odgovor

1. Mršav/a sam
2. Nisam ni debeo/la ni mršav/a
3. Debeo/la sam
4. Ne mogu da ocenim

ZZ4

ZZ4 **Kako procenjujete svoju fizičku aktivnost?**
[ANK] Jedan odgovor

1. Vrlo loša
2. Loša
3. Prosečna
4. Dobra
5. Vrlo dobra

ZZ5

ZZ5 **Kako procenjujete stanje svojih zuba?**
[ANK] Jedan odgovor

1. Vrlo loše
2. Loše
3. Prosečno
4. Dobro
5. Vrlo dobro

ZZ6

ZZ6 **Kakav je Vaš stav u vezi sa sopstvenim zdravljem?**

	Ne slažem se	Nisam siguran/a	Slažem se
1. Ja sam odgovoran/na za svoje zdravlje	1	2	3
2. Ako budem brinuo/la o sebi biću zdrav/a	1	2	3
3. Čak ako i budem vodio/la računa o sebi, lako ću se razboleti	1	2	3
4. Ako sam zdrav/a, to je čista sreća	1	2	3

ZZ7

ZZ7 **Poređajte navedene životne vrednosti od prvog do devetog mesta prema važnosti koje one imaju za Vas:**
(1 - najznačajnije, 2 - manje značajno, ...5,6,...
9 - najmanje značajno)

[ANK] Dajte ispitaniku kartice sa atributima i zamolite ga da na desnu stranu izdvoji najznačajniji atribut, a na levu stranu najmanje značajan atribut za njega. Potom neka između njih poređa ostale atribute po značajnosti.

1. Sreća u ljubavi	
2. Puno para	
3. Dobro zdravlje	
4. Zanimljiv posao	
5. Bavljenje politikom	
6. Biznis (unosan posao)	
7. Život u miru	
8. Sloboda	
9. Religija	

ZZ8

ZZ8 **Molim Vas da na skali od 0 do 10 ocenite zadovoljstvo svojim sadašnjim životom, gde 0 znači najgori život, a 10 znači najbolji život. Možete koristiti sve brojeve od 0 do 10.**
[ANK] Pokazati karticu ZZ8. Jedan odgovor.

Najgori život										Najbolji život
0	1	2	3	4	5	6	7	8	9	10

ZZ9

ZZ9 **Ako niste u potpunosti zadovoljni svojim sadašnjim životom, kojim delom ste najmanje zadovoljni?**

[ANK] Zaokružite najviše do tri odgovora. Pokazite karticu ZZ9.

-1- Bračnim partnerom
-2- Decom
-3- Roditeljima
-4- Prijateljima/komšijama
-5- Rođacima
-6- Kolegama na poslu
-7- Seksualnim životom
-8- Prihodima
-9- Svojim zdravljem
-10- Samoćom
-11- Samim sobom
-12- Poslom
-13- Političkom situacijom
- 95 - Drugo, navesti: _____
14. U potpunosti sam zadovoljan

ZZ10

ZZ10 **Da li postoji neka osoba na koju sigurno možete računati u svakoj situaciji (u slučaju razvoda, bolesti i dr.)?**

1. Ne
2. Da

NZ (Ne zna)

DEO MZ

DEO MZ - MENTALNO ZDRAVLJE

MZ1 **Da li ste bili napeti, pod stresom/pritiskom tokom prethodne 4 nedelje?**

1. Ne
2. Da, ponekad, ali ne više nego ostali ljudi
3. Da, više nego ostali ljudi
4. Da, moj život je skoro nepodnošljiv

MZ2

MZ2 **Da li ste tokom prethodne 4 nedelje imali emocionalne probleme (tuga, neraspoloženje, zabrinutost, potištenost)?**

1. Ne — MZ 5
2. Da — MZ 3

MZ3 **Ako je odgovor „da":**

	Ne	Da
1. Problemi su bili takvi da sam jedno vreme morao/la da skratim dužinu rada ili neku drugu aktivnost	1	2
2. Manje sam uradio/la nego što sam želeo/la	1	2
3. Nisam radio/la svoj posao ili bilo koju drugu aktivnost pažljivo kao obično	1	2

MZ 4

MZ4	**Da li su navedeni emocionalni problemi uticali na Vaše odnose u porodici, sa prijateljima, komšijama ili društvom?**	1. Nisu nimalo 2. Neznatno 3. Umereno 4. Veoma 5. Izuzetno su uticali	MZ 5

MZ5 **Koliko dugo ste se u toku prethodne 4 nedelje osećali na opisan način:**
[ANK] Pokazati karticu sa skalom MZ5. Pitati za sve opise iz tabele.

	Stalno	Najveći deo vremena	Dobar deo vremena	Neko vreme	Vrlo malo vremena	Nikad
1. Bio/la sam pun/a poleta	1	2	3	4	5	6
2. Bio/la sam veoma nervozan/a	1	2	3	4	5	6
3. Osećao/la sam se tako potišteno da ništa nije moglo da me oraspoloži	1	2	3	4	5	6
4. Osećao/la sam se spokojno i smireno	1	2	3	4	5	6
5. Osećao/la sam da imam puno energije	1	2	3	4	5	6
6. Bio/la sam tužan/a	1	2	3	4	5	6
7. Bio/la sam iscrpljen/a	1	2	3	4	5	6
8. Bio/la sam srećan/a	1	2	3	4	5	6
9. Osećao/la sam se umorno	1	2	3	4	5	6

DEO OA

DEO OA - MOGUĆNOST OBAVLJANJA AKTIVNOSTI U SVAKODNEVNOM ŽIVOTU

OA1	**Da li bolujete od neke dugotrajne bolesti ili imate nekih dugotrajnih zdravstvenih problema?**	1. Ne 2. Da	OA2
OA2	**Da li ste zbog zdravstvenih razloga, poslednjih 6 meseci ili duže, ograničeni u obavljanju uobičajenih aktivnosti (aktivnosti koje većina ljudi obično obavlja)?**	1. Ne 2. Da 3. Da, veoma	OA3

OA3 **Da li možete samostalno:** *[ANK] Čitati jednu po jednu tvrdnju i za svaku zaokružiti jedan odgovor.*

	Da, bez teškoća	Da, ali sa određenim teškoćama	Da, ali samo uz tuđu pomoć
1. Sesti i ustati sa stolice?	1	2	3
2. Leći i ustati iz kreveta?	1	2	3
3. Oblačiti se, svlačiti odnosno obuvati i izuvati?	1	2	3
4. Hraniti se i seći hranu u tanjiru?	1	2	3
5. Umiti se, oprati ruke?	1	2	3
6. Koristiti WC (nužnik) ?	1	2	3
7. Kupati se, tuširati?	1	2	3

OA 4

OA4	**Kakva je Vaša mogućnost kretanja?** *[ANK] Jedan odgovor*	1. Vezani ste za krevet 2. Krećete se uz pomoć invalidskih kolica	OA 8
		3. Krećete se uz pomoć pomagala (štap, štake, aparati, proteze) 4. Krećete se samostalno	OA 5
OA5	**Da li možete da pređete razdaljinu od 500m?** *[ANK] Jedan odgovor*	1. Da, bez teškoća 2. Da, ali uz manje teškoće	OA 7
		3. Da, ali uz velike teškoće 4. Ne, nisam u stanju	OA 6
OA6	**Koja je najveća daljina koju možete sami preći bez zaustavljanja i većeg zamaranja?** *[ANK] Jedan odgovor*	1. Nijedan korak 2. Samo nekoliko koraka 3. Više od nekoliko koraka, ali manje od 200 m 4. Više od 200 m, ali manje od 500 m	OA 7
OA7	**Da li možete da podignete i nosite 5 kilograma, na primer punu torbu namirnica?** *[ANK] Jedan odgovor*	1. Da, bez teškoća 2. Da, ali uz manje teškoće 3. Da, ali uz velike teškoće 4. Ne, nisam u stanju	OA 8

OA8 **Da li možete da:**

	Da, bez teškoća	Da, ali uz manje teškoće	Da, ali uz velike teškoće	Ne, nisam u stanju
1. Sa ili bez naočara/kontaktnih sočiva, prepoznate osobu na daljini od 4m?	1	2	3	4
2. Sa ili bez naočara/kontaktnih sočiva, čitate običan tekst u novinama?	1	2	3	4
3. Sa ili bez slušnog aparata, čujete tako da možete voditi razgovor sa jednom osobom?	1	2	3	4
4. Govorite?	1	2	3	4
5. Grizete i žvaćete čvrstu hranu (npr. jabuku)?	1	2	3	4

OA 9

OA9	**Da li primate invalidsku penziju?**	1. Ne 2. Da	DEO ZS

DEO ZS - KORIŠĆENJE ZDRAVSTVENE SLUŽBE I ZADOVOLJSTVO ZDRAVSTVENOM ZAŠTITOM

ZS1	**Kome se prvom obraćate kad imate zdravstveni problem?** *[ANK] Jedan odgovor*	1. Lekaru opšte medicine, medicine rada 2. Specijalisti 3. Privatnom lekaru 4. Narodnom iscelitelju (travaru, bioenergetičaru) 5. Nekom drugom (roditeljima, rođacima, prijateljima, deci) 6. Nikome, lečim se sam/a	ZS 2
ZS2	**Da li imate svog lekara (opšte medicine/medicine rada) ?**	1. Ne 2. Da	ZS 4 ZS 3
ZS3	**Da li ste zadovoljni Vašim lekarom?** *[ANK] Jedan odgovor. Pokazati karticu ZS3.*	1. Veoma sam nezadovoljan/na 2. Nezadovoljan/na sam 3. Nisam ni nezadovoljan/na ni zadovoljan/na 4. Zadovoljan/na sam 5. Veoma sam zadovoljan/na	ZS 4
ZS4	**Da li ste i koliko puta u toku prethodnih 12 meseci bili kod LEKARA OPŠTE MEDICINE/ MEDICINE RADA?**	1. Da, ____ puta 2. Bio/la sam pre više od godinu dana 3. Nikad nisam bio/la u životu	ZS 5 ZS 10
ZS5	**Koji je bio glavni razlog Vaše poslednje posete lekaru ?** *[ANK] Jedan odgovor*	1. Kontrola zdravlja **(kada ste bez tegoba),** sistematski pregled 2. Bolest, povreda 3. Samo da mi propiše lekove 4. Dobijanje potvrde 5. Dobijanje uputa za specijalistu, laboratoriju i slično 95. Drugo, navesti:_____	ZS 6

ZS6 **Koliko ste prilikom poslednjeg odlaska u prethodnih 12 meseci sa uputom lekara čekali na sledeće usluge u domu zdravlja?**

Pregledi	**Odmah sam primljen/a**	**Do nedelju dana**	**Do mesec dana**	**Više od mesec dana**	**Nisam koristio/la uslugu**
1. Laboratorijski pregled	1	2	3	4	5
2. Rendgen-preglede	1	2	3	4	5
3. EKG	1	2	3	4	5
4. Ultrazvuk	1	2	3	4	5

ZS 7

ZS7	**Da li ste i koliko puta u prethodnih 12 meseci bili kod LEKARA SPECIJALISTE (izuzimajući posete ginekologu) u domu zdravlja?**	1. Da, ____ puta 2. Bio/la sam pre više od godinu dana 3. Nikad nisam bio/la u životu	ZS 8 ZS 10

ZS8 **Koliko ste prilikom poslednjeg odlaska u prethodnih 12 meseci sa uputom za pregled specijaliste čekali da budete primljeni?**

Specijalista	**Primljen/a sam isti dan**	**Do nedelju dana**	**Do mesec dana**	**Više od mesec dana**	**Nisam bio/la**
1. Internista-kardiolog	1	2	3	4	5
2. Hirurg	1	2	3	4	5
3. Reumatolog	1	2	3	4	5
4. Urolog	1	2	3	4	5
5. Očni lekar	1	2	3	4	5
6. Ušni lekar	1	2	3	4	5
7. Neuropsihijatar	1	2	3	4	5
8. Fizijatar	1	2	3	4	5
95. Drugo, navesti:_____	1	2	3	4	5

ZS 9

ZS9 **Da li Vam je pri pružanju usluga traženo da sami nabavite:**

	Ne	Da	Nisam koristio/la	
1. Rendgen-filmove	1	2	3	
2. Reagense	1	2	3	
3. Lekove	1	2	3	
4. Sanitetski materijal	1	2	3	
5. Hirurški materijal	1	2	3	
95. Drugo, navesti:_____	1	2	3	ZS 10

ZS10 **Koji je glavni razlog što niste češće odlazili kod lekara?**
[ANK] Jedan odgovor

1. Bio/la sam zdrav/a
2. Nisam imao/la vremena
3. Gužva/dugo čekanje
4. Usluge lekara se plaćaju
5. Nemam poverenja u lekare
6. Daleko mi je
7. Odlazim često
95. Drugo, navesti: _____

ZS 11

ZS11 **Da li imate svog zubnog lekara (stomatologa)?**

1. Ne
2. Da

ZS 12

ZS12 **Da li ste i koliko puta u prethodnih 12 meseci bili kod zubnog lekara (stomatologa)?**

1. Da, ____ puta
2. Bio/la sam pre više od godinu dana

ZS 13

3. Nikad nisam bio/la u životu

ZS 14

ZS13 **Koji je glavni razlog Vaše poslednje posete zubnom lekaru (stomatologu)?**

1. Kontrola/savet/sistematski pregled.
2. Poliranje zuba i čišćenje kamenca
3. Problemi sa desnima/parodontopatija
4. Plombiranje zuba
5. Vađenje zuba
6. Zbog proteze
95. Drugo, navesti: _____

ZS 14

ZS14 **Koliko zuba Vam nedostaje?**

1. Nijedan

ZS 16

2. 1 - 5 zuba
3. 6 - 10 zuba
4. Više od 10 zuba, ali ne svi
5. Nemam nijedan zub

ZS 15

ZS15 **Da li imate protezu?**

1. Da, imam totalnu
2. Da, imam parcijalnu
3. Nemam protezu

ZS 16

ZS16 **Ukoliko ste u toku prethodnih 12 meseci koristili usluge hitne pomoći, navedite koliko ste poslednji put čekali na pomoć od trenutka poziva:**

1. Čekao/la sam ____ minuta
2. Ne sećam se
3. Nisam koristio/la usluge hitne pomoći

ZS 17

ZS17 **Da li ste tokom prethodnih 12 meseci koristili usluge privatnog lekara?**

1. Ne

ZS 20

2. Da

ZS 18

ZS18 **Koje specijalnosti je bio privatni lekar čije ste usluge koristili?**

	Ne	Da	
1. Lekar opšte medicine	1	2	
2. Stomatolog	1	2	
3. Ginekolog	1	2	
4. Internista	1	2	
5. Oftalmolog	1	2	
6. Hirurg	1	2	
7. Psihijatar	1	2	
8. Lekar neke druge specijalnosti	1	2	ZS 19

ZS19 **Navedite najznačajnije razloge zbog kojih ste koristili usluge privatnog lekara:**

[ANK] Zaokružiti najviše tri odgovora

-1- Kvalitetnije radi
-2- Nema čekanja
-3- Ljubazniji je
-4- Strpljiviji je
-5- Prinuđen/a sam jer određene preglede mogu da obavim samo kod privatnog lekara
-95- Drugo, navesti: _____

ZS 20

ZS20	**Da li ste i koliko puta u toku prethodnih 12 meseci bili na bolničkom lečenju? Izuzima se pratilac deteta i bolnički tretman vezan za porođaj**	1. Da, ____ puta 2. Bio/la sam pre više od godinu dana 3. Nikad nisam bio/la u životu	ZS 21 ZS20a

ZS20a.

☐ ***Ispitanik je žensko. ⇨ Preći na pitanje ZS25.***

☐ ***Ispitanik je muško. ⇨ Preći na pitanje ZS38.***

ZS21	**Koliko ste dugo čekali na prijem u bolnicu od trenutka kada ste dobili uput?** *[ANK] Ako ste bili više puta, ocenite poslednji prijem.*	1. Odmah sam bio primljen/a 2. Do nedelju dana 3. Do mesec dana 4. Više od mesec dana	ZS 22

ZS22 **Molimo Vas da ocenite boravak u bolnici :**

Predmet ocenjivanja	Loše	Osrednje	Dobro	
1. Čistoća bolničkih soba, posteljine	1	2	3	
2. Čistoća trpezarije	1	2	3	
3. Čistoća toaleta	1	2	3	
4. Kvalitet hrane	1	2	3	
5. Odnos lekara	1	2	3	
6. Odnos medicinskih sestara/tehničara	1	2	3	
7. Odnos ostalog osoblja	1	2	3	ZS 23

ZS23	**Da li ste bili zadovoljni bolničkim lečenjem?** *[ANK] Ako je bilo više bolničkih lečenja, neka ispitanik oceni poslednje bolničko lečenje. Jedan odgovor*	1. Veoma sam nezadovoljan/na 2. Nezadovoljan/na sam 3. Nisam ni nezadovoljan/na ni zadovoljan/na 4. Zadovoljan/na sam 5. Veoma sam zadovoljan/na	ZS 24
ZS24	**Kako se lečenje završilo?** *[ANK] Ako je bilo više bolničkih lečenja, neka ispitanik oceni poslednje bolničko lečenje. Jedan odgovor*	1. Ozdravio/la sam 2. Stanje se poboljšalo 3. Stanje je ostalo nepromenjeno 4. Stanje se pogoršalo 5. Ne mogu da ocenim	ZS24a

ZS24a.

☐ ***Ispitanik je žensko. ⇨ Preći na pitanje ZS25.***

☐ ***Ispitanik je muško. ⇨ Preći na pitanje ZS38.***

NA SLEDEĆIH 13 PITANJA ODGOVARAJU SAMO OSOBE ŽENSKOG POLA.

ZS25	**Da li ste i koliko puta u toku prethodnih 12 meseci bili kod ginekologa?**	1. Da, ____ puta 2. Bila sam pre više od godinu dana 3. Nikad nisam bila u životu	ZS 26 ZS 29
ZS26	**Koliko godina ste imali kada ste prvi put bili kod ginekologa?**	_____ godina	ZS 27
ZS27	**Koji je najčešći razlog Vaših poseta ginekologu?** *ANK] Jedan odgovor*	1. Tegobe 2. Kontracepcija 3. Kontrola zdravlja (kada ste bez tegoba), sistematski pregled 4. Trudnoća 5. Abortus 6. Sterilitet 95. Drugo, navesti: _____	ZS 28
ZS28	**Koliko često idete na ginekološke preglede, iako se osećate zdravi?**	1. Jedanput godišnje 2. Jedanput u 2 godine 3. Ređe 4. Ne idem	ZS 29
ZS29	**Da li ste do sada bili trudni?**	1. Ne 2. Da	ZS 33 ZS 30

ZS30	**Kakav je bio ishod Vaše poslednje trudnoće?**	1. Živorođeno dete/ca 2. Mrtvorođeno dete/ca 3. Spontani pobačaj 4. Namerni pobačaj 5. Trudnoća u toku	ZS 30a.
ZS30a.	**Da li je ste rađali u toku prethodnih 12 meseci?**	1. Ne	ZS32
		2. Da	ZS31

ZS31	**Posle Vašeg poslednjeg porođaja, da li Vas je u prvoj nedelji nakon izlaska iz porodilišta posetila u kući neka od sledećih osoba:**		**Ne**	**Da**	
		1. Doktor	1	2	
		2. Medicinska/patronažna sestra	1	2	ZS 32

ZS32	**Navedite ukupan broj namernih prekida trudnoća:**	_____	ZS 33
ZS33	**Da li bar jednom u toku meseca obavljate samopregled dojki?**	1. Ne 2. Da, na to me uputio lekar 3. Da, samoinicijativno	ZS 34
ZS34	**Kada Vam je poslednji put urađeno radiografsko snimanje dojki (mamografija)?**	1. U toku prethodnih 12 meseci 2. Pre 1 do 3 godine 3. Pre više od 3 godine 4. Pre više od 5 godina	ZS 35
		5. Ne sećam se 6. Nikad 7. Ne znam kakav je to pregled	ZS 36
ZS35	**Da li ste na mamografiju otišli:**	1. Samoinicijativno 2. Po savetu svog lekara 3. Po savetu lekara u okviru organizovanog ranog otkrivanja raka dojke	ZS 36
ZS36	**Kada Vam je poslednji put urađen Papanikolau test (test za procenu rizika od raka grlića materice)?**	1. U toku prethodnih 12 meseci 2. Pre 1 do 3 godine 3. Pre više od 3 godine 4. Pre više od 5 godina	ZS 37
		5. Ne sećam se 6. Nikad 7. Ne znam kakav je to test	ZS 38
ZS37	**Da li ste Papanikolau test uradili:**	1. Samoinicijativno 2. Po savetu svog lekara 3. Po savetu lekara u okviru organizovanog ranog otkrivanja raka grlića materice	ZS 38
ZS38	**Kakvo je u celini Vaše zadovoljstvo zdravstvenom službom?**	1. Veoma sam nezadovoljan/na 2. Nezadovoljan/na sam 3. Nisam ni nezadovoljan/na ni zadovoljan/na 4. Zadovoljan/na sam 5. Veoma sam zadovoljan/na	DEO LE

DEO LE - LEKOVI

LE 1	**Kako uzimate lekove?**	1. Po savetu lekara	LE3
		2. Samoinicijativno 3. Po savetu lekara i samoinicijativno	LE2
		4. Ne uzimam ih	Kraj

LE2	**Koje lekove uzimate samoinicijativno bez konsultacije, saveta lekara?**		**Ne**	**Da**	
		1. Vitamine, minerale (za jačanje organizma)	1	2	
		2. Biljne preparate (za jačanje organizma i lečenje bolesti)	1	2	
		3. Protiv bolova (glavobolje, zubobolje i sl.)	1	2	
		4. Za lečenje određenih bolesti (antibiotike, lekove za pritisak)	1	2	
		5. Za smirenje	1	2	
		6. Zbog nesanice	1	2	
		7. Za varenje	1	2	
		8. Protiv zatvora	1	2	
		9. Protiv začeća (kontraceptivna sredstva)	1	2	
		95. Drugo, navesti: _____	1	2	LE3

LE3	**Da li ste tokom prošle nedelje uzimali neke tablete, pilule ili druge oblike lekova?**		**Ne**	**Da**	
		1. Lekove za regulisanje krvnog pritiska	1	2	
		2. Lekove za regulisanje nivoa holesterola	1	2	
		3. Lekove za regulisanje nivoa šećera	1	2	
		4. Lekove protiv glavobolje i drugih bolova	1	2	
		5. Lekove protiv kašlja	1	2	
		6. Lekove za srce	1	2	
		7. Antibiotike	1	2	
		8. Sedative	1	2	
		9. Vitamine, minerale i slično	1	2	
		10. Kontraceptivna sredstva	1	2	LE4
LE4	**Kako najčešće nabavljate lekove?** *[ANK] Jedan odgovor*	1. Preko recepta 2. Kupujem ih u državnoj apoteci 3. Kupujem ih u privatnoj apoteci 4. Kupujem ih na pijaci 5. Dobijam ih od drugih (prijatelja, rođaka, roditelja, dece) 6. Ne mogu da ih nabavim jer ih nema 7. Ne mogu da ih nabavim jer su skupi 95. Drugo, navesti:_____			Kraj

DEO ON - OBJEKTIVNI NALAZ

ON1. TM (TELESNA MASA):	____
ON2. TV (TELESNA VISINA):	Visina u stojećem položaju (cm) ____
ON3. Identifikaciona šifra saradnika na merenju	Šifra saradnika merenja _____
ON4. Rezultat merenja:	1. Izmereno 2. Odsutno 3. Odbilo merenje 4. Nije bilo moguće izvršiti merenje *95. Drugo, navesti* _____

PRIMEDBE:

_____
_____
_____
_____
_____

ON5. Merenje krvnog pritiska: *[ANK] Vreme između merenja je 1 minut. Osoba kojoj se meri pritisak ne sme da menja položaj.*	**1. Sistolni krvni pritisak**	**2. Dijastolni krvni pritisak**
ON5_1. Merenje 1 **vreme prvog merenja: _ _:_ _ (čč/mm)**	I__I__I__I mm Hg	I__I__I__I mm Hg
ON5_2. Merenje 2	I__I__I__I mm Hg	I__I__I__I mm Hg
ON5_3. Merenje 3	I__I__I__I mm Hg	I__I__I__I mm Hg
ON6. Rezultat merenja:	1. Izmereno 2. Odsutno 3. Odbilo merenje 4. Nije bilo moguće izvršiti merenje *95. Drugo, navesti* _____	

PRIMEDBE:

_____
_____
_____
_____
_____

Datum

_____

IME I PREZIME ZDRAVSTVENOG RADNIKA

_____

	Anketu popunio:	1. Ispitanik 2. Anketar 3. Kombinovano 4. Član domaćinstva	
	PRIMEDBE ANKETARA		
	_____ _____ _____ _____ _____		
	Datum		
	IME I PREZIME ANKETARA		

UPITNIK ZA SAMOPOPUNJAVANJE ZA ODRASLE

DEO DO - INFORMACIONI PANEL UPITNIK ZA SAMOPOPUNJAVANJE ZA ODRASLE	
A Ovaj upitnik popunjava svaki član domaćinstva starosti 20 godina i više. *B Ovaj upitnik ispitanici popunjavaju samostalno i svi podaci su anonimni.* *C Potrebno je da se popuni poseban upitnik za svakog člana domaćinstva starosti 20 godina i više, koji živi u tom domaćinstvu.* *Upisati redni broj popisnog kruga u uzorku i redni broj domaćinstva u popisnom krugu, kao i broj reda ispitanika. Upisati ime i šifru anketara i datum anketiranja.* *Popuniti zajedno sa ispitanikom pitanja iz sekcije P – Probni deo, a zatim ispitanik nastavlja sam sa popunjavanjem upitnika.*	
DO1. Redni broj popisnog kruga u uzorku: ___ ___ ___	**DO2. Redni broj domaćinstva u popisnom krugu:** ___ ___
DO3. Broj reda člana domaćinstva: ___ ___ ___	**DO4. Dan / mesec / godina anketiranja:** ___ ___ / ___ ___ / **2 0 0 6**
DO5. Rezultat ankete rađene za samopopunjavanje: *[ANK] Šifre se odnose na ispitanika, tj. na to da li je ispitanik pristao na anketiranje.*	1. Upitnik je popunjen 2. Ispitanik nije kod kuće 3. Ispitanik odbija razgovor 4. Upitnik je delimično popunjen 5. Ispitanik nije u stanju da odgovara 95. Drugo, navesti: _____

Ponoviti uvodni pozdrav ukoliko to već nije učinjeno:

Poštovani,

Ovom prilikom želimo da Vam se najsrdačnije zahvalimo u ime istraživačke agencije Strategic Marketing Research što ste izdvojili svoje vreme i učestvovali u ovoj anketi.

Strategic Marketing Research garantuje i štiti vašu anonimnost. Podaci prikupljeni na ovaj način posmatraju se samo grupno i koristiće se jedino u svrhu ovog istraživanja. Ne postoji način da se bilo koji Vaš odgovor iz ove ankete poveže sa podacima o Vašem identitetu.

U slučaju pitanja molimo Vas kontaktirajte nas na broj 011 328 49 87, Natalija Biliškov.

Hvala Vam na saradnji!

P	**PROBNI DEO – ovaj deo popuniti zajedno sa ispitanikom**		
P1	**Da li lično poznajete osobe koje piju kafu?**	1. da 2. ne	P2
P2	**Da li ste vi lično ikada pili kafu?**	1. da → **pređite na pitanje P3 i dalje redom**	P3
		2. ne → **pređite na sledeću sekciju**	Sledeća sekcija
P3	**U kojoj godini života ste prvi put probali kafu?**	*Upišite godinu života u kojoj ste prvi put pili kafu - u _____ godini*	P4
P4	**Da li ste pili kafu tokom poslednjih 12 meseci?**	1. da 2. ne	P5
P5	**Da li ste pili kafu tokom poslednjih 30 dana?**	1. da 2. ne	P6
P6	**Koliko ste dana u proteklih 30 dana pili kafu?**	*Upišite broj dana, _____dana*	Sledeća sekcija

DEO PU - PUŠENJE

PU 1	**Da li Vi ili neko od članova Vašeg domaćinstva puši u kući?**	1. Ne, niko 2. Da, neko	PU2
PU2	**Koliko ste sati dnevno izloženi duvanskom dimu na Vašem radnom mestu?** **numeracija**	1. Više od 5 h 2. 1- 5 h 3. Manje od 1 h 4. Nisam izložen/a 5.....Ne radim van kuće	PU3
PU3	**Da li ste ikad pušili?**	1. Ne →**pređite na pitanje PU12**	PU12
		2. Da	PU4
PU4	**Da li ste tokom života popušili bar 100 cigareta?**	1. Ne 2. Da	PU5
PU5	**Da li ste ikad pušili svakodnevno (svaki dan u toku bar jedne godine)? Koliko godina ukupno?**	1. Ne 2. Da, ukupno _____ godina	PU6
PU6	**Da li sada pušite?**	1. Ne 2. Da, povremeno	PU7
		3. Da, svakodnevno →**pređite na pitanje PU8**	PU8
PU7	**Kada ste poslednji put pušili svakodnevno?**	1. Pre manje od 1 mesec 2. Pre 1 do 6 meseci 3. Pre 6 do 12 meseci 4. Pre 1 do 5 godina 5. Pre 5 do 10 godina 6. Pre više od 10 godina	PU8
		7. Nikad nisam pušio/la svakodnevno →**pređite na pitanje PU9**	PU9
PU8	**Koliko prosečno pušite u toku jednog dana, ili ste pušili pre nego što ste prestali da pušite svakodnevno?**	-1- Fabrički proizvedenih cigareta _____dnevno -2- Samostalno zavijene cigarete _____ dnevno -3- Lula duvana _____ dnevno -4- Cigare/cigarilosi _____dnevno	PU9
PU9	**Da li želite da prestanete da pušite?**	1. Ne 2. Da 3. Nisam siguran/na 4. Prestao/la sam	PU10
PU10	**Da li ste ikad ozbiljno pokušali da prestanete da pušite i niste pušili najmanje 24 sata?** **Ako je tako, kada poslednji put?**	1. Tokom prošlog meseca 2. Pre 1 do 6 meseci 3. Pre 6 do 12 meseci 4. Pre više od 12 meseci 5. Nikad	PU11
PU11	**Da li ste se obraćali za pomoć savetovalištu za odvikavanje od pušenja?**	1. Ne 2. Ne, nisam znao/la da postoji 3. Da	PU12
PU12	**Da li ste zabrinuti zbog štetnih posledica pušenja/duvanskog dima po Vaše zdravlje?**	1. Ne, nimalo 2. Ne previše 3. Da, pomalo 4. Da, veoma	DEO UA

DEO UA - UPOTREBA ALKOHOLA

UA1	**Koja se od navedenih izjava odnosi na Vas?** (IZABERITE JEDAN OD PONUĐENIH ODGOVORA).	1. Nikada nisam pio/la alkoholna pića (pivo, vino, žestoka pića, koktele i sl.) →**pređite na DEO UP** 2. Probao/la sam da pijem jednom ili dva puta →**pređite na DEO UP** 3. Pio/la sam, ali više ne →**pređite na DEO UP**	DEO UP
		4. Pijem alkoholna pića povremeno 5. Pijem alkoholna pića svakodnevno	UA2

UA2 **Koliko često sada pijete alkoholna pića** (POD OVIM SE PODRAZUMEVA I KAD POPIJETE SASVIM MALO)**:**

	Nikad	Nekoliko puta godišnje	2 – 3 puta mesečno	Jedanput nedeljno	2 – 3 puta nedeljno	Svaki dan
1. Pivo	1	2	3	4	5	6
2. Vino	1	2	3	4	5	6
3. Žestoka pića	1	2	3	4	5	6
4. Likere	1	2	3	4	5	6
5. Koktele	1	2	3	4	5	6

UA3

UA3	**Koliko ste čaša ili flaša sledećih pića popili tokom prošle nedelje?**	1. Flaša piva - 0,5 l |_|_| 2. Čaša vina - 0,2 l |_|_| 3. Čašica žestokog pića - 0,03 l |_|_| 4. Čašica likera - 0,03 l |_|_|		UA4
UA4	**Koliko često se dešava da popijete 6 ili više alkoholnih pića u toku jedne prilike?**	1. Nikad 2. Nekoliko puta godišnje 3. Jednom mesečno	4. Jednom nedeljno 5. Dnevno ili skoro svaki dan	DEO UP

DEO UP - UPOTREBA PSIHOAKTIVNIH SUPSTANCI

UP1 **Da li ste čuli za sledeća sredstva i šta mislite o njima?** (U SVAKOM REDU ZAOKRUŽITE JEDAN OD PONUĐENIH ODGOVORA)

	Nikad čuo/la	Čuo/la, ali ništa ne znam o njima	Bezopasna su ako se koriste pravilno	Uvek su štetna
1. Amfetamin (spid)	1	2	3	4
2. Barbiturati (sredstva za spavanje)	1	2	3	4
3. Kanabis (marihuana, hašiš)	1	2	3	4
4. Ekstazi	1	2	3	4
5. Kokain (koka)	1	2	3	4
6. Halucinogene droge (LSD)	1	2	3	4
7. Heroin	1	2	3	4
8. Morfijum	1	2	3	4
9. Krek	1	2	3	4
10. Rastvarači (lepak)	1	2	3	4
11. Sredstva za umirenje (npr. bensedin, librijum)	1	2	3	4
12. Sredstva protiv bolova (npr.trodon)	1	2	3	4
13. Kombinacija (npr. trodon i alkohol ili neka druga)	1	2	3	4

UP2

UP2 **Da li ste probali ili uzimate neko od navedenih sredstava?**

	Nikad	Probao/la 1 do 2 puta	Uzimao/la pre, sada ne	Uzimam povremeno	Uzimam svakodnevno
1. Lepak	1	2	3	4	5
2. Tablete (bensedin, trodon, amfetamin i dr.)	1	2	3	4	5
3. Marihuanu	1	2	3	4	5
4. Hašiš	1	2	3	4	5
5. Ekstazi	1	2	3	4	5
6. Kokain	1	2	3	4	5
7. Heroin	1	2	3	4	5

UP3

UKOLIKO NISTE NIKADA PROBALI NIJEDNO OD NAVEDENIH SREDSTAVA U PRETHODNOM PITANJU PRESKOČITE PITANJA UP3. I UP4. I PREĐITE NA DEO SP – SEKSUALNO PONAŠANJE

UP3	**Koliko ste imali godina kada ste prvi put probali:**	1. Lepak ....._____godina 2. Tablete (bensedin, trodon, amfetamin i dr.) ....._____godina 3. Marihuanu....._____godina 4. Hašiš....._____godina 5. Ekstazi....._____godina 6. Kokain....._____godina 7. Heroin....._____godina	UP4
UP4	**Gde ste prvi put probali neko od prethodno navedenih sredstava?**	1. Na žurci, u diskoteci, kafiću 2. Na ulici 3. U školi 4. U stanu svoga druga/rice ili svom stanu 95. Drugo, navesti _____	DEO SP

DEO SP - SEKSUALNO PONAŠANJE

SP1	**Da li ste stupili u seksualne odnose?**	1. Ne →**pređite na pitanje SP10**	SP10
		2. Da	SP2
SP2	**Sa koliko godina ste prvi put stupili u seksualne odnose?**	_____godina	SP3
SP3	**Da li ste imali/imate seksualne odnose sa osobom istog pola?**	1. Ne 2. Da	SP4
SP4	**Da li ste u prethodnih 12 meseci imali seksualne odnose?**	1. Ne →**pređite na pitanje SP10**	SP10
		2. Da	SP5
SP5	**Da li imate stalnog partnera (osobu sa kojom ste u bračnoj/vanbračnoj vezi)?**	1. Ne →**pređite na pitanje SP7**	SP7
		2. Da	SP6

SP6 **Da li Vi i Vaš stalni partner/ka koristite pri seksualnom odnosu neko od sredstava ili metoda za sprečavanje trudnoće (kontracepciju)?**

	Ne	Da, ponekad	Da, stalno
1. Pilulu	1	2	3
2. Intrauterinu spiralu	1	2	3
3. Lokalna hemijska sredstva (penu, AB film)	1	2	3
4. Kondom (prezervativ-gumicu)	1	2	3
5. Dijafragmu	1	2	3
6. Neplodne dane	1	2	3
7. Prekinut odnos	1	2	3
95. Drugo navesti šta_____	1	2	3

SP7

SP7	**Da li ste u prethodnih 12 meseci imali seksualne odnose sa osobom koja nije Vaš stalni partner?** (ISKLJUČUJE SEKS ZA NOVAC/USLUGU)	1. Ne →**pređite na pitanje SP10**	SP10
		2. Da	SP8
SP8	**Koliko takvih partnera ste imali u prethodnih 12 meseci?**	_____	SP9
SP9	**Da li je korišćen kondom prilikom poslednjeg seksualnog odnosa sa takvim partnerom?**	1. Ne 2. Da	SP10
SP10	**Da li ste čuli za virus koji se zove HIV i za bolest SIDU (AIDS) koju on izaziva ?**	1. Ne →**pređite na DEO NA**	DEO NA
		2. Da	SP11

SP11 **Šta mislite o sledećim izjavama?**

	Tačno	Netačno	Ne znam
1. Ljudi se mogu zaštititi od inficiranja HIV-om ako imaju samo jednog seksualnog partnera koji nije zaražen i nemaju druge partnere.	1	2	NZ
2. Ljudi se mogu zaštititi od inficiranja HIV-om pravilnom upotrebom kondoma prilikom svakog seksualnog odnosa.	1	2	NZ
3. Sida se može dobiti ujedom komarca.	1	2	NZ
4. Osoba koja izgleda zdrava može biti nosilac HIV-a.	1	2	NZ
5. Osoba se može inficirati HIV-om ako deli hranu sa inficiranom osobom.	1	2	NZ
6. HIV se može preneti sa majke na dete tokom trudnoće.	1	2	NZ
7. HIV se može preneti sa majke na dete prilikom porođaja.	1	2	NZ
8. HIV se može preneti sa majke na dete preko mleka prilikom dojenja.	1	2	NZ
9. Nastavniku koji ima virus, a još nije oboleo od side, treba dozvoliti da i dalje radi u školi.	1	2	NZ
10. I dalje treba kupovati hranu kod prodavca za koga ste saznali da ima sidu ili virus side.	1	2	NZ

SP12

SP12	**Da li ste se testirali na HIV?**	1. Ne →**pređite na pitanje SP14**	SP14
		2. Da	SP13
SP13	**Da li su Vam saopšteni rezultati?**	1. Ne 2. Da	SP14
SP14	**Da li znate mesto gde možete da se testirate na HIV?**	1. Ne 2. Da	DEO NA

DEO NA - NASILJE

NA1	**Da li ste u toku prethodnih 12 meseci bili izloženi nekom fizičkom nasilju?**		**Ne**	**Da**	
		1. U porodici	1	2	
		2. U školi/na radnom mestu	1	2	
		3. Na ulici	1	2	
		95. Drugo, navesti _____	1	2	NA2
NA2	**Da li ste u toku prethodnih 12 meseci bili izloženi nekom psihičkom maltretiranju (vređanju, ponižavanju, omalovažavanju, ismevanju, ucenjivanju...)?**		**Ne**	**Da**	
		1. U porodici	1	2	
		2. U školi/na radnom mestu	1	2	
		3. Na ulici	1	2	
		95. Drugo, navesti _____	1	2	NA3
NA3	**Ukoliko ste u toku prethodnih 12 meseci bili izloženi fizičkom nasilju ili psihičkom maltretiranju, da li ste se obraćali za pomoć?**		**Ne**	**Da**	
		1. Socijalnom radniku	1	2	
		2. Zdravstvenom radniku	1	2	
		3. SOS službi	1	2	
		4. Policiji	1	2	
		5. Roditelju, rođaku, prijatelju	1	2	
		6. Nastavniku, profesoru	1	2	
		95. Drugo, navesti_____	1	2	NA4
NA4	**Da li se do sada dešavalo da Vi nekoga :**		**Ne**	**Da**	
		1. Psihički maltretirate (vređate, ponižavate..)	1	2	
		2. Tučete	1	2	Kraj