1.Sadržaj

... while the individual man is insoluble puzzle, in the aggregate he becomes mathematical certainty.

Sherlock Holmes in "The sign of four"

1.Uvod

1.1.Klasifikacijone procedure u multivarijacionoj analizi

Multivarijacione metode su tehnike koje se koriste za analiziranje složenih skupova podataka. Njihova popularnost je sve veća jer one omogučavaju sprovođenje analize u slučaju kada postoji značajan broj nezavisnih i zavisnih promenljivih. Zbog teškoča rešavanja složenih istraživačkih zadataka sa metodama univarijacione analize i zbog dostupnosti statističkog softvera za sprovođenje multivarijacione analize, metode multivarijacione statistike su postale široko upotrebljavane.

Multivarijacioni modeli nisu slučajno stekli svoju popularnost. Njihova rastuća popularnost je rezultat i veće kompleksnosti savremenih istraživanja. U skoro svakoj disciplini, empirijska istraživanja gotovo nikada nisu ograničena samo na jednu zavisnu promenljivu.

Osnovne tehnike multivarijacione analize su: modeli multivarijacione linearne regresije, analiza glavnih komponenata, faktorska analiza, kanonička korelaciona analiza, diskriminaciona analiza i klaster analiza. U ovom radu analizirane su **tehnike za klasifikaciju podataka**, i to: **klaster analiza**, **diskriminaciona analiza** i **analiza klasifikacionog stabla**.

Da bi se dobila ideja za to kako, klaster analiza, diskriminaciona analiza i analiza klasifikacionog stabla imaju svoj udeo u generalnoj šemi analize multivarijacionih podataka, potrebno je proučiti: Tabelu 1.1. Ova tabela klasifikuje osnovne metode multivarijacione analize na osnovu odnosa koji se ispituju i tipa promenljivih.

Klaster analiza predstavlja grupu multivarijacionih tehnika čiji je primarni cilj da grupiše objekte na osnovu karakteristika koje oni poseduju. Poznata je i kao Q analiza, konstrukcija tipologije, analiza klasifikacije i numerička taksonomija. Različiti nazivi javljaju se zbog toga što metodi grupisanja imaju široku primenu u različitim disciplinama, kao što su psihologija, biologija, sociologija, ekonomija, inžinjerstvo i biznis. Iako su imena različita, svi metodi imaju zajednički cilj: klasifikaciju na osnovu odnosa između objekata koji su predmet grupisanja. Osnovna ideja klaster analize je da grupiše opservacije ili objekte u klastere, tako da su objekti u istom klasteru međusobno sličniji nego objekti koji se nalaze u ostalim klasterima. Ideja ove analize je da se maksimizira homogenost objekata unutar klastera, dok se u isto vreme maksimizira heterogenost između klastera.

Na Tabeli 1.1. klaster analiza spada u analize gde ne postoji uzročno - posledični odnos između zavisnih i nezavisnih promenljivih, već međuzavisnost između predmeta ili ispitanika.

Klaster analiza uzima u razmatranje opservacije i njihove karakteristike kao promenljive, ali ih ne razdvaja na zavisne i nezavisne promenljive. Na osnovu ovih promenljivih, sprovodi se grupisanje opservacija u zasebne grupe.

Tabela 1.1. Klasifikacija metode multivarijacione analize

Vid promenljive		**Kvantitativne promenljive**	**Kategorijske promenljive**
Metodi međuzavisnosti		▪ Glavne komponente ▪ Faktorska analiza ▪ Analiza grupisanja ▪ Kvantitativno višedimenziono proporcionalno prikazivanje	▪ Kategorijsko višedimenziono proporcionalno prikazivanje ▪ Loglinearni modeli
Metodi zavisnosti	**Jedna zavisna promenljiva**	▪ Višestruka korelacija ▪ Višestruka regresija	▪ Diskriminaciona analiza (zavisna promenljiva mora bidi kategorijska) ▪ Analiza klasifikacionog stabla ▪ Logit analiza
	Više zavisnih promenljivih	▪ Višedimenziona regresija ▪ Višedimenziona korelacija ▪ Kanonička korelaciona analiza	▪ Kanonička korelaciona analiza sa veštačkim promenljivama

Izvor: Multivarijaciona analiza, Kovačić Z., 1994, str.6.

Osnovni cilj **diskriminacione analize** je da oceni vezu između jedne kategorijske zavisne promenljive i skupa kategorijskih ili kvantitativnih nezavisnih promenljivih. Višestruka diskriminaciona analiza ima široku primenu u situacijama gde je primarni cilj da se identifikuje grupa kojoj objekat (na primer, osoba, kompanija ili proizvod) pripada. Potencijalna aplikacija ove tehnike je u predviđanju neuspeha ili uspeha novog proizvoda; u određivanju kreditnog rizika za jednu osobu; predviđanju da li će jedna kompanija biti uspešna. U svakom slučaju, objekti se klasifikuju u jednu grupu, a cilj je da se preko nezavisnih promenljivih koje je istraživač odabrao predvide i objasne razlozi zašto se objekat nalazi u određenoj grupi.

Na Tabeli 1.1. možemo uočiti da diskriminaciona analiza pripada grupi metoda gde se prikazuje odnos koji ima samo jednu kategorijsku zavisnu promenljivu i više nezavisnih promenljivih.

Analiza klasifikacionog stabla predstvalja metod kreiranja stabala klasifikovanja i odlučivanja, na osnovu kojih se sprovodi bolja identifikacija grupe, otkrivanje veze između grupa, i predviđanje budućih dešavanja. Upotrebom analize stabala klasifikovanja i odlučivanja mogu se doneti zaključci o segmentaciji, stratifikaciji, prognozi i redukciji podataka, zatim, mogu se proveriti promenljive, kao i identifikovati njihove interakcije. Osim toga može se izvršiti spajanje kategorija, kao i kategorizacija kvantitativnih promenljivih.

Na Tabeli 1.1. možemo uočiti da diskriminaciona analiza, kao i analiza klasifikacionog stabla pripada grupi sa jednom zavisnom i više nezavisnih promenjivih. Druga sličnost između ove dve analize je to da obe imaju kategorijske (ordinalne ili nominalne) zavisne promenljive. Ukoliko je zavisna promenljiva kvantitativna treba da se transformiše u kategorijsku promenljivu.

1.2.Pregled primena teorijskih modela u rešavanju empiriskih problema

Da bi se prikazala praktična upotreba klasifikacionih metoda, u radu su korištene tri različite baze podatka za različite probleme koji bi se mogli rešiti sa jednom od tri klasifikacione tehnike.

Prvo je prikazana klasifikaciona metoda **klaster analize**. U ovom delu su razmatrane tri metode klaster analize, i to: **hijerahijska klaster analiza**, k **– sredina klaster analiza** i **dvostepena klaster analiza**.

Za prikaz **hijerahijske klaster analize** izvršena je se klasifikacija opština Makedonije prema njihovim karakteristikama. Hijerahijska klaster analiza u softverskom paketu SPSS[1] prvo traži da istraživač izabere meru odstojanja, onda metod za povezivanje elementa kako bi se formirali klasteri, i da utvrdi koji broj klastera bi bio optimalan za korišćene podatke.

Cilj ove hijerahijske klaster analize je da se izvrši klasifikacija 84 opštine Makedonije na osnovu njihovih karakteristika, a preko dobijenih centroida klastera za selektovane promenljive kako bi se dobile informacije koje bi postale osnova za

[1] SPSS - od engleske reči “statistical software for social sciences”.

kreiranje nacionalne ekonomske i razvojne politike. To znači da bi se kreiralo nekoliko ekonomskih politika koje bi se primenile na grupe sastavljene od opština koje su međusobno slične. Tako, umesto da se svaka opština tretira zasebno, ili da se primenuje jedinstvena nacionalna politika, za svaku grupu opština dobijenu klaster analizom kreira se adekvatna razvojna politika. Na ovaj način potrebno je manje finansijskih resursa za kreiranja date politike, jer kreiranje razvojne politike posebno za svaku opštinu tražilo bi mnogo više finansiskih i ljudskih resursa, kao i vremena, nego kreiranje, tri ili četire razvojne politike.

Druga klaster metoda je k – **sredina klaster analiza**. Njena prednost u odnosu na hijerahijsku klaster analizu je to da hijerarhijska klaster analiza ne dozvoljava ponovo pregrupiranje objekata u drugi klaster, kad se već taj objekt svrsti u jedan klaster, i pored toga šta bi pregrupisavanje u novi klaster bilo bolje rešenje. Prednost k-sredina klaster analize je u tome što se pregrupisavanjem mogu dobiti bolja rešenja. Naravno, rezultati hijerahijske analize su vrlo važni, jer oni otkrivaju broj klastera i početne tačke klastera, koje su osnova za *k*-sredina klaster analizi. Za ovu analizu korišćena je ista baza podataka za makedonske opštine. Dobijeni rezultati mogu da se uporede i da se utvrdi specifičnost svakog klaster metoda.

Treća klaster metoda je **dvostepena klaster analiza**. Dvostepena klaster analiza kreira pod-klastere upotrebom hijerarhijske metode. Ukoliko se radi sa velikim bazama podataka, preporučuje se dvostepena klaster analilza, kao i u situaciji kada su posmatrane promenljive kategorijske. S obzirom da je baza podatka za makedonskih opština sastavljena od 82 opservacije, za ovu analizu koriščena je drugu veća baza podataka za 200 makedonskih kompanija. Cilj ove analize bio je da se dobiju klasteri i njihovi profili koji bi dali strukturu glavnih grupa makedonskih kompanija.

Prikaz klasifikacione tehnike **diskriminacione analize** dat je preko dva primera, od kojih jedan je za diskriminaciju dve grupe, a drugi je za diskriminaciju više grupa.

Za **diskriminacionu analizu dve grupe** primenićemo novu bazu podataka koja prikazuje grupu zemlje koje su, i grupu zemlje koje nisu članice Evropske unije i njihove osnovne ekonomske i demografske karakteristike. Cilj ove analize je da se sprovede diskriminaciona analiza za zavisnu promenljivu sa dve kategorije: *Zemlje članice Evropske Unije* i *Zemlje koje nisu članice Evropske Unije* na osnovu ekonomsko - demografskih karakteristika. Diskriminaciona analiza bi trebalo i da ukaže koja ekonomsko – demografska karakteristika najviše doprinosi razdvajanju

grupa, kako i da locira pripadnost posmatrane zemlje u jednu od dve razdvojene grupe.

Cilj **diskriminacione analize sa više grupa** je da se sprovede diskriminacija opština Makedonije prema kategorijskoj promenljivoj *Razvijenost* koja sadrži tri grupe, dobijene preko k-sredina klaster analize. Ciljevi ove analize su da se utvrde nezavisne promenljive koje najbolje razgraničavaju opštine i izvrši razvrstavanje opština u tri definisane grupe. Ukoliko diskriminaciona analiza daje dobre rezultate, to znači i da je podela opština po razvijenosti dobra i da se ista može koristiti pri kreiranju ekonomske politike zemlje za unapređivanje razvoja nerazvijenih i manje razvijenih opština.

Analiza klasifikacionog stabla koristi ponovo bazu podataka makedonskih opština sa ciljem da utvrdi pripadnost opština jednoj od tri kategorije zavisne promenljive – najrazvijenije opštine, srednje razvijene opštine i slabo razvijenije opštine. Ova analiza kreira i šemu klasifikacionog stabla iz koje se jasno može utvrditi struktura opština na osnovu njihovih karakteristika.

1.3.Osnovne hipoteze od kojih se polazi u istraživanju

U ovom radu se polazi od sledećih hipoteza, koje će se u toku istraživanja ispitati:

- Multivarijacione tehnike za klasifikaciju su korisni alat u rešavanju empiriskih ekonomskih problema i mogu dovesti do značajnih zaključaka čime predstavljaju bitno sredstvo kvantitativne analize.
- Multivarijacione tehnike klasifikacije, diskriminaciona analiza i klaster analiza, su korisne tehnike za rad sa bazama podataka.
- Primenom klaster analize moguće je izvršiti klasifikaciju 84 opštine Makedonije po njihovim karakteristikama, a preko centroida dobijenih klastera obezbediti informacije koje mogu da budu osnova za kreiranje različitih ekonomskih razvojnih politika.
- Klaster analiza se može koristiti za klasifikaciju 200 makedonskih kompanija da bi se dobila njihova osnovna struktura po veličini i delatnosti.
- Diskriminaciona analiza može da oceni vezu između kategorijske zavisne promenljive *Privredna razvijenost opština* i skupa kvantitativnih objašnjavajućih promenljivih makedonskih opština i da identifikuje grupe

kojima bi pripadale opštine. Diskriminaciona analiza uspešno klasifikuje zemlje članice Evropske unije i zemlje koje nisu članice Evropske unije preko osnovnih ekonomskih i demografskih indikatora. U primeru sa zemljama članicama Evropske unije koristimo diskriminacionu analizu - slučaj dve grupa, dok u primeru klasifikacije opština Makedonije koristimo diskriminacionu analizu - slučaj više grupa. Cilj korišćenja dve različite baze je da se prikaže široka upotrebljivost diskriminacione analize, i njena praktična primena u rešavanju sasvim različitih problema.

- Analizom klasifikacionog stabla možemo kreirati stablo klasifikovanja na osnovu koga se sprovodi bolja identifikacija grupa makedonskih opština. Ono takođe služi za otkrivanje postojećih veza između grupa.

Pored testiranja postavljenih hipoteza, ovaj rad ima za cilj i da stvori dobru analitičku osnovu za buduća istraživanja u ovoj oblasti. U istraživanju koristimo statistički softer SPSS.

From a drop of water, a logician could infer the possibility of an Atlantic or a Niagara without having seen or heard of one or the other. So all life is a great chain, the nature of which is known whenever we are shown a single link of it.

Sherlock Holmes in "Study in Scarlet"

2.Klaster analiza

2.1.Klaster analiza – Uvod

Klaster analiza pretstavlja skup multivarijacionih tehnika čiji je osnovni cilj da grupiše različite objekte na osnovu karakteristika koje poseduju, odnosno osnovna funkcija klaster analize je identifikacija grupa ili klastera. Cilj ove analize je da podeli skup objekta u određeni broj grupa ili klastera, tako da svi objekti jednoga klastera budu slični, dok se objekti pripadnici različitih klastera međusobno razlikuju.

Podatke iz matrice $Y = \left(y_{ij}\right)$ dimenzija $n \times p$, možemo predstaviti kao:

$$\underset{n\times p}{Y} = \begin{bmatrix} y_1^{'} \\ y_2^{'} \\ \vdots \\ y_n^{'} \end{bmatrix}$$

Cilj klaster analize je da razvije klasifikacioni algoritam koji će podeliti vrste matrice Y u k različitih grupa, odnosno klastera. Vrste matrice pretstavljaju objekte ili opservacije.

Tradicionalna upotreba klaster analize je u funkciji istraživačkih ciljeva i formiranja **taksonomije**, odnosno empirijski osnovana klasifikacija objekta. Klaster analiza može se koristi i pri formulisanju hipoteza u vezi strukture objekata. I pored toga što se ova tehnika smatra istraživačkom, ista nalazi primenu u slučajevima gde se predložena tipologija (na teoriji bazirana klasifikacija) upoređuje sa tipologijom dobijenom klaster analizom. U tom slučaju, ova se tehnika koristi kao potvrdna ili konfirmatorna tehnika.

Klaster analiza se koristi i za **simplifikaciju podataka**, tako što se analiziraju grupe sličnih objekata umesto svi individualni objekti. Uprošćena struktura dobijena iz klaster analize pronalazi odnose koji se ne mogu drugačije identifikovati.

Mogućnosti primene klaster analize su velike, i ista nalazi primenu u rešavanju različitih problema u brojnim disciplinama. I pored velikog broja prednosti ove analize, najčešće **kritike** ukazuju na to da je klaster analiza **deskriptivna** i da **nije zasnovana na statističkom zaključivanju**. To nije teško zaključiti kada je poznato da ova analiza nema statističku osnovu na bazi, bi se zaključci dobijeni na osnovu uzoraka, mogli aplicirati na celu populaciju. Rezultat ove analize se ne može generalizovati jer zavisi od promenljivih koje se koriste u analizi i posebno je osetljiv

na promene posmatranih promenljivih. Klaster analiza je istraživačka metodologija za analizu podataka. Ona zavisi od stepena slučajnog šuma[2] podataka, nestandardnih opservacija, izbora promenljivih u analizi i korišćenih mera bliskosti. Ne postoji jedinstven niti optimalan metod za analizu podataka primenom klaster analize.

2.2.Istraživački dizajn u klaster analizi

Pre početka procesa deljenja objekata ili opservacija, odnosno sprovođenja klaster analize, potrebno je definisati ciljeve istraživanja i sprovesti selekciju promenljivih. Nadalje, primena određene metode klaster analize, (dvostepena klaster analiza i klaster analiza k-sredina), zavisi od rasporeda objekata u bazi podataka. Različit raspored podataka daje različite rezultate, pa je potrebno podatke rasporediti metodom slučajnog izbora. Što je baza podataka manja problem redosleda objekata je veći, čak i ako su raspoređeni na slučajan način. Zato, pri sprovođenju analize, predlaže se preraspodela objekata.

Često klaster analizu moramo da primenimo u analizi numeričke promenljive. Ukoliko je bar jedna promenljiva kategorijska, onda se preporučuje dvostepena klaster analiza.

Potrebno je i odgovoriti na dva vrlo važna pitanja, da li postoje nestandardne opservacije i da li iste treba isključiti, kao i da li je potrebna standardizacija promenljivih. Nestandardne observacije mogu da budu opservacije koje nisu reprezentativne za populaciju, a koje su reprezentativne opservacije malih ili neznačajnih segmenta ili observacije koje nisu dovoljno zastupljene u uzorcima, ali pretstavljaju validne i relevantne grupe. Zbog ovih razloga, neophodan je preliminarni pregled podataka.

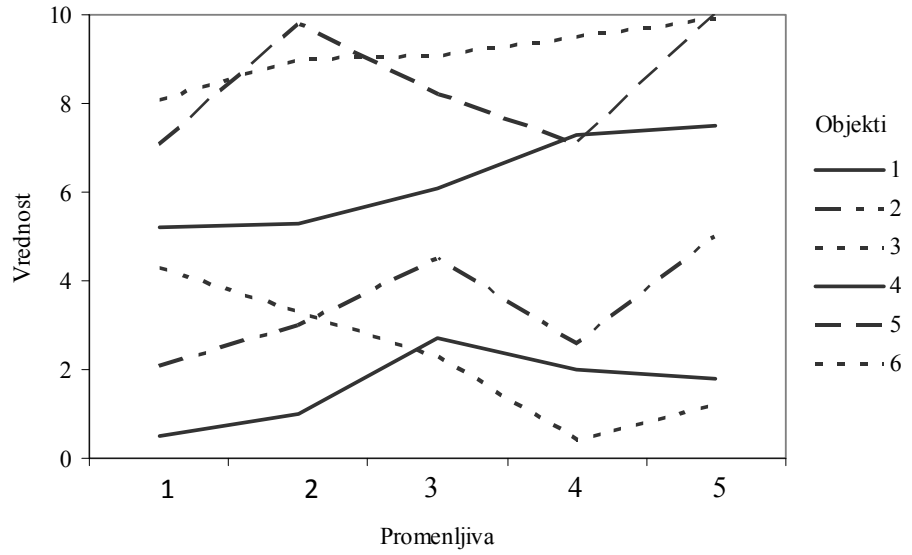
2.2.1.Nestandardne opservacije

Klaster analiza je veoma osetljiva na nestandardne opservacije, kao i na unos irelevantnih promenljivih u analizu. Grafički pristup pri skeniranju nestandarnih opservacija koristi takozvani dijagram profila (Slika 2.1.), gde se promenljive postavljuju na horizontalnu osu, dok se vrednosti promenljivih nalaze na vertikalnoj osi. Ovako se formiraju takozvani profili svih objekata i isti se prikazuju na grafikonu,

[2] Od engleske reči "random noise".

gde svaka linija predstavlja jedan objekat. Na Slici 2.1. je prikazan primer sa 6 opservacija. Nestandardne opservacije najčesce imaju ekstremne vrednosti kod jedne ili više promenljivih.

Slika 2.1. Dijagram profila



I pored toga što je grafički pristup jednostavan, postaje neefikasan za veliki broj opservacija. U ovom slučaju najadekvatnija je upotreba multivarijacione detekcije nestandardnih opservacija. Kada imamo više od dve promenljive, potrebno je da se objektivno izmeri multidimenzionalna pozicija svake opservacije u odnosu na neku zajedničku tačku. Za to koristimo Mahalanobisovu D^2 meru, koja meri rastojanje svake opservacije u multidimenzionalnom prostoru od središnjeg ili prosečnog centra opserviranih promenljivih. Visoka vrednost D^2 mere ukazuje na nestandardne opservacije.

Da bi definisali ovu meru, možemo pretpostaviti da imamo slučajnu promenljivu Y koja ima normalni raspored sa sredinom nula i varijansom jedan. Ukoliko imamo dve vrednosti promenljive, y_i i y_j, da bi uporedili rastojanje, potrebno je uzeti u obzir i varijansu slučajne promenljive. Tako je kvadratno rastojanje između y_i i y_j definisano kao:

$$D_{ij}^2 = (y_i - y_j)^2 / \sigma^2 = (y_i - y_j)(\sigma^2)^{-1}(y_i - y_j)$$

gde σ^2 je varijansa populacije.

2.2.2. Standardizacija promenljive

Drugo bitno pitanje u pripremi podataka za klaster analizu je dali je potrebno izvršiti standardizaciju podataka. Najčešća forma korišćena za standardizaciju promenljivih je svođenje svake promenljive na standardizovanu promenljivu sa Z vrednostima, gde se odstupanje svake vrednosti promenljive od sredine promenljive i deli sa standardnom devijacijom.

Preko standardizacije promenjive eliminiše se efekat različitih mernih skala. Potreba za standardizacijom se minimizira kada su sve promenljive prikazane na istoj mernoj skali, ali je standardizacija posebno važna kada se za promenljive koriste različite merne skale.

2.3.Pretpostavke klaster analize

Klaster analiza nije multivarijaciona tehnika koja se zasniva na statističkom zaključavanju, gde se preko statistike uzoraka ocenuju nepoznati parametri populacije. Klaster analiza je metod za kvantifikovanje strukturnih karakteristika jednog skupa objekta, opservacija. Ovaj metod, poseduje matematička svojstva, ali ne i statističku osnovu. Uslovi normalnosti i linearnosti koji su vrlo važni kod drugih tehnika, imaju vrlo mali značaj u klaster analizi. Zato je za klaster analizu bitno da se odredi da li:

- koristimo **reprezentativan uzorak;**
- postoji **multikolinearnost promenljivih;**
- postoji **autokorelacija objekata**.

Samo u retkim slučajevima klaster analiza se može sprovesti na podacima iz cele populacije. Uglavnom, klaster analiza se sprovodi na podacima iz uzoraka, tako da uzorak treba da bude reprezentativan za celu populaciju. Ovo je vrlo važno uslov,, jer samo u slučaju **reprezentativnih uzoraka** možemo generalizovati zaključke dobijene iz uzoraka na celu populaciju.

Multikolinearnost prestavlja stepen povezanosti nezavisne promenljive sa ostalim nezavisnim promenljivim. Povećanje multikolinearnosti smanjuje ili isključuje mogućnost definisanja uticaja jedne promenljive u određenoj analizi. U klaster analizi, efekat multikolinearnosti je sasvim drugog oblika jer multikolinearnost

predstavlja oblik indirektnog pondera. Multikolinearnost deluje ponderacioni proces koji nije očigledan, ali utiče na analizu. Zato je potrebno da se izvrši analiza za značajnosti multikolinearnosti, ukoliko ona postoji, ili da se redukuju promenljive u jednu promenljivu za svaki skup koreliranih promenljivih ili da se koristi mera odstojanja koja kompenzira korelaciju, što je Mahalanobisovo rastojanje.

Pretpostavka o multikolinearnost može se testirati preko kreiranja bivarijacione korelacione matrice za kvantitativne promenljive da bi se utvrdilo da li se koeficient korelacije značajno razlikuju od nule.

Pretpostavlja se da su objekti ili opservacije međusobno nezavisni. Vrednosti koje uzima objekat k ne bi trebalo da utiču na vrednosti za objekat $k+1$, odnosno, nema **autokorelacije objekata**.

2.4.Mere bliskosti

Koncept bliskoti je osnovni koncept klaster analize. Bliskost između objekata je empirijska mera korespodencije između objekata koje bi trebalo grupisati u klastere. Proces analize se odvija tako što se mera bliskosti izračunava za sve parove objekata, gde se bliskost bazira na profilu svake opservacije za karakteristike (promenljive) izabrane od strane istraživača. Na ovaj način se svaki objekat upoređuje sa bilo kojim drugim objektom preko mere bliskosti. Procedura klaster analize nastavlja da grupiše slične objekte u klastere.

Bliskosti može da se iskaže preko sličnosti ili različitosti. Ako mera bliskosti prikazuje sličnost, vrednost mere se povećava kada su dva objekta sličnija. Suprotno, ako mera bliskosti označava različitost, vrednost mere se smanjuje kad su dva objekta sličnija.

2.4.1.Mere različitosti

Neka y_i i y_j predstavljaju dva objekta u prostoru sa p – promenljivih. Mera različitosti zadovoljava sledeće uslove:

1) $d_{ij} \geq 0$ za sve objekte y_i i y_j;

2) $d_{ij} = 0$ ako i samo ako $y_i = y_j$;

3) $d_{ij} = d_{ji}$.

Prvi uslov ukazuje da mera nikad nije negativna. Drugi uslov ukazuje na to da je mera jednaka nuli kada su objekti jednaki međusobom, odnosno objekti su jednaki samo kada $d_{ij} = 0$ i u nijednoj drugoj situaciji. Treći uslov ukazuje da je mera simetrična, tako da mera različitosti koja upoređuje y_i sa y_j je ista kao i mera različitosti koja upoređuje objekte y_j i y_i. Mera različitosti koja zadovoljava ove uslove je polumetrička mera.

Za numeričke promenljive merene minimum na intervalnoj skali, najčesta mera različitosti je Euklidovo rastojanje među dva objekta. Ukoliko imamo matricu Y dimenzija $(n \times p)$ i $y_i^{'}$ vektor vrste dimenzija $(1 \times p)$, **kvadratno Euklidovo rastojanje** između dve vrste y_i i y_j je definisano kao:

$$d_{ij}^2 = (y_i - y_j)^{'}(y_i - y_j) = \left\| y_i - y_j \right\|^2 .$$

Matrica podataka $D = \left(d_{ij}\right)$ dimenzija $(n \times n)$ zove se **Euklidova matrica rastojanja**. Moguće je da različite merne jedinice promenljivih, utiču na to da određena promenljiva dominira u kvantifikovanju rastojanja. Euklidova matrica rastojanja je najefikasnija kada se promenljive iskazuju na istoj mernoj skali.

Ukoliko su promenljive različite i koriste različite merne skale, moguće je izvesti ponderaciju kvadratnih razlika preko $s_r^2 = \sum_{r=1}^{n} (y_{rs} - \bar{y}_{.s})^2 / (n-1), s = 1,2,\ldots, p$ gde s_r^2 i $\bar{y}_{.j}$ predstavljuju varijansu promenljive s i ocene sredine:

$$d_{ij}^2 = (y_i - y_j)^{'} (diag\ S)^{-1} (y_i - y_j) .$$

Ovaj proces eliminiše zavisnost analize od merne skale. Ali vrlo često prouzrokuje da rastojanja u klasterima budu veća od rastojanja među klasterima, i na ovaj način dolazi do prekrivanja klastera.

Euklidovo rastojanje je poseban slučaj **metrike Minkowski**, gde se mere različitosti mogu predstaviti kao:

$$d_{ij} = \left(\sum_{s=1}^{p} \left| y_{is} - y_{js} \right|^{\lambda} \right)^{1/\lambda} .$$

Za $\lambda = 1$ imamo meru zvanu **rastojanje gradskih blokova**, dok za $\lambda = 2$ imamo Euklidovo rastojanje. Rastojanje gradskih blokova je mera koja je manje osetljiva na nestandardne observacije.

U analizi se koriste još dve mere različitosti, i to sledeće:

Metrika Kanbera

$$d_{ij} = \sum_{s=1}^{p} \left\{ \frac{|y_{is} - y_{js}|}{(y_{is} + y_{js})} \right\},$$

Čekanovski koeficient

$$d_{ij} = \frac{\sum_{s=1}^{p} |y_{is} - y_{js}|}{\sum_{s=1}^{p} (y_{is} + y_{js})}.$$

Ove se mere koriste kada su podaci asimetrični ili/i kada postoje nestandardne observacije.

Prikazane mere se koriste u situaciji kada su promenljive kvantitativne. Za kategorijski tip podataka koji se mere na nominalnim ili ordinalnim skalama, situacija je složenija. U jednostavnom slučaju, pretpostavlja se da svaki red $y_i^{'}$ matrice Y sadrži samo binarne podatke. U tom slučaju, kvadratno Euklidovo rastojanje broji parove koji sadrže različite binarne vrednosti, (1 – 0) ili (0 – 1), dok parovi koji imaju iste binarne vrednosti, (1 – 1) ili (0 – 0), tretira jednako. Kada je promenljiva kodirana sa 0 ili 1, to ukazuje na odsustvo ili prisutnost određene karakteristike. Moguće kombinacije su prikazane u Tabeli 2.1.

Tabela 2.1. Raspored parova vrednosti binarnih promenljivih (tabela kontingencija)

		Objekat (opservacija) *j*		Ukupno
		1	0	
Objekat *i*	1	a	b	a+b
	0	c	d	c+d
Ukupno		a+c	b+d	p=a+b+c+d

Frekvencije b i c predstavljaju parove različitih binarnih vrednosti, dok frekvencije a i d predstavljaju parove koji imaju iste binarne vrednosti. Tako kvadratno Euklidovo rastojanje podeljeno sa p postaje:

$$\sum_{s=1}^{p}\left(y_{is}+y_{js}\right)^{2}/p=(b+c)/p=d_{rs}^{2}/p$$

za p binarnih promenljivih. Ukoliko se koristi metrika Minkowski, vrednost je ista za sve $\lambda \geq 1$. Isto tako, $(b+c)/p=1-(a+d)/p$. Veličina $(a+d)/p$ je mera sličnosti jer prikazuje proporciju parova koji imaju iste vrednosti binarnih promenljivih između dva binarna vektora sa dimenzijama $(1\times p)$. Za binarne promenljive, metrika Kambera je identična sa metrikom gradskih blokova.

Za evaluaciju metrike Čekanovski za binarne promenljive, sledi

$$d_{ij}=\frac{\sum_{s=1}^{p}\left|y_{is}-y_{js}\right|}{\sum_{s=1}^{p}(y_{is}+y_{js})}=\frac{b+c}{(a+b)+(a+c)}=1-\frac{2a}{2a+b+c}.$$

Veličina $2a/(2a+b+c)$ predstavlja koeficijent Čekanovski. Ovo je mera sličnosti gde se dvojni ponder dodeljuje (1 – 1) parovima, dok (0 – 0) parovi su isključeni iz imenioca i brojioca, čime postaju irelevantni za obračun.

2.4.2.Mere sličnosti

Za dva objekta y_i i y_j u p dimenzionalnom prostoru, mera sličnosti zadovoljava sledeće uslove:

1) $0 \leq s_{ij} \leq 1$, za sve objekte y_i i y_j;

2) $s_{ij}=1$ ako i samo ako $y_i=y_j$;

3) $s_{ij}=s_{ji}$.

Uslovi jedan i dva ukazuju na to da je mera uvek pozitivna i jednaka jedinici samo ako su objekti i i j identični, dok treći uslov ukazuje na to da je mera simetrična.

Ukoliko postoji mera sličnosti koja zadovoljava navedene uslove, uvek je moguće da se posmatra i mera različitosti, odnosno $d_{ij}=1-s_{ij}$. I suprotno, ukoliko je poznata mera različitosti, d_{ij}, moguće je da se konstruira mera sličnosti kao $s_{ij}=1/(1+d_{ij})$. Na ovaj način je moguće dobiti meru s_{ij} u zavisnosti od mere d_{ij}. Moguće je korišćenje mere sličnosti kao i mere različitosti u klaster analizi, da bi se grupisali vrste matrice podataka.

Poznata mera sličnosti je **Pirsonov koeficent korelacije** između objekta y_i i y_j, $i, j = 1,2,\ldots,n$, definisan kao:

$$q_{ij} \frac{\sum_{s=1}^{p}(y_{is} - \bar{y}_{i.})(y_{js} - \bar{y}_{j.})}{\left[\sum_{s=1}^{p}(y_{is} - \bar{y}_{i.})^2 \sum_{s=1}^{p}(y_{js} - \bar{y}_{j.})^2\right]}$$

gde $y_{i.} = \sum_s y_{is} / p$ i $y_{j.} = \sum_s y_{js} / p$. Koristi se simbol q_{ij} jer se izračunavaju korelacije redova matrice podataka. Ali, zbog $-1 \le q_{ij} \le 1$, prvi uslov nije ispunjen. Da bi se izvršila korekcija, koriste se i sledeće vrednosti $|q_{ij}|$ ili $1 - q_{ij}^2$.

Tabela 2.2. Mere sličnosti na bazi binarnih promenljivih

Mera sličnosti	Ponderacija istih parova i različitih parova kod binarnih promenljivih	Ime koeficienta
$\frac{a+d}{p}$	Jednaki ponderi uparivanih parova sa (0 – 0) parovima	Jednostavno uparivanje
$\frac{2(a+d)}{2(a+d)+b+c}$	Dvojni ponderi uparivanih parova sa (0 – 0) parovima	Dvojno uparivanje
$\frac{a+d}{a+d+2(b+c)}$	Dvojni ponderi neuparivanih parova sa (0 – 0) parovima	Rodžers-Tanimoto
$\frac{a}{p}$	Jednaki ponderi uparivanih parova bez (0 – 0) parova u imeniocu	Rasel-Rao
$\frac{a}{a+b+c}$	Jednaki ponderi uparivanih parova bez (0 – 0) parova u brojiocu	Žakard
$\frac{2a}{2a+b+c}$	Dvojni ponderi na (1 – 1) parove bez (0 – 0) parova u brojiocu ili imeniocu	Čekanovski-Sorensen-Dajs
$\frac{a}{a+2(b+c)}$	Dvojni ponderi za neuparovane parove bez (0 – 0) parova u brojiocu ili imeniocu	Kulezinski

Druga mera koja opisuje redove u matrici Y je **kosinus ugla** θ između vektora y_i i y_j, koja je za $i, j = 1,2,\ldots,n$ definisana kao:

$$\cos\theta = c_{ij} = y_i' y_j / \|y_i\| \|y_j\|.$$

Zbog $-1 \le c_{ij} \le 1$, prvi uslov mere sličnosti nije zadovoljen. Ako se normalizuju elementi svakog reda tako da $\|\tilde{y}_i\|^2 = \|\tilde{y}_j\|^2 = 1$, onda prema zakonu kosinusa:

$$\|\tilde{y}_i - \tilde{y}_j\|^2 = \|\tilde{y}_i\|^2 + \|\tilde{y}_j\|^2 - 2\|\tilde{y}_i\|^2 \|\tilde{y}_j\|^2 \cos\theta$$

i kvadratno rastojanje postaje:

$$d_{ij}^2 = 2(1 - c_{ij})$$

tako da se može primeniti mera različitosti da bi se izvršila klaster analiza redova matrice Y sa normalizovanim vrednostima.

Mere sličnosti kod binarnih promenljivih su važne u klaster analizi. Da bi se konstruisale mere sličnosti binarnih podataka koristi se tabela kontingencije binarnih promenljivih. Postavlja se pitanje kako ponderisati parove istih i parove različitih kodova binarnih promenljivih, jer par (1 – 1) može biti značajniji od para (0 – 0), jer prvi par ukazuje na prisutnost karakteristike, dok drugi ukazuje na odsustvo karakteristike. Moguća je i situacija kada se parovi (0 – 0) uopšte ne uzimaju u analizi. Da bi se omogućilo različito ponderiranje parova istih i parova različitih binarnih promenljivih, kao i tretman (0 – 0) parova, prikazano je nekoliko različitih odstojanja (Tabela 2.2).

2.5.Metodi klaster analize

Da bi se započeo postupak klaster analize, potrebno je da se najpre konstruiše matrica bliskosti. Ova matrica prikazuje jačinu odnosa između parova redova $Y'_{n\times p}$ ili matrice podataka $Y_{n\times p}$. Algoritmi dizajnirani za sprovođenje klaster analize su grupisani u dve grupe nazvane **hijerahijske** i **nehijerarhijske klaster metode**. Opšte rečeno, **hijerahijski metodi** generišu sekvencu klaster rešenja, počevši sa klasterom koji sadrži samo jedan objekat i kombinuje objekte dok svi objekti ne formiraju jedan klaster. Ovi metodi su nazvani **aglomeracijski hijerahijski metodi**. Drugi hijerahijski metodi počinju sa jednim klasterom i dele objekte sukcesivno da bi formirali klastere sa samo jednim objektom. Ovi metodi su nazvani **hijerahijski metodi deljenjem**. U nastavku su prikazani **aglomeracijski hijerahijski metodi**: **jednostruko povezivanje (najbliži sused)**, **kompletno ili potpuno povezivanje (najudaljeniji sused)**, **prosečno povezivanje (prosečno rastojanje)**, **metod centroida** i **Wardov metod**. Preferiran metod je prosečno povezivanje.

Kao i u ostalim metodama grupisanja, izvori greške i varijacije formalno se ne uzimaju u obzir u hijerarhijskim procedurama, što bi značilo da su metodi grupisanja osetljivi na nestandardne opservacije. Dobra ideja je sprovesti nekoliko metoda grupisanja, a za jedan metod, nekoliko načina za izbor rastojanja (sličnosti). Ako su

rezultati iz nekoliko metoda bar približno konzistentni, onda se može zaključiti da postoji "prirodno" grupisanje objekta. Hijerahijske metode udružuju objekte u najbliži klaster, u ranoj fazi, ali isti taj objekat ne može da bude pregrupisan u drugi klaster u kasnijoj fazi, iako je to bolje rešenje. Ovakvo pregrupisavanje je moguće kod nehijerahijske metode.

Prednost hijerahijske metode je to da ne treba da se zna broj klastera. Zato je ova tehnika poznata kako istraživačka, dok su nehijerahijske metode poznate kao potvrdne ili obrazložene[3] metode.

Nehijerahijske metode se koriste samo da bi se grupisali objekti u klastere.

2.5.1.Hijerahijska klaster analiza

U okviru **hijerahijske klaster analize**, posmatrani su aglomeracijski hijerahijski metodi i metodi povezivanja. Aglomeracijski hijerahijski metodi koriste elemente matrice bliskosti da generiraju dijagram u vidu drveta ili takozvani dendrogram, gde se kombiniraju objekti u klastere, počevši od objekata koji su najsličniji do objekata koji su najmanje slični da bi se na kraju dobio jedan jedini klaster.

Proces formiranja klastera, ukoliko je poznata matrica rastojanja $D_{n\times n}=(d_{ij})$, sprovodi se kroz sledeće korake:

1) Proces počinje sa n klastera, gde svaki klaster sadrži jedan objekt;
2) Traži se matrica različitosti D na osnovu koje se određuje najsličniji par elementa. Najsličniji par je predstavljen grupom d_{ij}, čiji su objekti i i j izabrani kao najsličniji;
3) Najsličniji par je predstavljen novim klasterom prema određenom kriterijumu. Na taj način, smanjuje se broj klastera za 1, brisanjem redova i kolona za objekte i i j. Izračunavaju se mere različitosti između formiranog klastera (ij) i svih ostalih klastera, preko određenog kriterijuma, i dodaje se red i kolona kod nove matrice različitosti.
4) Koraci 2 i 3 se ponavljaju $(n-1)$ puta, dok svi objekti ne formiraju jedan jedini klaster. U svakom koraku se identifikuju spojeni klasteri i vrednost različitosti na bazi koje se vrši spajanje klastera.

[3] Od engleske reči "confirmatory".

Sa menjanjem kriterijuma u koraku 3, dobijamo nekoliko aglomeracijskih hijerahijskih klaster metoda, ili algoritama grupisanja koji definišu kako je određena bliskost između klastera sa više objekta u procesu grupisanja.

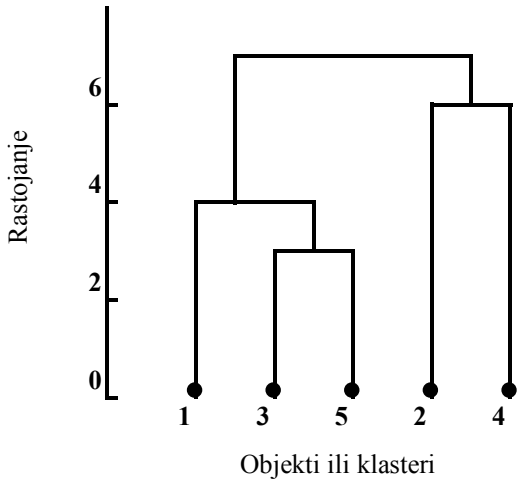
Pojedinačno ili jednostruko povezivanje (metod najbližeg suseda)

Da bi se implementirao metod najbližeg suseda, vrši se kombinovanje objekata u klastere koristeći najmanju različitost između klastera. Ako je i bilo koji element iz klastera R, $i \in R$, i s je bilo koji element u klasteru S, $j \in S$, rastojanje između R i S se izračunava kao:

$$d_{(R)(S)} = \min\left\{d_{ij} \middle| i \in R, j \in S\right\}.$$

Uz svaki korak procesa, kreira se dendrogram (Slika 2.2.), koji predstavlja grafički prikaz podređenih rastojanja na kojima se objekti povezuju. Grane drveta predstavljaju klastere ili objekte. Grane su povezane u čvorove čija pozicija na osi sličnosti (ili rastojanja) ukazuje na nivo povezivanja.

Slika 2.2. Dendrogram



Kompletno povezivanje (metod najudaljenijeg suseda)

Klaster procedura za kompletno povezivanje je ista kao i kod pojedinačnog povezivanja, osim što u svakoj fazi, rastojanje (sličnost) između dva klastera je određeno preko rastojanja (sličnosti) između dva elementa iz svakog klastera, koji su međusobno najudaljeniji.

Aglomeracijski algoritam počinje traženjem najmanjeg rastojanja u matrici rastojanja $D=(d_{ij})$ i nastavlja se spajanjem korespondirajućih objekata, i i j, da bi se dobio klaster (ij).

U ovom slučaju, ako $i \in R$ i $j \in S$, gde su R i S dva klastera, rastojanje između klastera R i S se izračunava kao:

$$d_{(R)(S)} = \max\{d_{ij} | i \in R, j \in S\}.$$

Prosečno povezivanje (prosečno rastojanje)

Input u algoritmu prosečnog povezivanja mogu biti rastojanja ili sličnosti. Proces započinje sa matricom rastojanja $D=(d_{ij})$ da bi se našli najsličniji objekti, i i j. Ovi se objekti spajaju u klaster (ij). Razlike između ovog klastera i nekog drugog klastera su determinisane klaster algoritmom. Umesto da se koristi minimum ili maksimum kao mera, izračunava se razdaljina između dva klastera preko prosečne vrednosti različitosti za svaki klaster

$$d_{(R)(S)} = \frac{\sum_i \sum_j d_{ij}}{n_R n_S}$$

gde $i \in R$, $j \in S$, i n_R i n_S predstavljaju broj objekta u svakom klasteru.

Metod centroida

Kod metoda prosečnog povezivanja, rastojanje između dva klastera je definisano kao prosečna vrednost mere različitosti. Ako se pretpostavi da klaster R sadrži n_R elemenata i klaster S sadrži n_S elemenata, onda centrodi za klastere koji sadrže dva objekta

$$\bar{y}_i = \frac{\sum_i y_i}{n_R} = \begin{bmatrix} \bar{y}_{i1} \\ \bar{y}_{i2} \\ \vdots \\ \bar{y}_{ip} \end{bmatrix} \quad \text{i} \quad \bar{y}_j = \frac{\sum_j y_j}{n_S} = \begin{bmatrix} \bar{y}_{j1} \\ \bar{y}_{j2} \\ \vdots \\ \bar{y}_{jp} \end{bmatrix}$$

i kvadratno Euklidovo rastojanje između dva klastera je $d_{ij}^2 = \|\bar{y}_i - \bar{y}_j\|^2$. Aglomeracijski proces metoda centroida počinje sa matricom rastojanja $D=(d_{ij})$. Zatim se dva najsličnija klastera udružuju preko ponderiranog proseka za dva klastera. Ako označimo novi klaster sa T, centroid ovog klastera je:

$$\bar{y}_t = (n_R \bar{y}_i + n_S \bar{y}_j)/(n_R + n_S).$$

Metod centroida je poznat i kako metod medijane, ako se koristi neponderisan prosek centroida, $\bar{y}_t = (\bar{y}_i + \bar{y}_j)/2$. Metod medijane je bolji metod kada je $n_R > n_S$ ili $n_S > n_R$.

Wardov metod

Wardov metod ima tendenciju da pronađe kompaktne klastere približno iste veličine, uz napomenu da klaster rešenje može da bude pod uticajem nestandardnih opservacija. Ovaj metod koristi hijerahijske procedure grupisanja koje minimiziraju gubitke informacija pri spajanju dve grupe. Gubitak informacije znači povećavanje vrednosti kriterijuma - suma kvadrata greške (SKG). Za jedan klaster, R, SKG_R suma kvadrata greške je suma kvadrata odstojanja svakog objekta klastera od sredine klastera (centroida). Ako je broj klastera k, onda je $SKG = SKG_1 + SKG_2 + \ldots + SKG_k$. Pri svakom koraku analize, svaka moguća unija klastera se uzima u obzir, i dva klastera čija kombinacija rezultira sa najmanjim povećivanjem SKG-a (ili minimalnim gubitkom informacije) se spajaju. Inicijalno, svaki klaster ima samo jedan objekat, tako da ako postoji ukupno n objekata, onda je $\mathrm{SKG_k} = 0,\ \mathrm{k} = 1,2,\ldots,\mathrm{n}$, tako da je $SKG = 0$. Suprotno, ako se svi klasteri nalaze u jednoj grupi sa n objekata, onda je vrednost SKG:

$$SKG = \sum_{i=1}^{n} (y_i - \bar{y})'(y_i - \bar{y}) = \sum_{i=1}^{n} \left\| y_i - \bar{y} \right\|^2$$

gde multivarijaciona mera je y_i koja označava i-ti objekat, dok $\bar{y}$ označava sredinu svih objekata.

2.5.2.Nehijerahijska k - sredina klaster analiza

U hijerarhijskoj klaster metodi broj klastera nije poznat unapred. Proces počinje sa matricom rastojanja, i kada se jednom objekat grupiše u klaster, ne vrši se njegova ponovna realokacija. Ovi se metodi mogu koristiti i za grupisanje objekata i za grupisanje promenljivih. **Nehijerahijske klaster metode** se koriste samo da bi se grupisali objekti. Proces počinje sa matricom originalnih podataka Y. Broj klastera k mora da bude poznat unapred, kao i centroidi klastera ili jezgra klastera[4], tako da se

[4] Od engleske reči “cluster seeds”.

opservacije mogu pregrupisati korišćenjem određenog kriterijuma, kao i označavanje kraja realokacije korišćenjem određenih pravila za zaustavljanje[5].

Najpopularniji nehijerahijski metod je k **- sredina klaster analiza**. Da bi se inicirala nehijerahijska klaster metoda, mora se najpre selektovati k centroida ili jezgra klastera. Jezgra klastera mogu da budu prvih k opservacija, ili prvih k opservacija pri definisanom nivou separacije, k slučajno izabranih tačaka. k inicijalnih tačaka mogu se zameniti na osnovu nekog kriterijuma zamene. Kada su tačke izabrane, razvija se grupisanje ili regrupiranje svakog objekta na osnovu kriterijuma konvergencije. Prema tome, osnovni koraci su:

1) Izbor k jezgara klastera;
2) Grupisanje svake opservacije od ukupno $n-k$ opservacija u najbliži centroid i ponovno izračunavanje centroida;
3) Korak iz tačke dva se ponavlja sve dok se sve opservacije ne grupišu ili dok razlike u centroidima klastera ne postanu dovoljno male.

Izbor jezgara klastera se može obaviti na dva načina. Prvi je kada klastere ne određuje istraživač, a to je slučaj kada su podaci bili analizirani korišćenjem neke druge multivarijacione metode. Najčešči primer je korišćenje hijerahijskog klaster algoritma da bi se dobio broj klastera a zatim generiranje jezgra klastera. Smatra se da ukoliko je broj klastera poznat, da je dostupna i informacija o osnovnim karakteristikama klastera. Drugi način dobijanja jezgara klastera je generiranje istih iz opservacija uzorka, na sistematski način ili jednostavno slučajnom selekcijom. Izbor jezgra klastera je vrlo važan jer se različita klaster rešenja dobijaju za različita jezgra klastera.

Primenom klaster analize za grupisanje objekta, mogu se kombinovati hijerahijske i nehijerahijske klaster metode. U prvom koraku, koristi se hijerahijska procedura da bi se identifikovala jezgra i broj klastera, koji bi bili input u nehijerahijskoj proceduri da bi se dobili bolji rezultati analize.

Ova klaster analiza pretpostavlja da je uzorak velik (na primer, $n > 200$)[6].

[5] Od engleske reči "stopping rule".

[6] Garson, D., 2009, *Cluster analysis from Statnotes: Topics in Multivariate analysis*, retrieved from http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm

2.5.3.Dvostepena klaster analiza

Dvostepena klaster analiza je metod koji se koristi kada se raspolaže sa velikom bazom podataka, jer hijerahijska i nehijerahijska klaster metoda nemaju istu efikasnost za velike baze. Ova analiza se koristi i za kategorijske i za numeričke promenljive, i nalazi primenu u analizi kategorijskih promenljivih sa tri ili više modaliteta.

Dvostepena klaster analiza je metod koj zahteva samo jedan prolaz kroz bazu podataka. Proces se sastoji od dva koraka: prvo se vrši inicijalno grupisanje objekta u manje pod-klastere, a zatim se ovi pod-klasteri tretiraju kao posebni objekti grupišu preko hijerahijskog klaster metoda. Moguće je da algoritam dvostepene klaster analize odredi broj klastera, ali i da broj klastera bude prethodno naznačen.

U dvostepenoj klaster analizi, ukoliko su jedna ili više promenljivih kategorijske, koristi se **mera rastojanja (prirodnog logaritma) funkcije verodostojnosti**[7], tako da se u klaster grupišu objekti sa najvećom vrednošću ove mere. Ako su sve numeričke promenljive kontinuirane, koristi se **Euklidovo rastojanje**, tako da se objekti grupišu u klaster koji ima najmanju vrednost Euklidovog rastojanja. SPSS algoritam koristi smanjivanje mere rastojanja (prirodnog logaritma) funkcije verodostojnosti kao meru rastojanja za kreiranje klastera, jer je ova mera kompatibilna kako za kategorijske, tako i za kontinuirane promenljive.

Mera rastojanja je potrebna u oba koraka, u koraku inicijalnih grupisanja, i u koraku samog grupisanja. Mera rastojanja (prirodnog logaritma) funkcije verodostojnosti predstavlja rastojanje koje je zasnovano na verovatnoći. Rastojanje između dva klastera je u relaciji sa smanjivanjem vrednosti (prirodnog logaritma) funkcije verodostojnosti, kada se oni stapaju u jedan klaster. U izračunavanju (prirodnog logaritma) funkcije verodostojnosti, normalna raspodela za kvantitativne promenljive, i multinomna raspodela za kategorijske promenljive, se podrazumeva (štaviše je i poželjna). Ipak, dvostepena klaster analiza daje dobre rezultate, i kada pretpostavka o normalnosti nije ispunjena. Isto tako, ova klaster analiza pretpostavlja da je uzorak velik. Takođe, podrazumeva se da su promenljive međusobno nezavisne, pritom se misli i na objekte ili opservacije. Rastojanje između klastera R i S definiše se kao:

$$d_{(R)(S)} = \xi_R + \xi_S - \xi_{(R,S)}$$

[7] Od engleske reči "log - likelihood distance measure"

gde su

$$\xi_v = -N_v \cdot \left(\left(\sum_{k=1}^{K^A} \frac{1}{2} \cdot \log(\hat{\sigma}_k^2 + \hat{\sigma}_{v \cdot k}^2) \right) + \left(\sum_{k=1}^{K^B} \hat{E}_{v \cdot k} \right) \right)$$

i

$$\hat{E}_{v \cdot k} = -\sum_{l=1}^{L_k} \left(\frac{N_{v \cdot k \cdot 1}}{N_v} \cdot \log\left(\frac{N_{v \cdot k \cdot 1}}{N_v} \right) \right)$$

sa sledećim oznakama:

K^A je broj kvantitativnih promenljivih u analizi;

K^B je broj kategorijskih promenljivih u analizi;

R_k je interval varijacije ili opseg k - te kvantitativne promenljive;

N je broj objekta u bazi podataka;

N_k je broj objekta u k - tom klasteru;

$\hat{\sigma}_k^2$ je ocenjena varijansa k - te kvantitativne promenljive nad svim podacima;

$\hat{\sigma}_{Rk}^2$ je ocenjena varijansa k - te kvantitativne promenljive u R - tom klasteru;

N_{Rkl} je broj objekta u R - tom klasteru, gde k -ta kategorijska promenljiva uzima l - tu kategoriju;

$d_{(R)(S)}$ je rastojanje između R - tog i S - tog klastera;

Ukoliko se zanemari $\hat{\sigma}_k^2$ u jednačini, rastojanje između klastera R i S biće jednako vrednosti smanjenja (prirodnog logaritma) funkcije verodostojnosti, kada se ta dva klastera spoje. Član $\hat{\sigma}_k^2$ je pridodat za rešenje problema koje nastaje, ukoliko je $\hat{\sigma}_{v \cdot k}^2 = 0$, jer se u tom slučaju dolazi do nedefinisane vrednosti prirodnog logaritma. To se dešava ukoliko su klasteri jednočlani.

Druga mera rastojanja, Euklidsko rastojanje, se može primeniti kada su sve promenljive kvantitativne. Euklidsko odstojanje između dve tačke je jasno definisano. Odstojanje između dva klastera je definisano Euklidskim rastojanjem između njihovih centroida. Centar klastera je definisan kao vektor sredina svih promenljivih, za dati klaster.

Postupak dvostepene klaster analize počinje sa prvim korakom, a to je stvaranje inicijalnih klastera. Ovaj korak se primenjuje u metodi redoslednog grupisanja. On analizira objekte redom i odlučuje, da li će se dati objekat pridružiti jednom od već

formiranih (postojećih) klastera, ili će formirati novi klaster. Ova odluka se donosi na osnovu kriterijuma rastojanja.

Analiza je implementirana konstrukcijom modifikovanog **drveta klaster osobina**. Drvo klaster osobina se sastoji od grana, a svaka grana se sastoji od grančica. Svaka grančica predstavlja finalni pod-klaster. Grane (koje se dalje ne granaju) kao i objekti koji se nalaze na toj grani, se upotrebljavaju, za veoma brzo i korektno klasifikovanje novih objekata. Svaku granu karakterišu posebne osobine, koje se grade na osnovu objekata, koji su na njoj. Ukoliko su promenljive kvantitativne razmatraju se sredina i varijansa, a ukoliko su promenljive kategorijske, razmatra se frekvencija svake kategorije. Za svaki objekat, pri određivanju na kojoj će grani biti smešten, polazi se od korena stabla (nulte grane). Prvo se analiziraju grane, da bi se utvrdilo, kojoj će grani pripasti objekat. Zatim se analiziraju podgrane (grančice date grane) da bi se uvidelo kojoj će od njih da pripadne, i tako redom sve dublje i dublje, u zavisnosti od razgranatosti stabla.

Kada se odredi kojoj će grani pripasti objekat, uzimaju se u obzir njegove karakteristike, i vrši se ponovni proračun osobina te grane. Ukoliko je objekat specifičan, u okviru date grane formira se nova podgrana. Ukoliko nema prostora da se na datoj grani formira nova podgrana ,onda se data grana deli na dve nove grane. Preraspodela objekata na dve nove grane se vrši na sledeći način: dva najudalenjija objekta postave se na po jednu granu, a zatim se svi preostali objekti redistribuiraju u odnosu na te dve grane, na osnovu kriterijuma bliskosti. Ukoliko stablo klaster osobina premaši dozvoljenu maksimalnu veličinu, ponovno se kreira stablo (na osnovu sadašnjeg stabla), tako što se poveća kriterijum bliskosti. Posledica takve analize je da su jedinice posmatranja sličnije među sobom, što dovodi do smanjivanja broja grana u drvetu i tako ostaje prostora za nove (još neobrađene) objekte. Ovaj proces je na snazi sve dok se ne obrade svi objekti.

Svi objekti koji pripadaju datoj grani su kolektivno predstavljeni osobinom date grane. Kada se objekat pridoda grani, onda se osobina date grane ažurira uzimajući u obzir i osobinu datog objekta i postojeću osobinu grane. Na ovaj način postojaće samo osobine grana, a ne i osobine individualnih objekata. Zbog toga je stablo klaster osobina mnogo manje od stabla klastera individualnih objekata.

Napomenimo da stablo klaster osobina može da zavisi od redosleda objekata smeštenih u bazi podataka. Da bi se minimizirao efekat redosleda objekata, predlaže se da se objekti u bazi razmeste na slučajan način.

Tretiranje nestandarnih opservacija podrazumeva da su nestandardne opservacije oni objekti koje se ne mogu kvalitetno oceniti pomoću nijednog klastera. Smatraćemo da su objekti nestandardni ukoliko pripadaju grani, na kojoj je broj objekata (relativno) manji od određenog procenta (po početnoj definicijii, 25%) veličine najveće grane stabla klaster osobina. Pre ponovnog kreiranja stabla, primenom ove procedure uočavaju se potencijalne nestandardne opservacije i zatim se sklanjaju sa strane. Nakon ponovnog kreiranja stabla, proveravamo da li se nestandardne opservacije mogu razvrstati na postojeće grane, bez mogućnosti stvaranja novih. Na završetku kreiranja stabla klaster osobina, objekti koji se nisu mogli razvrstati, smatraće se nestandardnim opservacijama.

Kada se započne proces grupisanja, postavlja se pitanje broja klastera. Odgovor zavisi od samih podataka u bazi. Karakteristika hijerarhijske klaster analize je da kreira niz rešenja u okviru jednog prolaza kroz podatke, sa jednim, dva, tri, i više klastera. Za *k*-sredina klaster algoritam potrebno je primeniti klasterizaciju više puta (svaki put za određeni broj klastera) da bi se generisao niz rešenja.

Za automatsko određivanje broja klastera, SPSS je razvio dvo-stepenu proceduru koja je kompatibilna sa hijerarhijskom klaster metodom. U prvom koraku, izračunava se *BIC* (Bajesov informacioni kriterijum) ili *AIC* (Ekejkov informacioni kriterijum), statistika za svaki različiti broj klastera u rešenju, koja se zatim upotrebljava za pronalaženje inicijalne ocene broja klastera. U drugom koraku, inicijalna ocena se popravlja pronalaženjem najvećeg povećanja u odstojanju između dva najbliža klastera u okviru svake etape hijerarhijskog grupisanja.

Statistike *BIC* i *AIC* za *R* klastera, su definisane kao:

$$BIC_R = -2 \cdot \sum_{i=1}^{R} \xi_R + m_R \cdot \log(N)$$

$$AIC_R = -2 \cdot \sum_{i=1}^{R} \xi_R + 2 \cdot m_R$$

gde je

$$m_R = R \cdot \left\{ 2 \cdot K^A + \sum_{k=1}^{K} (L_k - 1) \right\}$$

i gde je sa L_k označen broj modaliteta kod k - te kategorijske promenljive.

Uključivanje u klaster, ukoliko se modeliranje sprovodi bez nestandardnih opservacija, vrši se dodeljivanjem opservacije najbližem klasteru, na osnovu vrednosti mere rastojanja. Ukoliko se modeliraju nestandardne opservacije, onda koristimo rastojanje definisano (prirodnom logaritmom) funkcijom verodostojnosti.

Pretpostavimo da ekstremne opservacije slede normalnu raspodelu. Izračunavaju se dve funkcije verodostojnosti, jedna kada se objekat dodeljuje nestandardnom klasteru, a jedna kada se dodeljuje najbližem klasteru (koji nije nestandardni klaster). Objekat se onda dodeljuje onom klasteru koji ima veću vrednost (prirodnog logaritma) funkcije verodostojnosti. Ovaj postupak je ekvivalentan dodeljivanju objekata najbližem klasteru (koji nije nestandardni) ukoliko je odstojanje od njega manje od kritične vrednosti $C = \log(V)$, gde je $V = \left(\prod_k R_k\right) \cdot \left(\prod_m L_m\right)$. U ostalim slučajevima, objekat se klasifikuje kao nestandardna opservacija.

Objekat se dodeljuje najbližem klasteru (koji nije nestandardni) ukoliko je Euklidsko odstojanje između njih manje od kritične vrednosti $C = 2 \cdot \sqrt{\frac{\sum_{l=1}^{K^A} \hat{\sigma}_{k \cdot l}^2}{K^A}}$, inače se svrstava kao nestandardna opservacija.

Objekti za koje nedostaju neke vrednosti isključuju se iz analize.

2.6.Određivanje broja klastera

Vrlo važno pitanje u hijerahijskoj i nehijerahijskoj analizi je determiniranje broja klastera koji je najreprezentativniji za strukturu podataka. U hijerahijskoj analizi kreira se skup mogućih klaster rešenja, ali potrebno je odabrati jedno ili nekoliko rešenja koja bi najadekvatnije prikazala strukturu podataka. Ista odluka se donosi i u nehijerahijskoj analizi gde se najbolje rešenje bira između dva ili više ponuđenih klaster rešenja.

Ne postoji standardna objektivna procedura selekcije za izbor najboljeg rešenja. Razvijeni su mnogi kriterijumi, koji koriste kompleksne pristupe i oni su karakteristični za različite softverske programe. Jedna grupa ovih kriterijuma su mere promene heterogenosti. Ove mere slede promene i izračunavaju se u celom toku

spajanja klastera, a koristi se onda kada imamo značajan porast heterogenosti, jer se smatra da su prethodno spojeni klasteri najbolje rešenje. To je i logično, jer kada se spoje klasteri koji značajno povećavaju heterogenost, očigledno je da je prethodno rešenje bilo najbolje.

U regresionoj analizi, **koeficient determinacije** R^2, pokazuje procenat varijabiliteta objašnjenog regresijom u odnosu na ukupni varijabilitet, odnosno stepen do kojeg su varijacije zavisne promenljive objašnjenje preko varijacija nezavisne promenljive. Kod analize varijanse koeficijent R^2 je definisan kako odnos između sume kvadrata grupa i ukupne sume kvadrata i pretstavlja meru ukupnih varijacija zavisne promenljive koja je sadržana ili objašnjena preko sredina grupa. Tako, u klaster analizi možemo da konstruišemo R^2 i da ga izračunamo svaki put kad se broj klastera menja. Za n klastera, ukupna suma kvadrata je $T = \sum_{i=1}^{n} \|y_i - \bar{y}\|^2$ i suma kvadrata između klastera (SKG) k je $SKG_k = \sum_i \|y_i - \bar{y}_k\|^2$. Koeficijent R^2 za k klastera je definisan kao:

$$R_k^2 = \frac{T - \sum_k SKG_k}{T}$$

Za n klastera, važi da je $SKG_k = 0$, tako da je i $R^2 = 1$ da se broj klastera smanjuje od n do 1, klasteri bi trebali da se sve više razlikuju. Veliko smanjenje R^2 trebalo bi da ukaže na specifičnu tačku[8]. Pri spajanju klastera R i S može se izračunati razlika prirasta u R^2 ili

$$SR^2 = R_k^2 - R_{k-1}^2$$

nazvan **poluparcijalni** R^2 **indeks**. Statistika SR^2 upoređuje odnos $SKG_t - (SKG_r + SKG_s)$, gde su klasteri C_R i C_S spojeni da bi formirali C_T u ukupnoj sumi kvadrata $T = \sum_{i=1}^{n} \|y_i - \bar{y}\|^2$. Što je veći porast, veći je i "gubitak homogenosti", odnosno, klasteri su više separatisani.

U analizi koristimo nekoliko testova statistike da bi dobili stepen heterogenosti za svaki novi klaster nastao spajanjem dva prethodna klastera. Najčešće korišćena test statistika je **pseudo** F **statistika** koja upoređuje koliko je bolje rešenje sa k klastera u poređenju sa rešenjem sa $k-1$ klastera:

[8] Od engleske reči "distinct joint"

$$F_k^* = \frac{(T - \sum_k SKG_k)/(k-1)}{\sum_k SKG_k /(n-k)}.$$

Ako ova statistika ima visoke vrednosti, to ukazuje da je $k-1$ rešenje bolje od k rešenja.

Pseudo t^2 statistika se koristi za poređenje sredina spojenih klastera za sve promenjive uključene u analizu, da bi se utvrdilo dali su značajno razdvojene za bilo koji nivo klastering hijerarhije:

$$pseudot^2 = \frac{[SKG_t - (SKG_r + SKG_s)](n_R + n_S - 2)}{SKG_r + SKG_s}.$$

Ukoliko ova statistika ima vrednost statistički značajno veću od vrednosti ostalih rešenja, to je indikator da spajanje dva klastera kreira visoku heterogenost i da netreba spajati ova dva klastera.

2.7.Interpretacija, validacija i profiliranje klastera

Centroidi klastera, srednji profil klastera za sve promenljive uključene u analizu, su posebno korisni u fazi interpretacije klastera. **Interpretacija** uključuje ispitavanje izrazitih karakteristika za svaki profil klastera i identifikaciju značajne razlike između klastera. Klaster rešenja koja nemaju značajnu varijaciju treba da budu ispitana još jednom.

Treba ispitati da li centroidi klastera imaju sličnosti sa pretpostavljenim i očekivanim klaster rešenjima, baziranim na teoriji ili praktičnom iskustvu.

Validacija je od posebnog značaja u klaster analizi, jer klasteri opisno prikazuju svoju strukturu i potrebna je dopunska podrška za ispitivanje njihove relevantnosti. Ukrštena validacija empiriski potvrđuje dobijeno klaster rešenje preko kreiranja dva pod-uzoraka (slučajna podela uzoraka) i onda upoređuje klaster rešenja ovih uzoraka za konzistentost, preko broja dobijenih klastera i klaster profila.

Validacija se može dobiti i preko ispitivanja razlika promenljivih koje nisu uključene u klaster analizu, a za čiju analizu postoji teoretski ili praktičan razlog kako u većoj meri bi se objasnila varijacija između klastera.

Profiliranje opisuje karakteristike svakog klastera u cilju objašnjenje razlika klastera za različite dimenzije. Ovaj proces uključuje diskriminacionu analizu. Procedura počinje sa identifikacijom klastera. Koriste se podaci koji prethodno nisu

bili uključeni u klaster analizu kako bi se profilirale karakteristike svakog klastera. Ovo su najčešće demografske karakteristike. Upotrebom diskriminacione analize, upoređuje se prosečna vrednost profila klastera.

Ukratko, analiza profila se fokusira ne na direktnoj determinaciji klastera, već na karakteristikama klastera otkako su oni identifikovani.

2.8.Praktični primeri: Primena klaster analize u rešavanju empiriskih problema

2.8.1.Primena hijearhijske klaster analize u klasifikaciji opština Makedonije u odnosu na njihove karakteristike

Hijerahijska klaster analiza u softverskom paketu SPSS zahteva od istraživača da izabere meru rastojanja, zatim metod za povezivanje da bi se formirali klasteri, i onda da utvrdi koliki broj klastera bi bio najpogodniji za korišćene podatke. Pri tom, klasteri mogu grafički da se prikažu preko dendrograma ili preko dijagrama ledenica[9].

U radu hijerahijska klaster analiza je upotrebljena za klasifikaciju 84 opštine Makedonije na osnovu izabranih karakteristika, kako bi se preko centroida klastera dobile informacije koje bi bile osnova za kreiranje nacionalne ekonomske i razvojne politike. Za grupisanje makedonskih opština, posmatraćemo devet, demografskih i ekonomskih promenljivih: *Prirodni prirast*, *Broj prodavnica u maloprodaji*, *Broj individualnih poljoprivrednih domaćinstva*, *Ukupan broj stanovnika*, *Ukupan broj domaćinstava*, *Broj obrazovanih stanovnika*, *Ukupno ekonomski aktivno stanovništvo*, *Ukupan broj zaposlenih lica*, U*kupan broj nezaposlenih lica*. Podaci su dobijeni iz Državnog zavoda za statistiku i odnose se na pokazatelje o stanovništvu i domaćinstvima, iz popisa 2002. godine. Privredno-ekonomski pokazatelji datih opština odnose se na 2008. godinu.

S obzirom na to da su promenljive iskazane kao broj stanovnika, domaćinstava ili broj prodavnica, i da ne postoje velike razlike u mernim skalama, ne sprovedi se **standardizacija promenljivih**.

Sledeći korak u analizi je da se otkrije da li postoje nestandarne opservacije. Iz razloga što je ukupan broj opservacija 84 opština, grafički prikaz bi bio nedovoljno jasan. Umesto toga koristimo Mahalanobisovu D^2 meru.

[9] Od engleske reči "icicle plot".

Tabela 2.3. Identifikacija potencijalnih nestandarnih opština na osnovu Mahalanobisovog odstojanja

Opština	D^2	Opština	D^2	Opština	D^2	Opština	D^2
Karbinci	0,76	Pehcevo	1,61	Dolneni	3,08	Kičevo	8,16
Valandovo	0,88	Brvenica	1,63	Butel	3,10	Bogovinje	8,61
Krivogaštani	0,90	Čaška	1,63	Kriva Palanka	3,20	Lipkovo	10,06
Mak. Kamenica	0,90	Mogila	1,74	Staro Nagoričane	3,33	Saraj	11,05
Makedonski Brod	0,92	Češinovo	1,79	Berovo	3,46	Tearce	11,84
Čučer-Sandevo	0,93	Zelenikovo	1,81	Sveti Nikole	3,68	Šuto Orizari	11,93
Rosoman	0,94	Vranešnica	1,84	Aračinovo	3,73	Gazi Baba	12,95
Kruševo	1,00	Negotino	1,89	Delčevo	4,03	Kavadarci	13,98
Petrovec	1,00	Ilinden	2,02	Resen	4,19	Kisela Voda	17,07
Sopište	1,03	Dojran	2,13	Gevgelija	4,34	Radoviš	19,38
Drugovo	1,06	Demir Hisar	2,14	Vinica	4,47	Strumica	21,09
Rankovce	1,06	Oslomej	2,16	Želino	4,68	Kumanovo	22,66
Bogdanci	1,24	Probištip	2,17	Bosilovo	4,79	Struga	27,32
Konče	1,27	Novo Selo	2,19	Studeničani	5,85	Čair	30,23
Novaci	1,27	Centar Župa	2,23	Veles	6,15	Karpoš	32,18
Kratovo	1,31	Debarca	2,25	Gorče Petrov	6,55	Bitola	32,29
Zrnovci	1,34	Jegunovce	2,33	Štip	6,71	Gostivar	33,96
Gradsko	1,35	Plašnica	2,56	Debar	7,11	Prilep	34,50
Lozovo	1,40	Zajas	2,57	Vrapčište	7,21	Tetovo	35,47
Demir Kapija	1,41	Vevčani	2,94	Ohrid	8,08	Aerodrom	45,18
Mavrovo i Rostuša	1,55	Vasilevo	2,97	Kočani	8,15	Centar	50,08

Izvor: Rezultati dobijeni primenom SPSS - a

Na osnovu visoke vrednosti D^2 kao nestandardne opservacije se isključuju opštine Aerodrom i Centar. Ne postoji utvrđena vrednost koja definiše za koje vrednosti Mahalanobisove mere bi trebalo isključiti nestandardne opservacije, tako da je ovo odluka koju donosi sam istraživač.

Korelaciona matrica **Pearsonovog koeficijenta** uključenih promenljivih ukazuje da postoji korelaciona veza između promenljivih. To je i očekivano, naročito za promenljive *Broj stanovnika*, *Broj prodavnica u maloprodaji*, *Broj obrazovanih lica* i *Broj domaćinstava*. Najednostavna ekonomska logika ukazuje, u većini situacija, da postoji linearna porcionalna povezanost, odnosno, više stanovnika, više domačinstva, više prodavnica u maloprodaji.

Nakon nekoliko kombinacija mera rastojanja i aglomeracijskih hijerahijskih metoda klasifikacije, kao najbolje klaster rešenje dobijeno je na **bazi kompletnog povezivanja mera rastojanja iskazane kosinusom ugla**. Izabrali smo da je broj rešenja od 3 do 7 klastera, da bi kasnije utvrdili optimalno klaster rešenje.

Tabela 2.4. Pripadnost opština datom klasteru kod raznih klaster rešenja

Opština	Broj klastera					Opština	Broj klastera				
	7	6	5	4	3		7	6	5	4	3
1:Butel	1	1	1	1	1	42:Kičevo	2	2	2	2	2
2:Gazi Baba	1	1	1	1	1	43:Konče	4	4	4	3	3
3:Gorče Petrov	1	1	1	1	1	44:Kočani	5	1	1	1	1
4:Karpoš	1	1	1	1	1	45:Kratovo	4	4	4	3	3
5:Kisela Voda	1	1	1	1	1	46:Kriva Palanka	5	1	1	1	1
6:Saraj	2	2	2	2	2	47:Krivogaštani	4	4	4	3	3
7:Čair	2	2	2	2	2	48:Kruševo	5	1	1	1	1
8:Šuto Orizari	2	2	2	2	2	49:Kumanovo	2	2	2	2	2
9:Aračinovo	3	3	3	2	2	50:Lipkovo	3	3	3	2	2
10:Berovo	4	4	4	3	3	51:Lozovo	7	6	4	3	3
11:Bitola	1	1	1	1	1	52:Mavrovo i Rostuša	3	3	3	2	2
12:Bogdanci	5	1	1	1	1	53:Mak. Kamenica	5	1	1	1	1
13:Bogovinje	3	3	3	2	2	54:Makedonski Brod	4	4	4	3	3
14:Bosilovo	5	1	1	1	1	55:Mogila	4	4	4	3	3
15:Brvenica	2	2	2	2	2	56:Negotino	5	1	1	1	1
16:Valandovo	5	1	1	1	1	57:Novaci	4	4	4	3	3
17:Vasilevo	4	4	4	3	3	58:Novo Selo	5	1	1	1	1
18:Vevcani	2	2	2	2	2	59:Oslomej	3	3	3	2	2
19:Veles	5	1	1	1	1	60:Ohrid	1	1	1	1	1
20:Vinica	5	1	1	1	1	61:Petrovec	2	2	2	2	2
21:Vranešnica	6	5	5	4	3	62:Pehčevo	4	4	4	3	3
22:Vrapčište	3	3	3	2	2	63:Plašnica	3	3	3	2	2
23:Gevgelija	5	1	1	1	1	64:Prilep	5	1	1	1	1
24:Gostivar	2	2	2	2	2	65:Probištip	5	1	1	1	1
25:Gradsko	4	4	4	3	3	66:Radoviš	5	1	1	1	1
26:Debar	2	2	2	2	2	67:Rankovce	7	6	4	3	3
27:Debarca	7	6	4	3	3	68:Resen	5	1	1	1	1
28:Delčevo	5	1	1	1	1	69:Rosoman	7	6	4	3	3
29:Demir Kapija	4	4	4	3	3	70:Sveti Nikole	5	1	1	1	1
30:Demir Hisar	5	1	1	1	1	71:Sopište	2	2	2	2	2
31:Dojran	5	1	1	1	1	72:Staro Nagoricane	6	5	5	4	3
32:Dolneni	7	6	4	3	3	73:Struga	2	2	2	2	2
33:Drugovo	7	6	4	3	3	74:Strumica	5	1	1	1	1
34:Želino	3	3	3	2	2	75:Studeničani	3	3	3	2	2
35:Zajas	3	3	3	2	2	76:Tearce	3	3	3	2	2
36:Zelenikovo	2	2	2	2	2	77:Tetovo	2	2	2	2	2
37:Zrnovci	4	4	4	3	3	78:Centar Župa	3	3	3	2	2
38:Ilinden	5	1	1	1	1	79:Časka	7	6	4	3	3
39:Jegunovce	2	2	2	2	2	80:Česinovo	4	4	4	3	3
40:Kavadarci	5	1	1	1	1	81:Čučer - Sandovo	5	1	1	1	1
41:Karbinci	7	6	4	3	3	82:Štip	1	1	1	1	1

Izvor: Rezultati dobijeni primenom SPSS - a

Kako su dve nestandardne opservacije isključene, analiza je dalje rađena za 82 opština (Output statističkog softvera SPSS-a najpre prikazuje **matricu bliskosti**).

Tabela 2.5. Aglomeraciski koeficient

Faza	Udruživanje klastera		Koefi-cient	Faza u kojoj se klaster pojavljuje prvi put		Slede-ća faza	Faza	Udruživanje klastera		Koef-icient	Faza u kojoj se klaster pojavljuje prvi put		Slede-ća faza
	Klas. 1	Klas. 2		Klas. 1	Klas. 2			Klas. 1	Klas. 2		Klas. 1	Klas. 2	
1	40	46	1,000	0	0	30	42	14	28	0,998	0	23	56
2	3	5	1,000	0	0	38	43	17	47	0,998	0	0	69
3	31	66	1,000	0	0	23	44	1	3	0,998	4	38	62
4	1	2	1,000	0	0	44	45	6	8	0,998	0	0	58
5	11	60	1,000	0	0	20	46	55	57	0,998	0	7	63
6	61	71	1,000	0	0	32	47	67	79	0,998	0	0	65
7	57	80	1,000	0	0	46	48	22	35	0,998	0	19	61
8	29	45	1,000	0	0	37	49	36	39	0,998	32	0	72
9	52	76	1,000	0	0	40	50	19	40	0,998	26	30	64
10	53	68	1,000	0	0	18	51	15	24	0,998	0	24	71
11	19	74	1,000	0	0	26	52	25	37	0,997	37	0	59
12	27	33	1,000	0	0	53	53	27	41	0,997	12	41	70
13	9	75	1,000	0	0	29	54	12	30	0,997	22	0	68
14	10	54	1,000	0	0	35	55	16	48	0,997	25	28	64
15	20	65	1,000	0	0	18	56	14	20	0,997	42	33	68
16	24	73	1,000	0	0	24	57	21	72	0,997	0	0	79
17	16	70	1,000	0	0	25	58	6	26	0,997	45	0	67
18	20	53	1,000	15	10	33	59	10	25	0,997	35	52	69
19	35	78	1,000	0	0	48	60	18	42	0,996	0	31	71
20	11	82	1,000	5	0	38	61	13	22	0,996	40	48	66
21	48	81	1,000	0	0	28	62	1	4	0,996	44	0	76
22	12	23	0,999	0	0	54	63	43	55	0,996	0	46	73
23	28	31	0,999	0	3	42	64	16	19	0,995	55	50	74
24	24	77	0,999	16	0	51	65	32	67	0,995	0	47	70
25	16	56	0,999	17	0	55	66	9	13	0,995	39	61	78
26	19	44	0,999	11	0	50	67	6	7	0,995	58	0	75
27	41	51	0,999	0	0	41	68	12	14	0,993	54	56	74
28	48	58	0,999	21	0	55	69	10	17	0,992	59	43	73
29	9	34	0,999	13	0	36	70	27	32	0,992	53	65	77
30	40	64	0,999	1	0	50	71	15	18	0,992	51	60	72
31	42	49	0,999	0	0	60	72	15	36	0,991	71	49	75
32	36	61	0,999	0	6	49	73	10	43	0,989	69	63	77
33	20	38	0,999	18	0	56	74	12	16	0,986	68	64	76
34	13	59	0,999	0	0	40	75	6	15	0,985	67	72	78
35	10	62	0,999	14	0	59	76	1	12	0,983	62	74	80
36	9	50	0,999	29	0	39	77	10	27	0,979	73	70	79
37	25	29	0,999	0	8	52	78	6	9	0,975	75	66	80
38	3	11	0,999	2	20	44	79	10	21	0,962	77	57	81
39	9	63	0,999	36	0	66	80	1	6	0,939	76	78	81
40	13	52	0,999	34	9	61	81	1	10	0,917	80	79	0
41	41	69	0,998	27	0	53	42	14	28	0,998	0	23	56

Izvor: Rezultati dobijeni primenom SPSS – a

Ova matrica ukazuje na sva moguća rastojanja uključenih promenljivih. Kada je broj objekata ili opservacija velik, i ova matrica je velika, i iz tog razloga ovu matricu ne prikazujemo.

Tabela 2.4. prikazuje pripadnost objekta klasteru. Broj naznačen kod svake opštine je broj klastera kome pripada opština. U tabeli je prikazano pet mogućih rešenja, od 3 do 7 klastera.

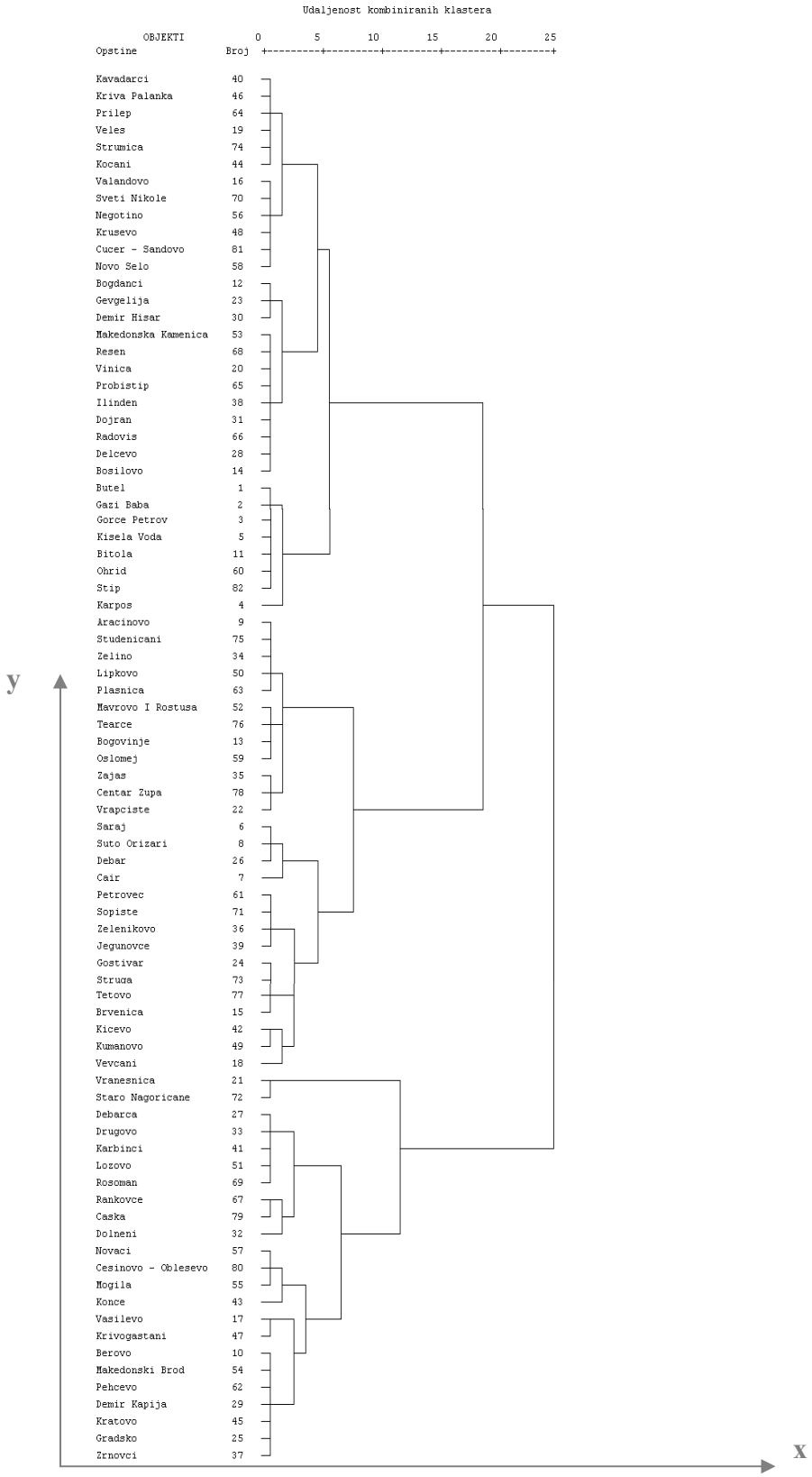
U Tabeli 2.5. je prikaz kao kriterijum o izbora najboljeg klaster rešenja je **aglomeraciskih koeficient**. U vrstama tabele data su faze grupisanja, numerisani od 1 do $(n-1)$. Faza $(n-1)$ obuhvata sve objekte u jednom klasteru. Kolone koje su označene "Kombinovani klasteri" prikazuju objekte ili broj već formiranih klastera koji se kombinuju u svakoj fazi.

Aglomeracijski koeficijent meri sličnost između dva najsličnija klastera u datom koraku analize. Kvalitetno rešenje analize je u onom koraku nakon koga dolazi do većeg smanjivanja vrednosti statistike "koeficijenta". U našoj analizi to bi bio korak 80 koji date opštine deli na tri klastera, pa kao konačno rešenje koristi se podela na tri klastera.

Na Slici 2.3. je prikaz celog procesa hijerahijske klaster analize pomoću **dendrograma**. Dendrogram prikazuje relativnu veličinu koeficijenta blizine u kombinovanju opština. Ako je vrednost koeficijenta blizine mala (ili vrednost koeficienta udaljenosti velika), to znači da su uključene opštine koji nisu međusobno slične, što predstavlja nepogodnu situaciju. Na grafičkom prikazu predstavljene su opštine na Y osi, a na X osi predstavljene su vrednosti koeficijenta blizine. Opštine koji su jako slične međusobno su bliže.

Opštine su svrstane u tri klastera. Prvi klaster broji 32 opštine, drugi klaster obuhvata 27 opština i treći klaster ima 23 opštine. U prvi klaster spadaju opštine Butel, Gazi Baba, Gorče Petrov, Karpoš, Kisela Voda, Bitola, Bogdanci, Bosilovo, Valandovo, Veles, Vinica, Gevgelija, Delčevo, Demir Hisar, Dojran, Ilinden, Kavadarci, Kočani, Kriva Palanka, Kruševo, Makedonska Kamenica, Negotino, Novo Selo, Ohrid, Prilep, Probištip, Radoviš, Resen, Sveti Nikole, Strumica, Čučer Sandovo i Štip. U drugom klasteru su opštine: Saraj, Čair, Šuto Orizari, Aračinovo, Bogovinje, Brvenica, Vevčani, Vrapčište, Gostivar, Debar, Želino, Zajas, Jegunovce, Kičevo, Kumanovo, Lipkovo, Mavrovo i Rostuša, Oslomej, Petrovec, Plašnica, Sopište, Struga, Studeničani, Tearce, Tetovo, Centar Župa.

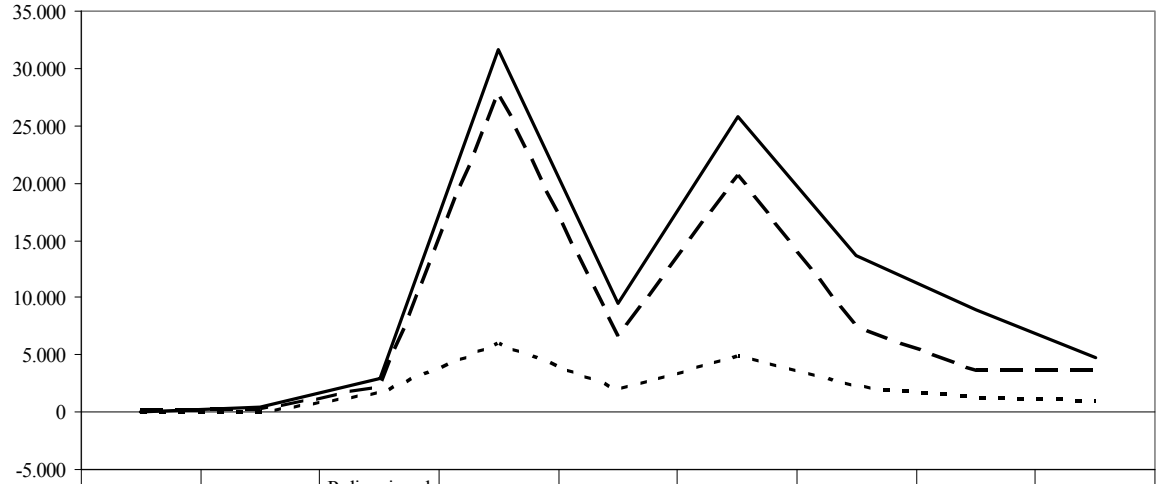
Slika 2.3. Dendrogram makedonskih opština



Izvor: Rezultati dobijeni primenom SPSS – a

U trećem klasteru su opštine Berovo, Vasilevo, Vranešnica, Gradsko, Debarca, Demir Kapija, Dolneni, Drugovo, Zrnovci, Karabincki, Konče, Kratovo, Krivogaštani, Lozovo, Makedonski Brod, Mogila, Novaci, Pehčevo, Rankovce, Rosoman, Staro Nagoričane, Čaška, Češinovo.

Slika 2.4. Centroidi klastera



	Prirodni prirast	Prodavnica u maloprodaji	Poljoprivredna domacinstva	Broj stanovnika	Broj domacinstva	Obrazovanih stanovnika	Ekonomsko aktivna lica	Broj zaposlenih	Broj nezaposlenih
Klaster 1	-10,75	389,00	2.887,66	31.605,38	9.530,81	25.776,94	13.690,53	8.975,25	4.715,28
Klaster 2	142,30	258,04	2.208,70	27.936,07	6.617,44	20.674,78	7.354,19	3.654,52	3.699,67
Klaster 3	-24,43	51,13	1.721,13	6.064,35	1.903,87	4.908,00	2.241,96	1.290,96	951,00

Izvor: Rezultati dobijeni primenom SPSS – a

Za bolju interpretaciju klastera, izračunavaju se centroidi klastera za uključene promenljive. Centroidi su prestavljeni na Slici 2.4. u formi profila klastera. Jasno se vidi da su prvi i drugi klaster sličniji međusobom, za razliku od trećeg klastera koj je prilično različit. Kao nestandardne opservacije iz analize su isključene opštine Aerodrom i Centar. Ovo su opštine koje pripadaju glavnom gradu Makedonije, Skoplju, i izuzetno su razvijene, postoji veliki broj stanovnika, mnogo prodavnica u maloprodaji, mnogo obrazovanih lica i mnogu zaposlenih lica. Njihovo isključenje iz analize omogučilo je bolje rezultate u formiranju klastera.

Preko izračunatih vrednosti centroida, može se zaključiti da su opštine u prvom klasteru najrazvijene opštine u Makedoniji. Ovde spadaju uglavnom velike opštine, gde je prosečni broj stanovnika 31605 . Broj obrazovanih stanovnika u klasteru je visok, 25776 stanovnika, a ovaj klaster isto tako ima najveći broj zaposlenih 8975,

kao i ekonomsko aktivnih lica, 13690. Problem u ovom klasteru je negativni prirodni priraštaj[10] stanovnika, -10,75, odnosno pojava demografskog starenja populacije.

Opštine iz drugog klastera su slične opštinama iz prvog, iako su relativno manje razvijene. Ovom klasteru pripadaju manje opštine, jer prosečan broj stanovnika u klasteru je 27936. Interesantno je da broj zaposlenih i broj nezaposlenih lica je skoro isti, što bi značilo da ove opštine imaju visoku nezaposlenost. Opštine iz prvog i trećeg klastera, ove opštine imaju visoki prirodni priraštaj, 142,3. Kada uključimo promenljivu broj albanskih stanovnika i izračunamo centroide ove promenljive za tri klastera, dobijemo da prosečan broj albanskih stanovnika u prvom klasteru je 1283 u drugom 16987 i u trećem klasteru 298. Ovo objašnjava činjenicu zašto postoji visoki prirodni prirast u drugom klasteru, jer je veći broj dece karakteristika albanskih domaćinstava. Zaključak je da drugi klaster broji razvijene opštine sa (relativno) značajnim delom albanskog stanovništva.

Treći klaster su opštine sa manjem brojem stanovnika, u proseku 6064. Osnovna delatnost stanovnika ovih opština je poljoprivreda. To može da se zaključi, jer je broj poljoprivrednih domaćinstva visok, odnosno procenat poljoprivrednih domaćinstva u ukupnom broju domaćinstava je 90,4%. Ovo su relativno nerazvijene opštine, gde je broj prodavnica u maloprodaji veoma mali, kao i veliki negativni prirodni priraštaj od -24,43 stanovnika. To ukazuje da je u ovim opštinama veoma izražen proces demografskog starenja stanovništva .

Na osnovu ovih karakteristika, izvršena je klasifikacija 84 opština Makedonije i dobijene su informacije koje mogu biti osnova za kreiranje efikasne nacionalne ekonomske i razvojne politike.

2.8.2.Primena klaster analize k - sredina za klasifikaciju opština Makedonije

Klaster analiza k - sredina se primenjuje kada su u pitanju velike baze podataka, jer za razliku od hijerahijske klaster analize, ova analiza ne vrši prethodne proračune, ne nalazi matrice rastojanja ili sličnosti između svih parova objekata. Klaster analiza k - sredina omogućuje premeštanje objekta iz jednog klastera u drugi, kroz seriju ponovljenih (iterativnih) koraka u procesu dobijanja konačnog rešenja, jer ovaj metod pretstavlja klastering metod realokacija.

[10] Prirodni priraštaj prestavlja razliku između ukupnog broja rođenih stanovnika i ukupnog broja umrelih stanovnika. Iskazuje se u broj stanovnika.

k - sredina klaster analiza koristi Euklidovo rastojanje. Istraživač mora unapred da naznači broj klastera. Izbor jezgra klastera može se obaviti na dva načina. Prvi je kada se klasteri odrede od strane istraživača, kada se najpre koriste hijerahijski klaster algoritmi da bi se dobio broj klastera i onda se od tih rezultata generiraju jezgra klastera. Drugi način dobijanja jezgra klastera je njihovo generiranje iz uzorka, ili na sistematski način ili jednostavno preko slučajne selekcije. Izbor jezgra klastera je vrlo važan jer za različita jezgra klastera dobijaju se različita klaster rešenja.

Primer koji sledi prikazuje primenu k - sredina klaster analize u klasifikaciji makedonskih opština. Baza podataka je ista kao i kod hijerahijske klaster analize, a i odabrane promenljive. Postupak pripreme za klaster analizu, odnosno, sve odluke oko standardizacije promenljivih, nestandardnih opservacija, kao i pretpostavke o reprezentivnosti uzoraka, multikolinearnosti promenljivih i autokorelaciji objekata važe kako i kod hijerahijske analize. Cilj k-sredina klaster analize je isti kao i kod hijerahijske klaster analize, a to je da se izvrši klasifikacija 84 opštine Makedonije na osnovu ispitanih karakteristika, a preko centroida klastera za selektiranje promenljivih kako bi se dobile informacije koje bi bile osnova za kreiranje nacionalne ekonomske i razvojne politike. Na ovaj način mogu se uporediti rezultati primene dva različita metoda.

Hijerarhijska klaster analiza ne dozvoljava pregrupisavanje objekta u drugi klaster, kad se za taj objekat odredi pripadnost jednom klasteru, i pored toga ako bi pregrupisavanje u novi klaster predstavljalo bolje rešenje. Prednost k-sredina klaster analize je u tome da omogučava da se pregrupisavanjem dobije bolje rešenje. Naravno, rezultati hijerahijske analize su vrlo važni, jer oni otkrivaju broj klastera i početna jezgra klastera, koje su osnova u k-sredina klaster analize.

Za k-sredina klaster analizu, jezgra klastera ili centroidi klastera su oni iz hijerahijske analize, prikazane na Slici 2.4. SPSS softver omogućava da se rešenja kreiraju na dva načina:

1. kada se naznače jezgra klastera dobijena iz prethodne analize
2. kada program slučajno izabere tačke klastera.

Sprovedena je analiza na oba načina, ali prvi način je dao bolje rezultate, što je i bilo očekivano, jer već postoji dobra osnova – hijerahijska analiza, za izbor početnnih jezgara klastera. U sledećem delu rada prikazani su rezultati dobijeni preko inicijalnih jezgara klastera iz hijerahijske klaster analize.

Najpre, SPSS prikazuje tabelu koja sadrži vrednosti inicijalnih jezgara klastera. To su vrednosti na osnovu kojim je dobijena Slika 2.4. Sledeći rezultati su prikazani u Tabela 2.6.

Tabela 2.6. Prikaz promena u procesu grupisanja

Iteracija	Promena centroida klastera		
	1	2	3
1	37662,708	6867,836	2174,586
2	11534,688	2900,046	758,158
3	2496,292	1651,442	247,462
4	0,000	572,981	261,378
5	0,000	571,381	260,149
6	0,000	0,000	0,000

Izvor: Rezultati dobijeni primenom SPSS – a

Konvergencija nastaje kod malih ili nepostojećih vrednosti u centrodu klastera. U našem primeru postoji 6 promena (iteracija).

U Tabeli 2.7. je dat prikaz pripadnosti svake opštine klasteru i njeno rastojanje od centroida klastera. Opštine koje imaju velike vrednosti rastojanja su nestandardne opservacije i manje su reprezentativni objekti u klasteru nego opštine koje imaju manje vrednosti rastojanja.

Osim pripadnosti klasteru, postoji i prikaz centroida novih klastera. Na osnovu ovih centroida može se izvršiti profiliranje novih klastera.

Može se uočiti da se rezultati k - sredina klaster analize razlikuju od rezultata hijerarhijske klaster analize. Uglavnom, promene su nastale u drugom klasteru, koji je u hijerarhijskoj klaster analizi bio veoma sličan prvom klasteru, dok kod k - sredina klaster analize, ovaj klaster se razlikuje od prvog klastera.

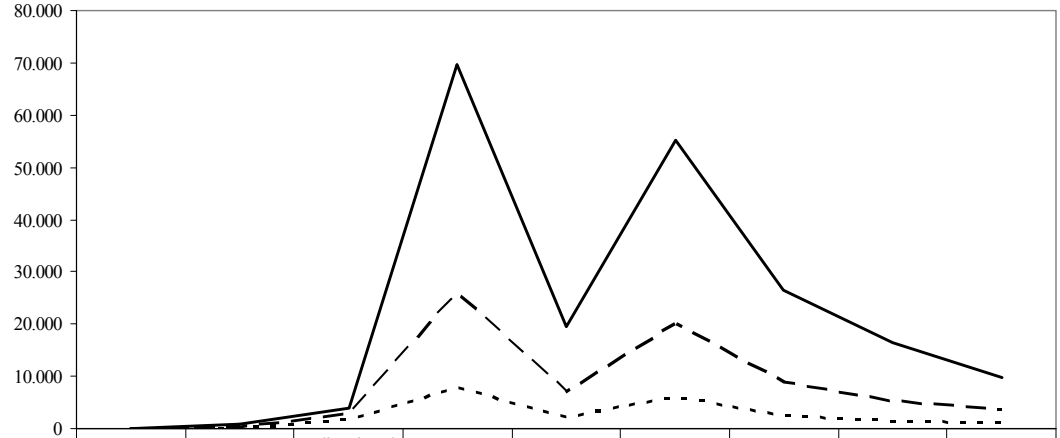
Tabela 2.7. Pripadnost opštine klasteru

Opština	Klaster	Rastojanje	Opština	Klaster	Rastojanje
1:Butel	2	15391,017	42:Kičevo	2	6323,897
2:Gazi Baba	1	6182,804	43:Konče	3	5771,198
3:Gorče Petrov	2	25170,307	44:Kočani	2	19611,935
4:Karpoš	1	14050,502	45:Kratovo	3	4336,915
5:Kisela Voda	1	15849,579	46:Kriva Palanka	2	6003,614
6:Saraj	2	11079,988	47:Krivogaštani	3	2004,910
7:Čair	1	13107,391	48:Kruševo	3	2868,929
8:Šuto Orizari	2	8308,346	49:Kumanovo	1	48582,522
9:Aračinovo	3	4304,297	50:Lipkovo	2	6201,449
10:Berovo	3	9462,532	51:Lozovo	3	6692,478
11:Bitola	1	42225,438	52:Mavrovo i Rostuša	3	1690,797
12:Bogdanci	3	2685,931	53:Mak. Kamenica	3	1060,052
13:Bogovinje	2	5961,812	54:Makedonski Brod	3	835,752
14:Bosilovo	3	9729,333	55:Mogila	3	1394,644
15:Brvenica	3	10131,541	56:Negotino	2	8245,118
16:Valandovo	3	6611,872	57:Novaci	3	5556,556
17:Vasilevo	3	6612,924	58:Novo Selo	3	5704,028
18:Vevcani	3	7383,784	59:Oslomej	3	3348,614
19:Veles	1	18274,127	60:Ohrid	1	17623,594
20:Vinica	2	7482,386	61:Petrovec	3	592,217
21:Vranešnica	3	8815,758	62:Pehčevo	3	2825,959
22:Vrapčište	2	6183,221	63:Plašnica	3	5039,938
23:Gevgelija	2	5307,821	64:Prilep	1	16215,851
24:Gostivar	1	15599,351	65:Probištip	2	12345,992
25:Gradsko	3	5426,897	66:Radoviš	2	5675,362
26:Debar	2	10197,892	67:Rankovce	3	5039,784
27:Debarca	3	2884,393	68:Resen	2	11532,998
28:Delčevo	2	10492,607	69:Rosoman	3	4815,897
29:Demir Kapija	3	4239,123	70:Sveti Nikole	2	9105,422
30:Demir Hisar	3	4079,259	71:Sopište	3	3093,104
31:Dojran	3	5901,064	72:Staro Nagoricane	3	4066,082
32:Dolneni	3	7874,145	73:Struga	1	17917,124
33:Drugovo	3	6058,138	74:Strumica	1	19323,142
34:Želino	2	7426,026	75:Studeničani	3	11261,085
35:Zajas	3	4696,181	76:Tearce	2	7360,403
36:Zelenikovo	3	5262,777	77:Tetovo	1	19850,347
37:Zrnovci	3	6128,001	78:Centar Župa	3	3248,793
38:Ilinden	3	11706,700	79:Časka	3	1247,206
39:Jegunovce	3	3919,113	80:Česinovo	3	1030,789
40:Kavadarci	2	20587,334	81:Čučer - Sandovo	3	1172,247
41:Karbinci	3	5170,049	82:Štip	1	28650,537

Izvor: Rezultati dobijeni primenom SPSS – a

Tabela 2.8. sadrži rezultate analize varijanse (ANOVA) za svaku promenljivu. Pri tom, F statistika prikazujea koliko svaka promenljiva doprinosi u diskriminaciji klastera. Nesignifikantne promenljive mogu da se odstrane iz analize, jer ne doprinose diferencijaciji klastera.

Slika 2.5. Centroidi klastera dobijeni klaster analizom *k* -sredina



	Prirodni prirast	Prodavnica u maloprodaji	Poljoprivredna domacinstva	Broj stanovnika	Broj domacinstva	Obrazovanih stanovnika	Ekonomsko aktivna lica	Broj zaposlenih	Broj nezaposlenih
Klaster 1	79,36	884,93	3.925,07	69.732,57	19.473,29	55.303,71	26.386,64	16.501,93	9.884,71
Klaster 2	74,55	255,68	2.853,68	25.920,27	7.041,95	20.073,50	8.954,68	5.203,86	3.750,82
Klaster 3	4,02	56,02	1.606,39	7.796,17	2.171,67	6.089,04	2.648,02	1.523,04	1.124,98

Izvor: Rezultati dobijeni primenom SPSS – a

Tabela 2.8. Analiza varijanse

Promenljiva	Klaster		Greška		F	Značajnost
	Sredina kvadrata	Stepeni slobode[11]	Sredina kvadrata	Stepeni slobode	Sredina kvadrata	Stepeni slobode
Prirodni prirast	53020,615	2	16301,616	79	3,252	0,044
Prodavnica u maloprodaji	3687678,666	2	38846,692	79	94,929	0,000
Privrednih domaćinstva	32867505,281	2	1776290,135	79	18,503	0,000
Stanovnici	20695705572,361	2	74424463,575	79	278,077	0,000
Domaćinstva	1612077121,089	2	5687971,796	79	283,419	0,000
Obrazovanih stanovnika	13048862987,091	2	44716364,282	79	291,814	0,000
Ekonomski aktivno stanovništvo	3028978464,011	2	16299214,417	79	185,836	0,000
Zaposlenih lica	1204380591,186	2	8588427,993	79	140,233	0,000
Nezaposlenih lica	414525692,891	2	3412999,533	79	121,455	0,000

Izvor: Rezultati dobijeni primenom SPSS – a

Ovi rezultati F – testova služe kasnije za dobijanje informacije koji objekti propadaju kojim grupama da bi se minimiziralo rastojanje u grupama, i da bi se maksimiziralo rastojanje između grupa, zbog čega su F - testovi korisni samo za

[11] Od engleske reči "degrees of freedom".

istraživačke ciljeve, a ne kao konvencionalni testovi značajnosti, jer postoje pretpostavke koje nisu ispunjene zbog samog procesa grupisanja.

U navedenom primeru sve promenljive koje se koriste u klaster analizi su statistički značajne. Najveću grešku ima promenljiva *Broj stanovnika*, šta znači da ova promenljiva najmanje doprinosi u difrencijaciji klastera. Ova tabela uglavnom se koristi da bi se proverila veličina grešaka.

Tabela 2.9. Raspored opština po klasterima

Klaster	1	14,000
	2	22,000
	3	46,000
Validne opservacije		82,000
Nedostajuće opservacije		0,000

Izvor: Rezultati dobijeni primenom SPSS – a

Tabela 2.9. je korisna za prikaz broja objekata u svakom klasteru. U situaciji kada broj objekta u klasteru nije balansiran, ili kada postoje klasteri sa manjim brojem objekata, istraživač eksperimtiše da bi dobio različito rešenje. Različiti rezultati mogu se dobiti preko postavljanja različitih inicijalnih jezgara klastera, ili preko menjanja broja klastera, pa čak i preko korišćenja baze podataka u kojoj objekti imaju različit redosled.

Konačni rezultat na Slici 2.6. prikazuje opštine razvrstane u tri klastera. Moguče je primetiti da su opštine primenom k - sredina klaster analize promenile klaster, u odnosu na rezultate hijerarhijske analize.

Dobijeni rezultati k - sredina klaster analize pokazuju da u prvom klasteru ima 14 opština, dok je prvi klaster u hijerarhijskoj klaster analizi imao 32 opštine. 9 opština je ostalo u prvom klasteru, 13 opština je otišlo u drugi klaster, 10 opština je prešlo u treći klaster, dok je 5 opština iz drugog klastera došlo u prvi klaster.

Sada, drugi klaster ima 22 opština, dok je u hijerarhijskoj analizi imao 27 opština. 9 opština je ostalo u drugom klasteru, 13 opština je prešlo u treći klaster, 5 opština je prešlo u prvi klaster, dok je 13 opština došlo iz prvog klastera.

Treći klaster broji 46 opština, dok je u hijerahijskoj analizi brojao 23 opštine. 23 opštine su ostale u istom, trećem klasteru, 10 opština je došlo iz prvog klastera, i 13 opština je došlo iz drugog klastera.

Slika 2.6. Poređenje rezultata hijerarhijske i k-sredina klaster analize

Klaster 1	
Hijerahijska analiza	**K-sredina analiza**
Gazi Baba	Gazi Baba
Karpoš	Karpoš
Kisela Voda	Kisela Voda
Bitola	Bitola
Veles	Veles
Ohrid	Ohrid
Prilep	Prilep
Strumica	Strumica
Štip	Štip
Butel	Butel
Gorče Petrov	Gorče Petrov
Vinica	Vinica
Gevgelija	Gevgelija
Delčevo	Delčevo
Kavadarci	Kavadarci
Kočani	Kočani
Kriva Palanka	Kriva Palanka
Negotino	Negotino
Probištip	Probištip
Radoviš	Radoviš
Resen	Resen
Sveti Nikole	Sveti Nikole
Bogdanci	Bogdanci
Bosilovo	Bosilovo
Valandovo	Valandovo
Demir Hisar	Demir Hisar
Dojran	Dojran
Ilinden	Ilinden
Kruševo	Kruševo
Mak.Kamen.	Mak.Kamen.
Novo Selo	Novo Selo
Cučer Sand.	Cučer Sand.

Klaster 2	
Hijerahijska analiza	**K-sredina analiza**
Saraj	Saraj
Šuto Orizari	Šuto Orizari
Bogovinje	Bogovinje
Vrapčište	Vrapčište
Debar	Debar
Želino	Želino
Kičevo	Kičevo
Lipkovo	Lipkovo
Tearce	Tearce
Aračinovo	Aračinovo
Brvenica	Brvenica
Vevčani	Vevčani
Zajas	Zajas
Zelenikovo	Zelenikovo
Jegunovce	Jegunovce
Mav. i Rost.	Mav. i Rost.
Oslomej	Oslomej
Petrovec	Petrovec
Plašnica	Plašnica
Sopište	Sopište
Studeničani	Studeničani
Centar Župa	Centar Župa
Čair	Čair
Gostivar	Gostivar
Kumanovo	Kumanovo
Struga	Struga
Tetovo	Tetovo

Klaster 3	
Hijerahijska analiza	**K-sredina analiza**
Berovo	Berovo
Vasilevo	Vasilevo
Vranešnica	Vranešnica
Gradsko	Gradsko
Debarca	Debarca
Demir Kapija	Demir Kapija
Dolneni	Dolneni
Drugovo	Drugovo
Zrnovci	Zrnovci
Karbinci	Karbinci
Konče	Konče
Kratovo	Kratovo
Krivogaštani	Krivogaštani
Lozovo	Lozovo
Mak. Brod	Mak. Brod
Mogila	Mogila
Novaci	Novaci
Pehčevo	Pehčevo
Rankovce	Rankovce
Rosoman	Rosoman
S. Nagoričane	S. Nagoričane
Čaška	Čaška
Češinovo	Češinovo

Centroidi novih klastera omogućuju profiliranje novih klastera, različitih od profila dobijenih u hijerarhijskoj klaster analizi. Prvi klaster uglavnom broji sve velike opštine, jer prosečan broj stanovnika po opštini u ovom klasteru je 69732. Ovo su razvijene opštine jer imaju najveći broj prodavnica u maloprodaji, 884, manji broj poljoprivrednih domaćinstava, 3925 ili ukupno 20,16% od svih domaćinstava. Da su ekonomski najrazvijenije opštine ukazuje i činjenica da je broj zaposlenih lica 16501, ili 62,54%, a samim tim ovaj klaster je sa najmanjim brojem nezaposlenih lica. Prirodni priraštaj je pozitivan i iznosi 79,36 stanovnika. Interesantan je podatak da su sve ove opštine imale visoku vrednost Mahalanobisovog odstojanja, i mogle bi da se označe i kao nestadardne opservacije. U primeru su samo opštine Aerodrom i Centar

isključene kako nestandardne opservacije. Ovo ukazuje na to da je grupisanje dobro izvedeno.

Drugi klaster ima 22 opštine, i posmtarno po broju stanovnika, ovu su manje opštine, jer je prosečni broj stanovnika u njima, 25920. I ove opštine imaju pozitivan prirodni priraštaj od 74,55 stanovnika, ali su manje razvijene jer broje 255 prodavnica u maloprodaji. Visoki je broj i poljoprivrednih domačinstva, 2853 ili 40,52%. Broj zaposlenih lica je manji nego u prvom klasteru, i iznosi 5203 lica, ili 58,11%.

Treći klaster je najveći i ima ukupno 46 opština. Čine ga najmanje i najnerazvijenije opštine. Prosečan broj stanovnika je 6089, dok je prirodni priraštaj veoma mali, 4 stanovnika. Prosečan broj prodavnica u maloprodaji je 56. Ovo su uglavnom opštine sa visokim brojoj poljoprivrednih domačinstva, 1606 ili 73,97%. Broj zaposlenih lica je manji u poređenju sa prethodnim klasterima, 1523 stanovnika ili 57,52%, šta znači da je nezaposlenost u ovim opštinama najveća, i iznosi 1124 stanovnika ili 42,48%.

Razlike u rešenjima iz hijerarhijske klaster analize i k-sredina klaster analize se javljaju, iz razloga šta hijerarhijski proces ograničava rezultate, jer ne dozvoljava da opštine menjanju klastere, ukoliko već pripadaju nekom klasteru.

Postavlja se pitanje, koji su rezultati bolji, da li oni dobijeni hijerarhijskom klaster analizom ili iz k-sredina klaster analizom. Hijerarhijska analiza koristi mere bliskosti a analiza k-sredina ne koristi ove mere, već uzima početna jezgra, centroide, i približava sve slične objekte u klaster čiji centroid je najmanje udaljen od objekata na osnovu Euklidovog rastojanja. Prednost k-sredina klaster analize je u tome da može da pregrupiše već jednom grupisani objekat, ukoliko ustanovi da je za taj objekt bolji drugi klaster. To je zato što se pri svakom spajanju objekata u klaster, izračunava novi centroid. Odluku koji će se rezultat korisiti donosi sam istraživač, koji najbolje poznaje cilj analize i teoretsku pozadinu problema.

2.8.3.Primena dvostepene klaster analize u definisanju klastera 200 makedonskih kompanija u odnosu na njihove karakteristike

Dvostepena klaster analiza kreira pod-klastere upotrebom hijerarhijske metode. Ukoliko se radi sa velikom bazom podataka, preporučuje se dvostepena klaster analilza, kao i u situaciji kada su podaci katerogorijski. Ova analiza u SPSS-u daje značajni broj izlaznih rezultata, kako i grafikone za važnost promenljivih.

Za prikaz dvostepene klaster analize koristićemo bazu podataka 200 makedonskih kompanija. Ova baza podataka je dobijena od Centralnog registra Makedonije, i sadrži 200 najuspešnih makedonskih kompanija. Cilj analize je da se dobije struktura glavnih grupa makedonskih kompanija definisanjem klastera kompanija na osnovu sledećih kvantitativnih promenljivih, *Ukupni prihod u 2007*, *Ukupni prihod u 2006*, *Stopa rasta prihoda 2007/2006*, *Profit pred oporezivanje u 2007* i *Broj zaposlenih*, kao i jedne kategorijske promenljive, *Vrsta industrije*.

Za pripremu dvostepene klaster analize, prvo se postavlja pitanje **standardizacije promenljivih**. Ova analiza vrši standardizaciju kontinuiranih promenljivih, ali konačni rezultati su prikazani u originalnim vrednostima. Pitanje oko nestandardnih opservacija se rešava sa Mahalanobisovom D^2 merom. Ova mera ukazuje na to da su manje od desetak kompanija potencijalne nestandardne opservacije. Ove kompanije su zadržane u analizi.

Korelaciona matrica **Pearsonovih koeficijenata** uključenih promenljivih ukazuje da postoji povezanost između promenljivih. To se i očekuje, posebeno između prihoda i profita. I u ovom primeru ne postoji situacija gde nekoliko promenljivih čine jedan ili više faktora, ako faktor definišemo kao skupiua koreliranih promenljivih. Ukoliko se uoči postojanje multikolinearnosti, potrebno je da se reduciraju promenljive u okviru svakog faktora, ili skupa koreliranih promenljivih. Zbog toga, analiza može da se nastavi i ako nije ispunjen ovaj preduslov.

Analiza uključuje 5 kvantitativnih promenljivih i 1 kategorijsku promenljivu. Kada su jedna ili više promenljivih kategorijske, onda se koristi **mera rastojanja (prirodnog logaritma) funkcije verodostojnosti**. Ukoliko su sve promenljive kvantitativne, onda se koristi Euklidovo rastojanje. Za određivanje broja klastera koristi se ili **Bajesov** ili **Ekejkov informacioni kriterijum**. Broj klastera može i automatski da se odredi od strane istraživača. Analiza prikazuje sledeće rezultate:

Statistika prikazana u Tabeli 2.10. daje vrednosti Bajesovog kriterijuma i promene Bajesovog kriterijuma za moguća rešenja broja klastera. U automatsko određivanje broja klastera, ponuđeni su različiti kriterijumi, odnosno najbolje rešenje broja klastera je ono koje ima najmanju vrednost za Švarc-Bajesov informacioni kriterijum, ili najmanju vrednost Ekejkovog informacionog kriterijuma. Statistički softver SPSS bira ono rešenje koje daje vrlo značajnu vrednost promene Bajesovog kriterijuma i veliku vrednost za stope mere rastojanja. Studije u kojima su sprovedene simulacije

potvrđuju da stopa mere rastoranja, koja predstavlja kombinovani kriterijum, daje bolje rezultate od individualne vrednosti dobijene preko Bajesovog ili Ekejkovog informacionog kriterijuma.

Tabela 2.10. Rezultati za automatskog grupisanja

Broj klastera	Švarc-Bajesov kriterium (BIC)	Promena BIC-a (a)	Stopa promene BIC-a (b)	Stopa mere rastojanja (c)
1	1272,747			
2	990,722	-282,026	1,000	1,839
3	884,535	-106,186	0,377	1,152
4	806,071	-78,464	0,278	2,168
5	825,571	19,500	-0,069	1,906
6	884,966	59,395	-0,211	1,212
7	952,054	67,088	-0,238	1,394
8	1029,408	77,354	-0,274	1,034
9	1107,607	78,199	-0,277	1,146
10	1189,023	81,416	-0,289	1,209
11	1274,238	85,216	-0,302	1,015
12	1359,726	85,488	-0,303	1,318
13	1449,533	89,808	-0,318	1,373
14	1543,036	93,503	-0,332	1,189
15	1638,112	95,076	-0,337	1,126

(a) Promene su iz prethodnog broja klastera u tabeli.

(b) Stope promene prikazuju relativnu promenu za rešenje dobijeno dvostepenom klaster analizom.

(c) Stope promene mere rastojanja se baziraju na postojećem broju klastera u uporedbi sa prethodnim brojem klastera.

Izvor: Rezultati dobijeni primenom SPSS – a

U navednom primeru, najbolje rešenje je sa 4 klastera, jer daje visoku vrednost stope mere rastojanja. Ovo rešenje nije u saglasnosti sa drugim merama, Bajesovog informacionog kretiriuma, a to je zbog toga jer SPSS algoritam daje rešenje na osnovu vrednosti stope mere rastojanja, čak i kada ovo rešenje nije isto sa rešenjem Bajesovog informacionog kriterijuma. To je upravo slučaj sa analiziranim podacima. Kada se rešenja razlikuju, SPSS algoritam smatra da dobit od informacije koja bi se dobila ukoliko bi se uzelo više klastera od naznačenog broja klastera Bajesovom informacionim kriterijumom nije dovoljno velika da bi kompenzirala povećanu kompleksnost modela.

Tabela 2.14. Pripadnost kompanija klasterima

Kompanija/Klaster		Kompanija/Klaster	
Okta AD Skopje	1	Adrijus DOOEL Skopje	2
Feni industries AD Kavadarci	1	MZT Hepos AD Skopje	2
Makpetrol AD Skopje	1	ADE Skopsko pivo Tetovo	2
Makedonski telekom AD Skopje	1	Swisslion-Agroplod AD Resen	2
EVN Macedonia AD Skopje	1	Ekstra-Skopsko Kosel DOOEL Ohrid	2
T-Mobile Macedonia AD Skopje	1	Metro AD Skopje	2
Macedonian power plants JSC Skopje	1	Bas tuti friuti DOOEL Skopje	2
Arcelormittal Skopje (CRM) AD Skopje	2	Brako DOO Veles	2
Arcelormittal Skopje (HRM) AD Skopje	2	Fabika Karpos AD Skopje	2
Makstil AD Skopje	2	Mlekara Zdravje Radovo, Strumica	2
Usje AD Skopje	2	Metalopromet DOOEL Strumica	2
Pivara Skopje AD Skopje	2	Prototip DOOEL Skopje	2
Alkaloid AD Skopje	2	MZT Learnica AD Skopje	2
Igm-trade DOO Kavadarcki	2	Helmateks AD Strumica	2
Dojran stil DOO Dojran	2	Tondach-Makedonija AD Vinica	2
Skopski leguri DOOEL Skopje	2	Kiro Kucuk AD Veles	2
11 Oktomvri AD Kumanovo	2	Zdenka DOOEL Negotino	2
Tutunski kombinat AD Skopje	2	Fersped AD Skopje	3
Dil Petrol DOOEL Stip	2	Euro tabak DOO Skopje	3
Toplifikacija AD Skopje	2	Cosmofon AD Skopje	3
Brilijant DOOEL Stip	2	Granit AD Skopje	3
Silmak DOOEL Tetovo	2	Makedonski aviotransport AD Skopje	3
Swisslion DOO Skopje	2	Tec Negotino AD Negotino	3
Mlekara AD Bitola	2	NLB Lizing DOOEL Skopje	3
Zito luks AD Skopje	2	Knauf-radika AD Debar	3
Vinarska vizba Tikves AD Skopje	2	JP za stop. so stanben i deloven prostor Skopje	3
Bomex DOO Skopje	2	Terna AD Skopje	3
Pekabesko AD Skopje	2	Makosped AD Skopje	3
ZK Pelagonija AD Bitola	2	Makoten DOOEL Gevgelija	3
F.I. Vitaminka AD Prilep	2	EFT Makedonija DOOEL Skopje	3
EMO AD Ohrid	2	Sasa DOOEL Makedonska kamenica	3
Teteks AD Tetovo	2	Alma-m DOO Skopje	3
Mega DOOEL Skopje	2	Makedonska posta AD Skopje	3
EMO DOOLE Ohrid	2	JP Vodovod i kanalizacija Skopje	3
Prilepska pivarnica AD Prilep	2	Bucim DOOEL Radovis	3
Droga kolinska DOOEL Skopje	2	Beton AD Skopje	3
Vest DOOEl Bitola	2	Media print Makedonija DOO Skopje	3
Strumica tabak AD Strumica	2	Pexim DOOEL Skopje	3
Promes DOO Skopje	2	Mavrovoinzenerig DOOEL Skopje	3
Alayans uan Makedonija AD Kavadarci	2	Peas Macedonia Skopje	3
FI Blagoj Gorev JSC Veles	2	JP Komunalna higiena Skopje	3
MIK Sveti Nikole DOO Sveti Nikole	2	DS Iskra steel construction DOO Kumanovo	3
Imperijal-tabako AD Valandovo	2	Indo minerals&metals DOOEL Skopje	3
Kontihidroplast DOOEL Gevgelija	2	Neocom AD Skopje	3
Ideal sipka DOO Bitola	2	Makinvest DOO Skopje	3
Pucko petrol DOO Makedonski brod	2	Dzesas insaat tidzared i sanaji AS Podruz. SK	3
Komuna AD Skopje	2	Alfeks inzenering DOO Skopje	3
Riomk bomeks - refraktori AD Pehcevo	2	DGU Pelister Bitola DOO Bitola	3
TGS Tehnicki gasovi AD Skopje	2	On.net DOO Skopje	3
Evropa AD Skopje	2	Pakom kompani DOOEl Skopje	3

Kompanija/Klaster		Kompanija/Klaster	
Leov kompani DOOEL Veles	2	Dauti komerc AD Skopje	3
Makprogres DOO Vinica	2	Mlaz AD Bogdanci	3
FHL Mermeren kombinat AD Prilep	2	Makedonija Turist AD Skopje	3
Pavor DOOEL Veles	2	Kiro D. Dandaro AD Bitola	3
Bunar petrol DOO Gostivar	3	Centro union DOO Skopje	4
Kvalitet-prom DOOEl Kumanovo	3	Gorenje DOOEL Skopje	4
Senker DOOEL Skopje	3	Podravka DOOEL Skopje	4
Pelagonija Inzinering DOOEL Skopje	3	Energomarket DOO Skopje	4
Publicis DOO Skopje	3	Tabako-promet BM DOOEL Valandovo	4
Int trejd DOOEL Kocani	3	Eurotrejd DOO Skopje	4
Lukoil Macedonia LTD Skopje	4	Montenegro DOO Gostivar	4
Tinex-mt DOOEL Skopje	4	Ka-dis DOO Skopje	4
Veroupulos DOOEL Skopje	4	Kolid kompani AS DOO, s. Kolesino N. Selo	4
Porshe Makedonija DOOEL Skopje	4	German PX DOO Skopje	4
Gemak-trade DOOEL Skopje	4	Gamatroniks DOOEL Skopje	4
ZEGIN DOO Skopje	4	Automakedonija AD Skopje	4
Skopski Pazar AD Skopje	4	Marija treid DOO Veles	4
Euro aktiva DOO Skopje	4	Gross prom DOO Skopje	4
KAM DOOEL Skopje	4	Grosist DOOEL Bitola	4
Makautostar DOOEL Skopje	4	Agrokumanovo AD Kumanovo	4
Nelt st DOOEL Skopje	4	Swisslion Mak DOO Skopje	4
AD D-r Panovski Skopje	4	Agroefodia DOOEL Strumica	4
Pharmacy Zegin farm Skoopje	4	Kemo-farm DOOEL Skopje	4
Promedika DOO Skopje	4	Avto kuka DOO Skopje	4
Euro media DOO Skopje	4	Krka-farma DOOEL Skopje	4
Euroimpex DOO Skopje	4	Zito DOOEL Veles	4
KIK DOO Skopje	4	Mepso AD Skopje	*
Ekspanda DOOEL Skopje	4	Kameni most komunikacii AD Skopje	*
Makoil DOOEL Skopje	4	ADG Mavrovo Skopje	*
Automobile sk DOOEL Skopje	4	Zito vardar AD Veles	*
Elektroelement DOO Skopje	4	JP Makedonski sumi Skopje	*
Jaka 80 AD Radovis	4	Jaka tabak AD Radovis	*
Replek AD Skopje	4	Hypo-alpe-adria-lizing DOOEL Skopje	*
Ramstore Macedonia DOO Skopje	4	4 Noemvri AD Bitola	*
Filip Moris Skopje DOOEL Skopje	4	Tutunski kombinat - cigari DOOEL Prilep	*
Toyota avto centar DOOEL Skopje	4	Public transportation enterprise Skopje	*
Mako-market DOO Skopje	4	Haier Makedonija trejd DOOEL Skopje	*
Merkur Makedonija DOO Skopje	4	JP Makedonijapat Skopje	*
Makedonija Lek DOO Skopje	4	Germanos Telekom AD Skopje	*
Mi-da motors DOO Skopje	4	Lek DOOEL Skopje	*
Libra 1 AG Skopje	4	Lotarija na Makedonija AD Skopje	*
Euromilk DOO Skopje	4	JP Makedonska radio televizija Skopje	*
Tediko super DOOEL Skopje	4	TCG learnica DOOEL Ohrid	*
Internacional food bazar DOO Skopje	4	AD Negotino Negotino	*
Avtonova DOO Skopje	4	Prima.mk DOO Skopje	*
Rudine-mm DOO Skopje	4	Fruktal mak AD Skopje	*
JUS MB DOO Skopje	4	Fabrika za kvasec i alkohol AD Bitola	*
Makkar DOO Skopje	4	SAF Energiferzorgungslezngen GmbH podr.	*
Inter tobako DOOEL Skopje	4	Alfa kopi DOOEL Skopje	*
Kola DOOEL Skopje	4	MIA Beverages DOO Skopje	*

** Isključene opservacije radi nedostatka podataka*

Izvor: Rezultati dobijeni primenom SPSS – a

Tabela 2.11. sadrži informaciju o broju i veličinu klastera. Takođe, prikazan je i broj isključenih kompanija. Za 24 kompanije raspoloživi podatci nisu bili dovoljni za njihovo grupisanje u klastere.

Tabela 2.11. Raspored kompanija po klasterima

Klaster	Broj objekta	% od kombinovanih klastera	% od ukupno
1	7	4,0%	3,5%
2	64	36,4%	32,0%
3	43	24,4%	21,5%
4	62	35,2%	31,0%
Kombinovani klasteri	176	100,0%	88,0%
Isključene opservacije	24		12,0%
Ukupno	200		100,0%

Izvor: Rezultati dobijeni primenom SPSS – a

Tabela 2.12. prikazuje centroide srednje vrednosti, za sve analizirane kvantitativne promenljive za svaki klaster zasebno.

Tabela 2.12. Centroidi klastera za kvantitativne promenljive

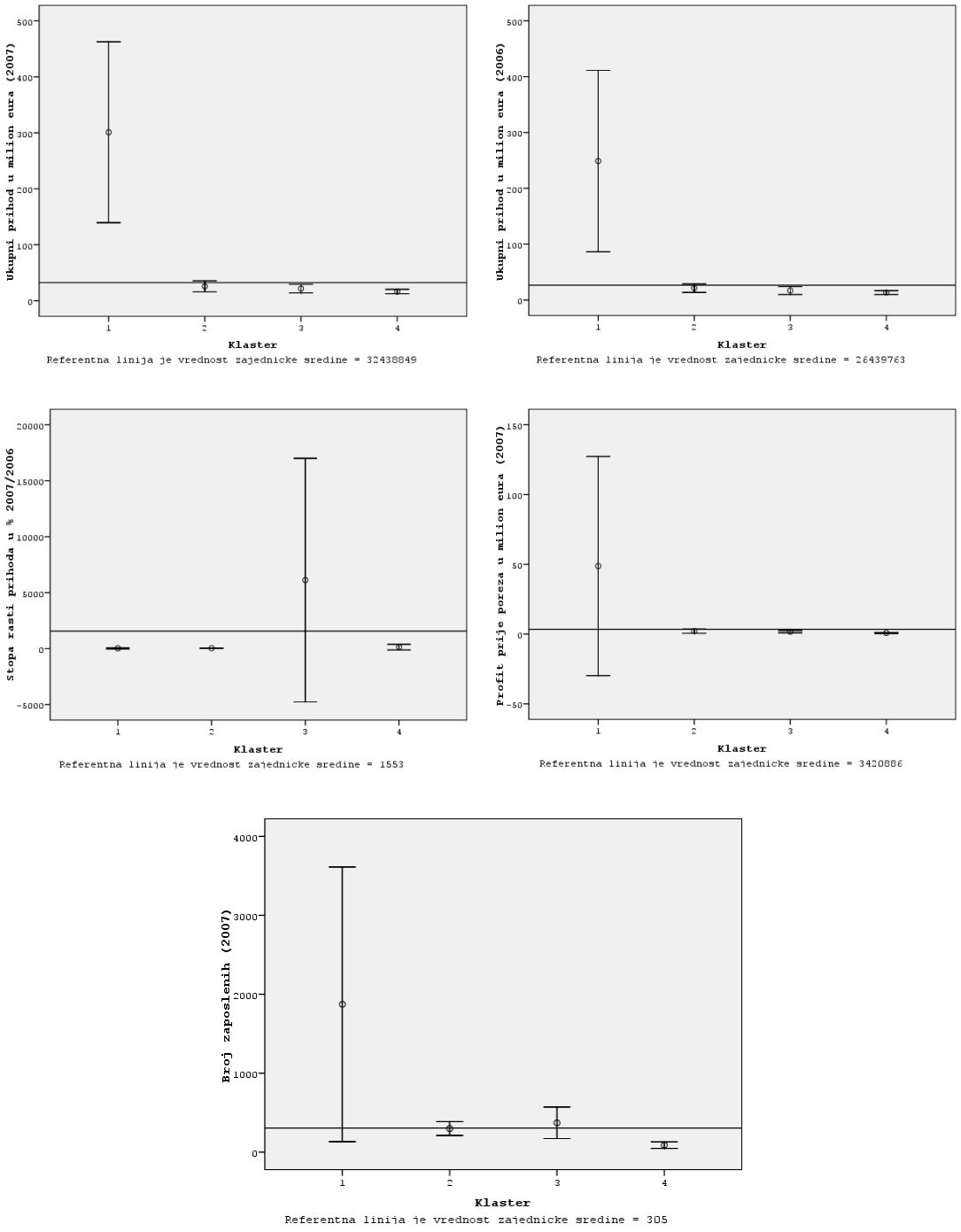
Promenljiva	Klaster 1	Klaster 2	Klaster 3	Klaster 4
Ukupni prihod u EUR (2007)	301.044.003	25.714.168	21.771.629	16.452.301
Ukupni prihod u EUR (2006)	248.912.033	21.215.261	16.912.291	13.322.724
Stopa rasti prihoda 2007/2006	27	33	6.119	129
Profit pre oporezivanja u EUR (2007)	48.756.644	2.103.284	1.714.597	845.832
Broj zaposlenika (2007)	1.872	298	371	88

Izvor: Rezultati dobijeni primenom SPSS – a

Prvi klaster broji 7 kompanija koje su po centroidima klastera: najveće, najprofitabilnije kompanije, sa najvećim brojem zaposlenih i umerenom stopom rasta. Ovo su kompanije koje se bave proizvodnjom, električnom energijom i telekomunikacijama. To znači da su ovo najveće i najuspešnije kompanije u Makedoniji.

Drugi klaster je najveći i broji 64 kompanija koje imaju značajno niži prihod i profit od kompanija iz prvog klastera. Ove kompanije imaju umereni rast i manji broj zaposlenih. Ovo su kompanije srednjeg stabilnog rasta.

Slika 2.8. 95% interval pouzdanosti za analizirane kvantitativne promenljive unutar klastera



Legenda

○ - Aritmetička sredina

I - Granice 95% intervala pouzdanosti

Izvor: Rezultati dobijeni primenom SPSS – a

Treći klaster broji 43 kompanija koje imaju nešto niži prihod i profit od kompanija iz drugog klastera. Broj zaposlenih je nešto veći nego kod kompanija iz drugog klastera, ali razlika ovoga klastera je u tome što ima visoku stopu rasta prihoda. To ukazuje da su ovo kompanije u rastu, sa potencijalom da se razviju u još uspešnije kompanije.

Četvrti klaster broji 62 kompanije, koje imaju najniži prihod, profit i broj zaposlenih. Značajno je napomenuti da ove kompanije imaju visoki rast prihoda. Ovo je klaster koji ima manje kompanija od drugog klastera, ali stopa rasta prihoda je značajno visoka, što ponovo znači da se može očekivati rast i kod ovih kompanija.

Tabela 2.13. Raspored kompanija po klasterima i oblastima delatnosti

Frekvencije	Klaster	Oblast delatnosti (kod)										
		3	4	5	6	7	8	9	10	11	12	15
Apsolutne frekvencije	Klaster 1	0	3	2	0	0	0	2	0	0	0	0
	Klaster 2	0	64	0	0	0	0	0	0	0	0	0
	Klaster 3	3	0	3	9	0	2	9	2	12	1	2
	Klaster 4	0	0	0	0	62	0	0	0	0	0	0
	Kombinovano	3	67	5	9	62	2	11	2	12	1	2
Relativne frekvencije	Klaster 1	0	4,5	40	0	0	0	18,2	0	0	0	0
	Klaster 2	0	95,5	0	0	0	0	0	0	0	0	0
	Klaster 3	100	0	60	100	0	100	81,8	100	100	100	100
	Klaster 4	0	0	0	0	100	0	0	0	0	0	0
	Kombinovano	100	100	100	100	100	100	100	100	100	100	100

Izvor: Rezultati dobijeni primenom SPSS – a

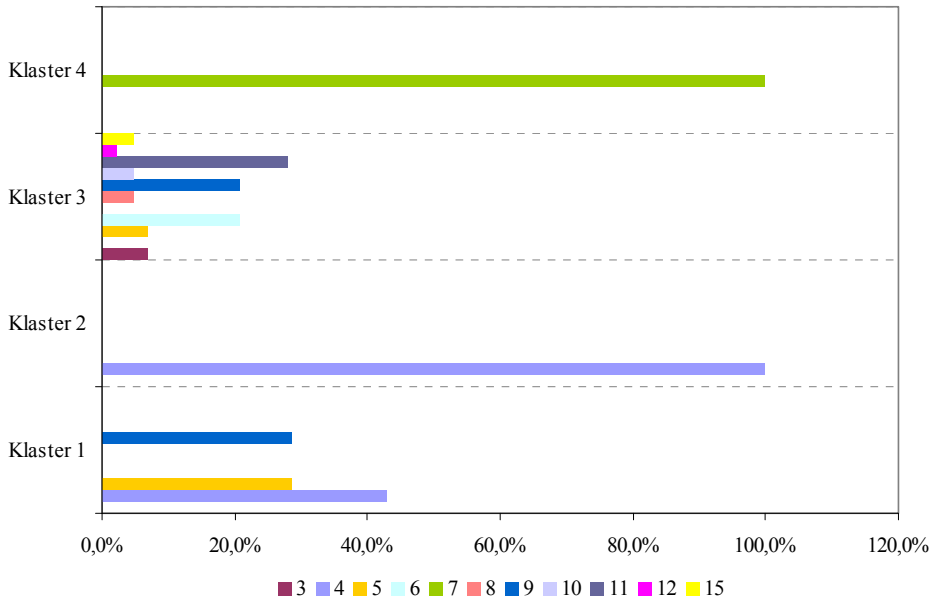
Kodovi oblasti su sledeće: 1 - poljoprivreda, šumarstvo i lov, 2 - ribarstvo, 3 - rudarstvo i rude, 4 - proizvodnja, 5 – električna energija, gas, snabdevanje vodom, 6 - građevina, 7 - trgovina na malo i veliko, popravka motornih vozila, proizvodi za ličnu upotrebu i upotrebu u domaćinstvu, 8 - hoteli i restorani, 9 - transport, skladištenje i komunikacije, 10 - finansisko posredništvo, 11 - nekretnine, zakup, i biznis aktivnosti, 12 - javna administracija i odbrana, socijalno osiguranje, 13 - obrazovanje, 14 - zdravstvo i zdravstveno osiguranje, 15 - ostale društvene i socijalne aktivnostii, lične usluge.

Kreirana je po jedna tabela za svaku kategorijsku promenljivu. Iz Tabele 2.13. se može videti da u prvom klasteru dominiraju kompanije iz oblasti sa kodom 4, a to je proizvodnja, zatim oblasti sa kodom 5, električna energija, i oblasti sa kodom 9, telekomunikacije. U drugom klasteru su sve kompanije iz proizvodnje. U trećem klasteru su kompanije iz skoro svih delatnosti. U četvrtom klasteru su sve kompanije iz oblasti sa kodom 7, trgovina.

U tabeli nema kompanija u oblastima delatnosti 1, 2, 13 i 14. To je ili zbog toga što analizom nisu obuhvaćene kompanije iz određene oblasti, ili kompanije iz ove oblasti nisu grupisane, jer su isključene zbog podataka koji nedostaju.

Raspored kompanija po klasterima je prikazan i grafički (Slika 2.7.), da bi se što bolje uvidela struktura modaliteta kategorijske promenljive klasterima.

Slika 2.7. Procentualna zastapljenost oblasti delatnosti po klasterima



Izvor: Rezultati dobijeni primenom SPSS – a

Analizirane kvantitativne promenljive prikazane su na dijagramima (Slika 2.8.).

Sa obzirom da se analiza zasniva na uzorku date su intervalne ocena sa pouzdanošću 0,95. Prvi grafikon prikazuje sredine za promenljivu *ukupan prihod u 2007. godini*. Može se videti da prvi klaster ima značajno veće vrednosti od ostalih klastera, gde je ova vrednost približno slična. Ista situacija važi i za promenljive: *Ukupni prihod u 2006*, *Profit pre oporezivanja* i *Broj zaposlenih*. Jedino *Stopa rasta prihoda* ima značajno veći interval poverenja u trećem klasteru, za razliku od ostalih klastera koji imaju umereni rast.

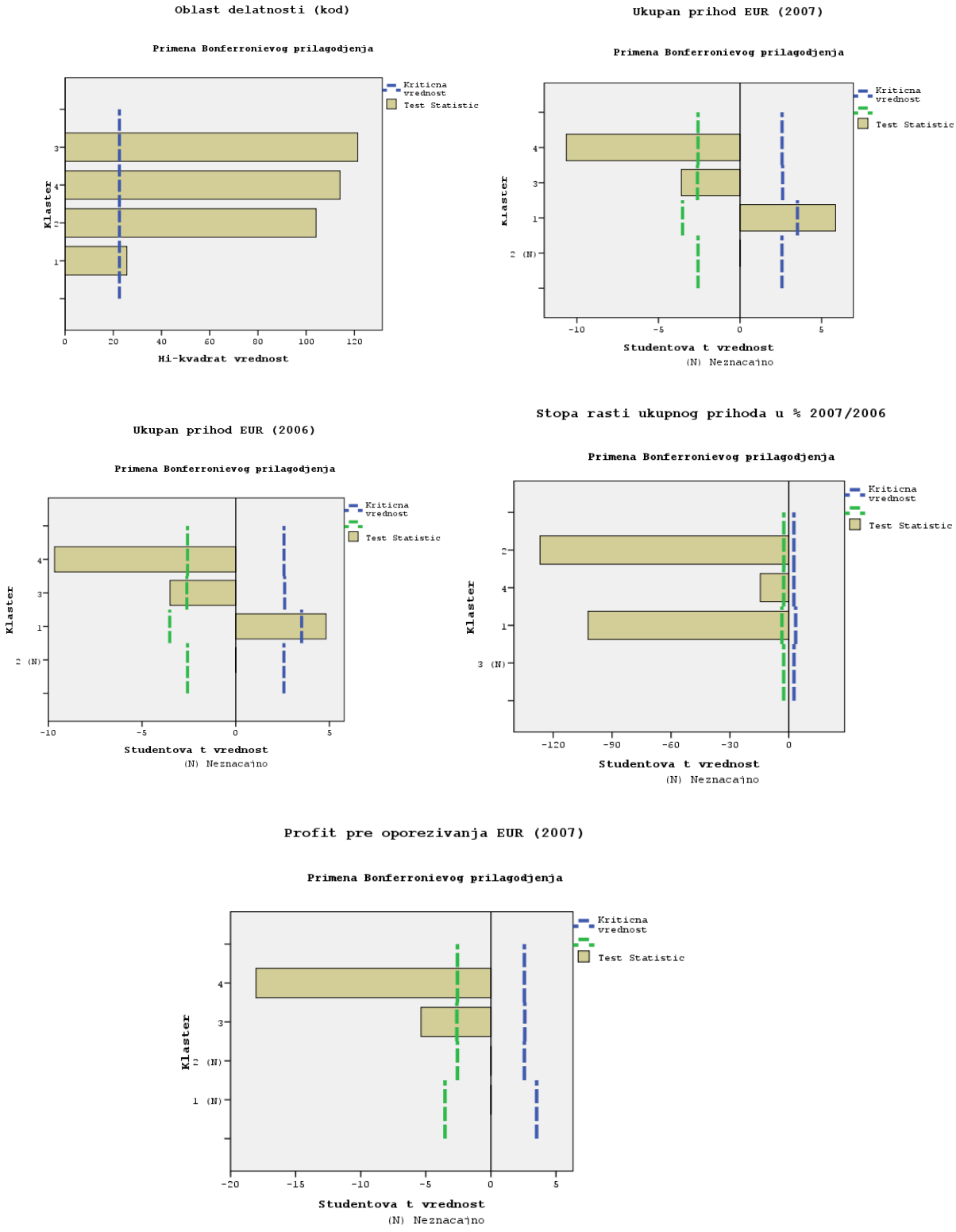
SPSS nudi još jednu grupu grafikona koji su rezultat dvostepene klaster analize i koji prikazuju značajnost promenljivih. Na X osi se prikazuje vrednost χ^2 ili Studentove t - statistike, a na Y osi promenljiva. Grafički prikaz sadrži i bar liniju.

Bar linija duža od kritične vrednosti promenljiva je kada je statistički značajna za diferenciranje klastera.

Prva promenljiva je kategorijska promenljiva za *Oblast delatnosti*. Vrednost χ^2 statistike ukazuje da je ova promenljiva statistički značajna za diferenciranje svih klastera, posebno za klastere 3, 4 i 2, a manje za klaster 1. Druga promenljiva, *Ukupan prihod u evrima za 2007. godinu*, i treća promenljiva, *Ukupan prihod u evrima za 2006. godinu*, značajno doprinose diferenciranju klastera 4, 3 i 1, dok ne doprinose u diferenciranju klastera 2. Četvrta promenljiva, *Stopa rasta ukupnog prihoda*, ima prilično veliki uticaj u diferenciranju klastera 2 i 1, njen je uticaj manji na diferenciranje klastera 4, dok nema nikakvog uticaja u diferenciranje klastera 3. Promenljiva *Profit pre oporezivanja u evrima za 2007. godinu*, značajno diferencira klastere 4 i 3, dok nije značajan za klastere 2 i 1. Poslednja promenljiva, *Broj zaposlenih za 2007. godinu*, značajna je samo u diferenciranju klastera 4, dok nema uticaj na klastere 3, 2 i 1.

U Tabeli 2.14. su prikazane svih 200 kompanija i njihova pripadnost klasterima.

Slika 2.9. Značajnost promenljivih po klasterima



Izvor: Rezultati dobijeni primenom SPSS – a

... if a gentleman walks into my rooms smelling of iodoform, with a black mark of nitrate of silver upon his right fore-finger, and a bulge on the side of his-hat to show where he has secreted his stethoscope, I must be dull indeed, if I do not pronounce him to be an active member of the medical profession.

Sherlock Holmes in "A Scandal in Bohemia"

3.Diskriminaciona analiza – Uvod

3.1.Diskriminaciona analiza – Uvod

Ukoliko postoje dve ili više grupa ili populacija i skup promenljivih, istraživač može imati cilj (zadatak) da utvrdi podskup promenljivih i odgovarajuće funkcije tog podskupa koje bi omogućile maksimalno razdvajanje između centroida grupa. Istraživačka[12] multivarijaciona procedura koja određuje promenljive i kreira smanjeni skup funkcija poznatih kao diskriminante ili diskriminantne funkcije zove se **diskriminaciona analiza**. Ova analiza razvija pravila za alokaciju ili dodeljivanje objekata ili opservacija jednoj ili više grupa. Multivarijaciona tehnika koja je bliže povezana sa diskriminacionom analizom zove se **klasifikaciona analiza**, pa se zato često tretiraju zajedno. **Ciljevi diskriminacione i klasifikacione analize** su sledeći:

1) Opisati, grafički (u tri ili manje dimenzija) ili algebarski, karakteristike koje razgraničavaju objekte (opservacije) iz nekoliko poznatih populacija. Traže se "diskriminante" koje imaju takve numeričke vrednosti da maksimalno razdvajaju populacije;

2) Sortiraju objekte (opservacije) u dve ili više definisanih grupa. Izvesti pravila koja se mogu koristiti da bi se optimalno razvrstali novi objekti u prethodno definisane grupe.

Diskriminaciona analiza rezultira izvođenjem diskriminacione funkcije. Diskriminaciona funkcija je linearna kombinacija dveju (ili više) nezavisnih promenljiva koje najbolje razdvajaju objekte (opservacije) u prethodno definisane grupe. Diskriminacija se sprovodi izračunavanjem pondera diskriminacione funkcije za svaku nezavisnu promenljivu, kako bi se maksimizirale razlike među grupa. Jednačina diskriminacione funkcije glasi:

$$Z_{jk} = a + W_1 X_{1k} + W_2 X_{2k} + \ldots + W_n X_{nk}$$

gde je:

Z_{jk} diskriminacioni skor diskriminacione funkcije j za objekt k;

a odsečak[13];

W_i diskriminacioni ponder nezavisne promenljive i;

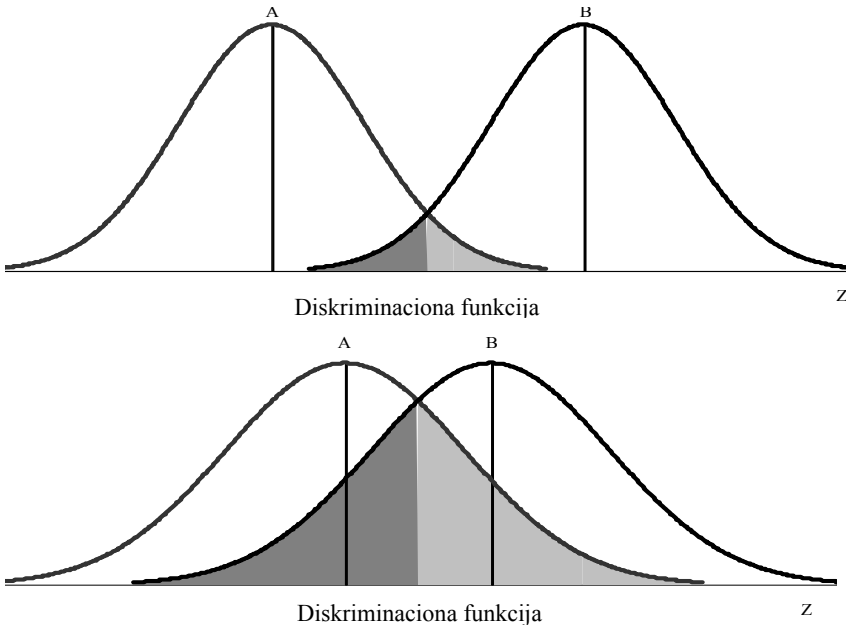
X_{ij} vrednost nezavisne promenljive i kod objekta k.

[12] Od engleske reči "exploratory".
[13] Od engleske reči "intercept".

Diskriminacioni skor svakog objekta predstavlja zbir vrednosti dobijenih množenjem svake nezavisne promenljive sa odgovarajućim diskriminacionim ponderom. Specifičnost diskriminacione analize je u tome da postoje jedna ili više diskriminacionih funkcija, tako da svaki objekat može imati jedan ili više diskriminacionih skorova.

Diskriminaciona analiza je statistička tehnika koja se koristi za testiranje hipoteze da su jednake aritmetičke sredine dveju ili više grupa (za određeni skup nezavisnih promenljiva). Izračunavanjem aritmetičke sredine diskriminacionih skorova svih objekata u jednoj grupi, dobija se grupna aritmetička sredina koja se zove centroid. Upoređavanjem centroida između grupa uočavaju se grupne razlike.

Slika 3.1. Jednodimenzionalne raspodele diskriminacionih skorova



Izvor: Multivariate Data Analysis, Hair, J.F., W.C. Black, B.J. Babin, R.E. Anderson i R.L., 2006, p. 275

Test statističke značajnosti diskriminacione funcije predstavlja generalizovanu meru rastojanja između centroida grupa. Dobija se upoređavanjem diskriminacionih skorova grupa. Ako je zajednička površina rasporeda diskriminacionih funkcija mala, gornji prikaz na Slici 3.1., diskriminaciona funkcija dobro razdvaja grupe. Ako je zajednička površina rasporeda diskriminacionih funkcija velika, donji prikaz na Slici

3.1., funkcija je slab diskriminator grupa. Ovaj koncept za dva rasporeda diskriminacionih skorova, prikazan je na Slici 3.1.

Obojene oblasti koje su zajedničke predstavljaju situacije pogrešno klasificiranih objekata iz grupe A i grupe B, i suprotno.

3.2.Istraživački dizajn diskriminacione analize

Diskriminaciona analiza počinje **definisanjem istraživačkih ciljeva**, koji mogu biti: ispitivanje statističke značajnosti razlike između centroida za skup promenljivih za dve ili više prethodno definisane grupe, određivanje koja nezavisna promenljiva najviše doprinosi razdvajanjem grupa, određivanje broja i struktura dimenzije diskriminacije između grupa formiranih iz skupa nezavnisnih promenljivih, kao i definisanje procedura za klasifikacije objekata (opservacija) u grupu na osnovu njihovih skorova.

Da bi se primenila diskriminaciona analiza istraživač treba da **naznači koja od promenljivih je zavisna, a koje su promenljive nezavisne**. Zavisna promenljiva treba da bude kategorijska, dok su nezavisne promenljive kvantitativne. Zavisna promenljiva može imati dve ili više kategorija (modaliteta) koje se međusobno isključuju, a zajedno obuhvataju sve moguće vrednosti. Ako dve ili više grupa imaju slične profile, diskriminaciona analiza ne može dobro da profiliše svaku grupu. Zato, istraživač treba da odredi zavisne promenljive, tako, da one jasno reflektuju razlike kod nezavisnih promenljivih.

Istraživač treba da teži manjem broju grupa. I pored toga što izgleda kao logično rešenje, povećavanje broja grupa dovodi do povećavanja kompleksnosti u profiliranju i klasifikovanju. Diskriminaciona analiza može oceniti m [14] diskriminacionih funkcija, tako da, ako je broj grupa veći, imali bi i više diskriminacionih funkcija, šta povećava kompleksnost u identifikaciji osnovnih dimenzija diskriminacije.

Posle izbora zavisne promenljive, sledi izbor nezavisnih promenljivih. Promenljive se mogu identifikovati prethodnim istraživanjem, iz teoretskog modela koji je osnova istraživanja ili intuitivno.

Kao i kod ostalih multivarijacionih tehnika, tako i kod diskriminacione analize rezultati zavise od **veličine uzoraka** koji se analizira. Diskriminaciona analiza je jako

[14] $m = (\text{broj grupa} - 1)$

osetljiva odnos između veličine uzoraka i broja nezavisnih promenlljiva. Preporučuje se upotreba minimum pet opservacija za svaku nezavisnu promenljivu.

Osim ukupnog broja opservacija, treba uzeti u obzir i **veličinu uzoraka po grupama**. Najmanja grupa mora da ima više jedinica od broja nezavisnih promenljivih. Praktičan savet je da svaka grupa (kategorija) ima više od dvadeset opservacija, ali moraju se uzeti u obzir i odnosi veličina grupa. Velike varijacije između veličina grupa imaju uticaj na ocene diskriminacionih funkcija i klasifikaciju opservacija. Kod klasifikacione analize, veće grupe imaju neproporcionalno veću šansu za klasifikaciju.

Za **validaciju** diskriminacione analize, preferira se deljenje uzoraka u dva manja pod-uzorka. Prvi pod-uzorak, **uzorak za analizu** koristi se za izračunavanje diskriminacione funkcije. Drugi pod-uzorak, **uzorak za proveru**, koristi se za testiranje diskriminacione funkcije. Bitno je da svaki pod-uzorak ima odgovarajuću veličinu da bi potvrdio zaključke. Prethodne preporuke za veličinu uzoraka su validne i kod pod-uzoraka. Uobičajeni postupak određivanja pod-uzoraka je slučajna deoba ukupnog uzorka u dva pod-uzoraka. Ovaj metod validacije funkcije poznat je kao ukrštena validacija ili validacija podeljenog uzoraka.

Najpopularniji pristup je da se ukupni uzorak podeli na dva jednaka pod-uzoraka (50%–50%). Neki istraživači koriste i odnos 60%–40% ili 75%–25% između uzorka za analizu i uzorka za proveru, u zavisnosti od veličine ukupnog uzorka.

3.3.Pretpostavke diskriminacione analize

Kao i ostale multivarijacione tehnike, tako i diskriminaciona analiza je zasnovana na nekoliko osnovnih **pretpostavki**. Osnovne pretpostavke u izvođenju diskriminacione funkcije su: **nezavisnost promenljivih**, **normalnost nezavisnih promenljivih** i **homogenost kovarijacione matrice** (kovarijacione matrice populacije za diskriminacione promenjive su jednake).

Karakteristika podataka koja ima efekat na rezultate je **multikolinernost između nezavisnih promenljivih**. Multikolinearnost označava da dve ili više nezavisih promenljivih imaju visoku korelaciju, tako da jedna promenljiva može objasniti ili predvideti drugu promenljivu i ima mali doprinos u objašnjavanju cele grupe podataka. Pretpostavka o nezavisnost promenljivih je ključna. Da bi se ispunila ova

pretpostavka potrebno je ispitati korelacionu matricu nezavisnih promenljivih. Korelacije koje imaju aposolutnu vrednost manju od 0,3 ispunjavaju ovu pretpostavku. Korelacije koje imaju apsolutnu vrednost jednaku ili veću od 0,3 ne ispunjavaju datu pretpostavku.

Neispunjavanje pretpostavke multivarijacione normalosti ima uticaj na ocenu diskriminacionih funkcija. Jedan od načina rešavanja ovog problema je transformacija podataka da bi se smanjila neusaglašenost kovarijacionih matrica. U mnogim slučajevima ovaj metod nema efekta i modeli trebaju biti detaljno provereni. Ukoliko je zavisna promenliva binarna, onda bi trebalo koristiti logističku regresiju.

Nejednake kovarijacione matrice imaju negativan efekat i na klasifikacioni proces, naročito ako su uzorci mali. Ovaj efekt se može minimizirati povecavanjem uzorka. Diskriminaciona analiza je veoma otporna na neispunjavanje pretpostavke o **homogenosti kovarijacionih matrica** ukoliko je odnos između veličine najveće grupe i veličine najmanje grupe manji od 1,5. Istraživači mogu testirati ovu pretpostavku preko Boksove M statistike. Ukoliko ova pretpostavka nije ispunjena, onda je verovatnije da će se objekti rasporediti (klasifikovati) u grupu koja ima veću disperziju.

Implicitna pretpostavka je da su odnosi promenljivih linearni. U slučaju nelinearnih odnosa sprovodi se specifična transformacija promenljivih koja eliminiše nelinearne efekte. **Nestandardne opservacije** imaju značajan efekat na tačnost klasifikacije. Zato, podatke treba ispitati i eliminisati sve nestandardne opservacije.

Novija istraživanja ukazuju da diskriminaciona analiza daje dobre rezultate, i kad pretpostavke nisu ispunjenje, osim pretpostavke za multivarijacionu normalnost i homogenost kovarijacione matrice.

3.4.Diskriminacija i klasifikacija – slučaj sa dve grupe (populacije)

Kod diskriminacione analize za dve grupe pretpostavlja se da imamo dva nezavisna uzorka (grupe) π_1 i π_2 iz dve nezavisne p - dimenzionalne populacije, objekti ili opservacije klasificiraju se na osnovu vrednosti p slučajnih promenljivih $X^{'} = \left[X_1, X_2, \ldots, X_p\right]$. Realizacije promenljive X se do neke mere razlikuju između uzoraka (grupa). Ove dve grupe mogu biti predstavljene preko funkcije gustine $f_1(x)$ i $f_2(x)$.

Pravila za alokaciju ili klasifikaciju forumulišu se na bazi uzoračkih realizacija tako što se ispituju razlike u analiziranim karakteristikama slučajno izabranih objekata koji pripadaju dvema populacijama. Skup svih mogućih ishoda je podeljen u dva dela, R_1 i R_2. Ako nova opservacija pripada delu R_1, onda je alocirana u populaciju π_1, i ako nova opservacija pripada delu R_2, onda se alocira u populaciju π_2.

Pravila klasifikacije ne mogu garantovati da je metod klasifikacije idealan, jer je moguće da nema jasnog razgraničenja između izmerenih karakteristika populacije, odnosno postoji mogućnost da se grupe preklapaju. U ovom slučaju je moguće da se objekat iz π_2 klasifikuje u π_1 ili objekat iz π_1 klasifikuje u π_2. Sledi da je osnovna ideja analize da se kreira pravilo (ili odrede regije R_1 i R_2) koje minimizira mogućnost ove nesavršenosti. Moguće je da je veća verovatnoća pojavljivanja iz jedne populacije nego iz druge jer je taj objekat član populacije koja je veća. Optimalno pravilo klasifikacije uzima u obzir ove verovatnoće.

Drugi aspekt koji optimalna klasifikacija treba da uzme u obzir su mogući troškovi pogrešne klasifikacije.

Neka su $f_1(x)$ i $f_2(x)$ funkcije gustine za vektore slučajne promenljive X dimenzije $p \times 1$ za populacije π_1 i π_2 respektivno. Objekat koji uzima skup vrednosti x mora biti dodeljen u π_1 ili u π_2. Neka je $\Omega = R_1 \cup R_2$ prostor uzoraka, odnosno skup svih mogućih opservacija x. Ako je R_1 skup vrednosti x koje klasifikuju objekte u π_1 onda $R_2 = \Omega - R_1$ obuhvata preostale x vrednosti koje klasifikuju objekte u π_2. Svaki objekat mora da se klasifikuje u samo jednu od dve populacije, skupovi R_1 i R_2 se međusobno isključuju, a zajedno obuhvataju sve vrednosti.

Uslovna verovatnoća, $P(2|1)$, važi za klasifikaciju objekta u π_2 koji potiče iz π_1, funkcija glasi:

$$P(2|1) = P(X \in R_2 \mid \pi_1) = \int_{R_2=\Omega-R_1} f_1(x)dx .$$

Slično, uslovna verovatnoća, $P(1|2)$, za klasifikaciju objekta u π_1 kada je taj objekat iz π_2 se izračunava kao:

$$P(1|2) = P(X \in R_1 \mid \pi_2) = \int_{R_1} f_2(x)dx .$$

Znak integrala označava površinu određenu funkcijom gustine $f_1(x)$ u intervalu regiona R_2. Slično, znak integrala u drugoj formuli označava površinu formiranu funkcijom gustine $f_2(x)$ u intervalu regiona R_1. Ovo je ilustrovano na Slici 3.2.

Ako je p_1 prethodna verovatnoća[15] za π_1 i p_2 je prethodna (a priori) verovatnoća za π_2, i pritom vazi $p_1 + p_2 = 1$, tada verovatnoća tačno ili netačno klasificiranih objekata se može dobiti kao proizvod prethodne (a priori) i uslovne verovatnoće:

P(Opservacija je tačno klasifikovana u π_1) =

= P(opservacija dolazi iz π_1 i tačno je klasifikovana u π_1) = $P(X \in R_1 \mid \pi_1)P(\pi_1) = P(1 \mid 1)p_1$

P(Opservacija je pogrešno klasifikovana u π_1) =

= P(opservacija dolazi iz π_2 i pogrešno je klasifikovana u π_1) = $P(X \in R_1 \mid \pi_2)P(\pi_2) = P(1 \mid 2)p_2$

P(Opservacija je tačno klasifikovana u π_2) =

= P(opservacija dolazi iz π_2 i tačno je klasifikovana u π_2) = $P(X \in R_2 \mid \pi_2)P(\pi_2) = P(2 \mid 2)p_2$

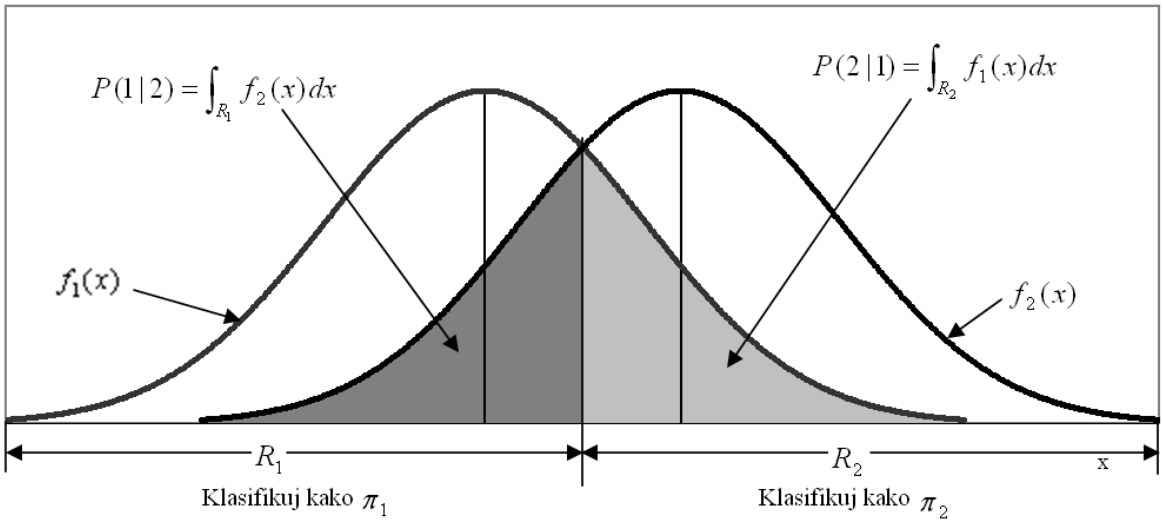
P(Opservacija je pogrešno klasifikovana u π_2) =

= P(opservacija dolazi iz π_1 i pogrešno je klasifikovana u π_2) = $P(X \in R_2 \mid \pi_1)P(\pi_1) = P(2 \mid 1)p_1$

Klasifikacione šeme su često vrednovane preko njihovih verovatnoća za pogrešnu klasifikaciju, ali ovaj postupak ignorše troškove pogrešne klasifikacije. Na primer, i mala verovatnoća, kao što je $P(2 \mid 1) = 0{,}06$ može biti značajna, ako je trošak za pogrešnu klasifikaciju ekstremno velik.

[15] A priori verovatnoća.

Slika 3.2. Verovatnoća pogrešne klasifikacije za hipotetičke klasifikacione regione



Izvor: Applied Multivariate Statistical Analysis, Johnsong, N., and D. Wichern, 2002, p. 580

Troškovi pogrešne klasifikacije mogu se definisati preko matrice troškova:

	Klasifikovati kao:	
Originalna populacija:	π_1	π_2
π_1	0	c(2|1)
π_2	c(1|2)	0

Troškovi imaju vrednost 0 za tačnu klasifikaciju, c(1|2) kada je opservacija iz π_2 netačno klasifikovana u π_1, i c(2|1) kada je opservacija iz π_1 netačno klasifikovana u π_2.

Prosečan ili očekivan trošak pogrešne klasifikacije[16] dobija se kao proizvod nedijagonalnih elemenata matrice troškova i njihove verovatnoće pojavljivanja, odnosno:

$$\text{ECM} = \text{c}(2|1)\text{P}(2|1)\text{p}_1 + \text{c}(1|2)\text{P}(1|2)\text{p}_2 .$$

Dobro pravilo klasifikacije trebalo bi da ima malu vrednost ECM.

[16] ECM - od engleske reči "expected cost of misclassification".

Regioni R_1 i R_2 koji minimiziraju ECM su definisani vrednostima x za koje važe sledeće nejednakosti:

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$\begin{pmatrix}\text{stopa}\\ \text{gustine}\end{pmatrix} \geq \begin{pmatrix}\text{stopa}\\ \text{troškova}\end{pmatrix}\begin{pmatrix}\text{stopa}\\ \text{prethodne}\\ \text{verovatnoće}\end{pmatrix}$$

$$R_2 : \frac{f_1(x)}{f_2(x)} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$\begin{pmatrix}\text{stopa}\\ \text{gustine}\end{pmatrix} < \begin{pmatrix}\text{stopa}\\ \text{troškova}\end{pmatrix}\begin{pmatrix}\text{stopa}\\ \text{prethodne}\\ \text{verovatnoće}\end{pmatrix}$$

Posebni slučajevi regiona sa minimalnim očekivanim troškovima su sledeći:

1) $p_2 / p_1 = 1$ (jednake a priori verovatnoće)

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \left(\frac{c(1|2)}{c(2|1)}\right) \qquad R_2 : \frac{f_1(x)}{f_2(x)} < \left(\frac{c(1|2)}{c(2|1)}\right)$$

2) $c(1|2)/c(2|1) = 1$ (jednakvi troškovi pogrešne klasifikacije)

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \left(\frac{p_2}{p_1}\right) \qquad R_2 : \frac{f_1(x)}{f_2(x)} < \left(\frac{p_2}{p_1}\right)$$

3) $p_2 / p_1 = c(1|2)/c(2|1) = 1$ ili $p_2 / p_1 = 1/c(1|2)/c(2|1)$ (jednake a priori verovatnoće i jednaki troškovi pogrešne klasifikacije)

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq 1 \qquad \text{i} \qquad R_2 : \frac{f_1(x)}{f_2(x)} < 1.$$

Kada a priori verovatnoće nisu poznate, često se pretpostavlja da su one jednake, i da pravilo za minimalne ECM koristi samo količnik (stopu) funkcija gustine i količnik troškova pogrešne klasifikacije. Ako stopa troškova pogrešne klasifikacije nije poznata, onda se pretpostavlja da njena vrednost iznosi 1, i stopa funkcije gustine se upoređuje sa stopom a priori verovatnoće.

3.4.1.Klasifikacija dve normalno distribuirane populacije sa jednake kovarijacione matrice ($\Sigma_1 = \Sigma_2 = \Sigma$)

Klasifikacione procedure na osnovu normalnih populacija dominiraju u statističkoj praksi jer su jednostavne i imaju prilično visoku efikasnost. Ako

pretpostavimo da su $f_1(x)$ i $f_2(x)$ multivarijacione normalne gustine raspodele, gde prva ima vektor sredina μ_1 i kovarijacionu matricu Σ_1, a druga ima vektor sredina μ_2 i kovarijacionu matricu Σ_2, onda se u posebnom slučaju jednakostih kovarijacionih matrica ($\Sigma_1 = \Sigma_2 = \Sigma$) dobijaju prilično jednostavne linearne klasifikacione statistike.

Ako se pretpostavi da su zajedničke gustine $X' = [X_1, X_2, \ldots, X_p]$ za populacije π_1 i π_2 pretstavljene preko

$$f_i(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i)\right] \text{ za } i = 1,2.$$

Pretpostavka je i da su populacioni parametri μ_1, μ_2 i Σ poznati. Onda, kada poništimo izraze $(2\pi)^{p/2}|\Sigma|^{1/2}$, minimalni ECM za regione su odrađeni kao:

$$R_1 : \exp\left[-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)\right] \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$R_2 : \exp\left[-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)\right] < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

Kada su regioni R_1 i R_2 poznati, može se konstruirati sledeće klasifikaciono pravilo. Ako populacije π_1 i π_2 imaju višedimenzionalne normalne gustine, pravilo alokacije koje minimizira ECM je sledeće:

- alociraj x_0 u π_1 ako

$$(\mu_1 - \mu_2)'\Sigma^{-1}x_0 - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

- alociraj x_0 u π_2 u suprotnom slučaju.

U praksi, vrednosti za populacije μ_1, μ_2 i Σ nisu poznate, tako da ova pravila treba modificirati. Predloženo je da se izvrši zamena populacionih parametara sa vrednostima uzoraka. Ako se pretpostavi da postoji n_1 opservacija za višedimenzionalnu slučajnu promenljivu $X' = [X_1, X_2, \ldots, X_p]$ iz π_1 i n_2 opservacija iz π_2 gde $n_1 + n_2 - 2 \geq p$, onda su odgovarajuće matrice podataka:

$$\underset{(n_1 \times p)}{X_1} = \begin{bmatrix} x_{11}' \\ x_{12}' \\ \vdots \\ x_{1n_1}' \end{bmatrix} \quad \underset{(n_2 \times p)}{X_2} = \begin{bmatrix} x_{21}' \\ x_{22}' \\ \vdots \\ x_{2n_2}' \end{bmatrix}$$

Iz ovih matrica podataka, vektori sredina uzoraka i kovarijacione matrice su definisani sledećim izrazima

$$\underset{(p\times 1)}{\bar{x}_1} = \frac{1}{n_1}\sum_{j=1}^{n_1} x_{1j}\,, \quad \underset{(p\times p)}{S_1} = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(x_{1j}-\bar{x}_1)(x_{1j}-\bar{x}_1)'$$

$$\underset{(p\times 1)}{\bar{x}_2} = \frac{1}{n_2}\sum_{j=1}^{n_2} x_{2j}\,, \quad \underset{(p\times p)}{S_2} = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(x_{2j}-\bar{x}_2)(x_{2j}-\bar{x}_2)'$$

Kako se pretpostavlja da populacije imaju istu kovarijacionu matricu Σ, kovarijacione matrice uzoraka S_1 i S_2 se kombiniraju da bi se dobila jedna, nepristrasna ocena Σ. Ponderisani prosek

$$S_{zdr} = \left[\frac{n_1-1}{(n_1-1)+(n_2-2)}\right]S_1 + \left[\frac{n_2-1}{(n_1-1)+(n_2-2)}\right]S_2$$

je nepristrasna ocena Σ ako matrice podataka X_1 i X_2 sadrže slučajne uzorke iz odgovarajućih populacija π_1 i π_2.

Ukoliko zamenimo $\bar{x}_1$ sa μ_1 i μ_2 sa $\bar{x}_2$, i Σ sa S_{komb} dobijamo klasifikaciono pravilo za uzorak (ili ocenjeni minimalni trošak pogrešne klasifikacije (ECM) za dve populacije sa normalnim rasporedima):

Alociraj x_0 u π_1 ako

$$(\bar{x}_1-\bar{x}_2)'S_{zdr}^{-1}x_0 - \frac{1}{2}(\bar{x}_1-\bar{x}_2)'S_{zdr}^{-1}(\bar{x}_1+\bar{x}_2) \geq \ln\left[\left(\frac{c(1\,|\,2)}{c(2\,|\,1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

i alociraj x_0 u π_2 u suprotnom slučaju.

Ako u prethodnoj jednačini imamo

$$\left(\frac{c(1\,|\,2)}{c(2\,|\,1)}\right)\left(\frac{p_2}{p_1}\right) = 1$$

onda $\ln(1) = 0$, i pravilo ocenjenog ECM-a za dve normalno distribuirane populacije svodi se na upoređenje skalarne promenjive

$$\hat{y} = (\bar{x}_1-\bar{x}_2)'S_{zdr}^{-1}x = \hat{a}'x$$

ocenjene za x_0, sa brojem

$$\hat{m} = \frac{1}{2}(\bar{x}_1-\bar{x}_2)'S_{zdr}^{-1}(\bar{x}_1+\bar{x}_2) = \frac{1}{2}(\bar{y}_1-\bar{y}_2)$$

gde

$$\bar{y}_1 = (\bar{x}_1-\bar{x}_2)'S_{zdr}^{-1}\bar{x}_1 = \hat{a}'\bar{x}_1$$

i

$$\bar{y}_2 = (\bar{x}_1 - \bar{x}_2)' S_{zdr}^{-1} \bar{x}_2 = \hat{a}' \bar{x}_2$$

Tako, ocenjeno pravilo za minimum ECM-a za dve normalno distribuirane populacije je jednako kreiranju dve univarijacione populacije za y vrednosti dobijene preko formiranje linearne kombinacije opservacija iz populacije π_1 i π_2 i onda dodeljivanjem nove opservacije x_0 u π_1 ili π_2, zavisno od toga da li $\hat{y}_0 = \hat{a}' x_0$ uzima poziciju desno ili levo od središne tačke $\hat{m}$ za dve univarijacione sredine $\bar{y}_1$ i $\bar{y}_2$.

3.4.2.Fišerov pristup za klasifikaciju dvie populacije

Fišer je dobio izraz linearne statističke klasifikacije koristeći sasvim drugi postupak. Fišerova ideja je bila da transformiše multivarijacione opservacije x u univarijacione opservacije y, tako da su y vrednosti izvedene iz populacije π_1 ili π_2 maksimalno razdvojene. Fišer je predložio da se preko linearne kombinacije vrednosti za x kreriraju vrednosti y jer one nisu ništa drugo već njihove funkcije, a lakše su za upotrebu. Fišerov pristup ne bazira se na pretpostavci da su populacije normalne, ali ipak pretpostavlja da su populacione kovarijacione matrice jednake, jer koristi ocenu zajedničke kovarijacione matrice.

Linearnom kombinacijom vrednosti za x dobijaju se vrednosti $y_{11}, y_{12}, \ldots, y_{1n_1}$ za opservacije iz prve populacije i vrednosti $y_{21}, y_{22}, \ldots, y_{2n_2}$ za opservacije iz druge populacije. Razdvajanje ova dva skupa univarijacionih vrednosti y se vrši izračunavanje razlika između $\bar{y}_1$ i $\bar{y}_2$ izraženih u jedinicama standardne devijacije, odnosno

$$\text{razdvajanje} = \frac{|\bar{y}_1 - \bar{y}_2|}{s_y}, \quad \text{gde } s_y^2 = \frac{\sum_{j=1}^{n_1} (y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2} (y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2}$$

je zajednička ocena varijanse. Cilj je da se selektiraju linearne kombinacije za vrednosti x da bi se dobilo maksimalno razdvajanje sredina uzoraka $\bar{y}_1$ i $\bar{y}_2$.

Linearna kombinacija $\hat{y} = \hat{a}' x = (\bar{x}_1 - \bar{x}_2)' S_{zdr}^{-1} x$ (Fišerova linearna diskriminaciona funkcija) maksimizira odnos

$$\frac{\begin{pmatrix}\text{Kvadrat rastojanja između}\\ \text{sredina uzoraka za y}\end{pmatrix}}{(\text{Varijansa promenljive y})} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{\left(\hat{a}'\bar{x}_1 - \hat{a}'\bar{x}_1\right)}{\hat{a}' S_{zdr} \hat{a}} = \frac{\left(\hat{a}' d\right)^2}{\hat{a}' S_{zdr} \hat{a}}$$

za sve vektore koeficienta $\hat{a}$ gde $d = (\bar{x}_1 - \bar{x}_2)$. Maksimalna vrednost koeficijenta je $D^2 = (\bar{x}_1 - \bar{x}_2)' S_{zdr}^{-1} (\bar{x}_1 - \bar{x}_2)$.

Pravilo alokacije na osnovu Fišerove diskriminacione funkciju glasi

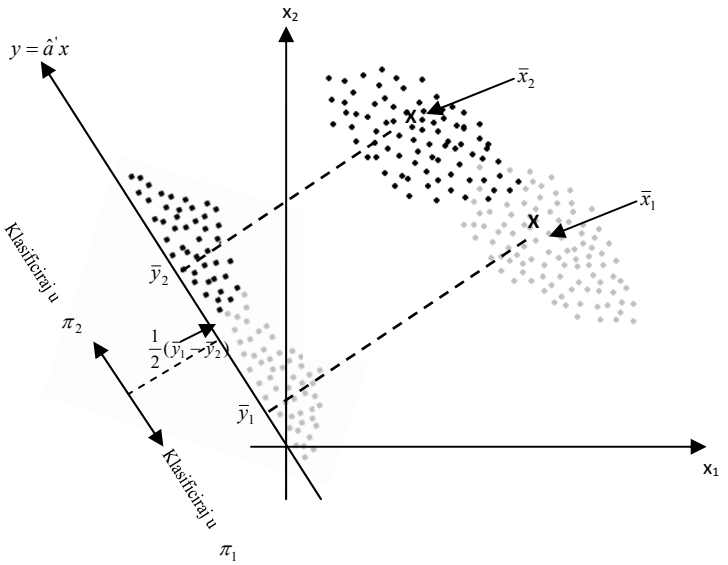
Alociraj x_0 u π_1 ako

$$\hat{y}_0 = (\bar{x}_1 - \bar{x}_2)' S_{zdr}^{-1} x_0 \geq \hat{m} = \frac{1}{2} (\bar{x}_1 - \bar{x}_2)' S_{zdr}^{-1} (\bar{x}_1 + \bar{x}_2)$$

ili

$$\hat{y}_0 - \hat{m} \geq 0$$

Slika 3.3. Prezentacija Fišerove procedure diskriminacije za dve populacije sa dve promenljive



Izvor: Applied Multivariate Statistical Analysis, Johnsong, N., and D. Wichern, 2002, p. 592

Alociraj x_1 u π_2 ako

$$\hat{y}_0 < \hat{m}$$

ili

$$\hat{y}_0 - \hat{m} < 0$$

Sve tačke dijagrama rasturanja (Slika 3.3.) su projektovane na liniji sa pravcom $\hat{a}$. Ova procedura se ponavlja dok se uzorci ne razdvoje maksimalno.

3.4.3.Evaluacija funkcije klasifikacije

Jedan važan način za ocenivanje performanse bilo koje klasifikacione procedure bitno je da se izračunaju njene "stope greške", ili verovatnoće pogrešne klasifikacije. Kada su rasporedi populacija poznati, verovatnoće pogrešne klasifikacije se izračunavaju relativno jednostavno. Nasuprot tome rasporedi populacije su vrlo retko poznati, pa su prikazane stope grešaka povezane sa klasifikacionim funkcijama uzoraka. Kada je jednom dobijena klasifikaciona funkcija, mera njene performanse se koristi za sledeće uzorake.

Ukupna verovatnoća pogrešne klasifikacije[17] je

$$TPM = p_1 \int_{R_2} f_1(x)dx + p_2 \int_{R_1} f_2(x)dx$$

Najmanja vrednost ovog izraza, dobijena preko izbora R_1 ili R_2, zove se optimalna stopa greške[18].

$$\text{OER} = p_1 \int_{R_2} f_1(x)dx + p_2 \int_{R_1} f_2(x)dx$$

gde su R_1 ili R_2 su determinisane preko

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \left(\frac{p_2}{p_1} \right) \qquad R_2 : \frac{f_1(x)}{f_2(x)} < \left(\frac{p_2}{p_1} \right)$$

Tako, OER je stopa greške za minimalne vrednosti pravila ukupne verovatnoće pogrešne klasifikacije (TPM).

Performans klasifikacione funkcije uzoraka može se izračunati preko realne stope grešaka[19]

$$AER = p_1 \int_{\hat{R}_2} f_1(x)dx + p_2 \int_{\hat{R}_1} f_2(x)dx$$

gde $\hat{R}_1$ i $\hat{R}_2$ prestavljaju regione klasifikacije određene uzorcima veličine n_1 i n_2. Na primer, ukoliko se koristi klasifikaciona funkcija

$$(\bar{x}_1 - \bar{x}_2)' S_{zdr}^{-1} x_0 - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S_{zdr}^{-1} (\bar{x}_1 + \bar{x}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)} \right) \left(\frac{p_2}{p_1} \right) \right]$$

regioni $\hat{R}_1$ i $\hat{R}_2$ su definisani za skup vrednosti x koje zadovoljuju sledeće nejednačine.

[17] TMP – od engleske reči "total probability of misclassification".
[18] OER – od engleske reči "optimum error rate".
[19] AER – od engleske reči "actual error rate".

$$\hat{R}_1 : (\bar{x}_1 - \bar{x}_2)' S_{zdr}^{-1} x - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S_{zdr}^{-1}(\bar{x}_1 + \bar{x}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

$$\hat{R}_2 : (\bar{x}_1 - \bar{x}_2)' S_{zdr}^{-1} x - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S_{zdr}^{-1}(\bar{x}_1 + \bar{x}_2) < \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

Realna stopa greške pokazuje performanse klasifikacione funkcije pri izvljačenju drugih uzoraka. Kao i optimalnu stopu greške, i realnu stopu greške je teško izračunati jer zavisi od nepoznatih funkcija gustine $f_1(x)$ i $f_2(x)$. Ipak, moguće je izračunati ocenu realne greške.

Postoji mera performanse koja ne zavisi od rasporeda populacija i koja se može izračunati za bilo koju klasifikacionu proceduru. Ova se mera zove očigledna stopa greške[20] i definisana je kao deo opservacija koje se nalaze u analiziranim uzorcima, a koje su pogrešno klasifikovane preko diskriminacione funkcije uzorka.

Očigledna stopa greške može se lako izračunati preko matrice konfuzije, koja pokazuje faktičku pripadnost naspurot predviđene pripadnosti po grupama. Za n_1 opservacija iz π_1 i n_2 opservacija iz π_2, matrica konfuzije ima sledeći oblik

	Predviđena pripadnost grupi		
Faktička pripadnost grupi	π_1	π_2	
π_1	n_{1C}	$n_{1M} = n_1 - n_{1C}$	n_1
π_2	$n_{2M} = n_2 - n_{2C}$	n_{2C}	n_2

gde je:

n_{1C} = broj tačno klasifikovanih objekata iz π_1 u π_1

n_{1M} = broj pogrešno klasifikovanih objekata iz π_1 u π_2

n_{2C} = broj tačno klasifikovanih objekata iz π_2 u π_2

n_{2M} = broj pogrešno klasifikovanih objekata iz π_2 u π_1

Očigledna stopa greške je

$$APER = \frac{n_{1M} + n_{2M}}{n_1 + n_2}$$

ili proporcija opservacija iz uzoraka za analizu koji su pogrešno klasificirani.

[20] APER – od engleske reči "apparent error rate".

3.5.Diskriminacija i klasifikacija u slučaju više grupa (populacija)

Kao i u prethodnom poglavlju, razmatraju se teoretska optimalna pravila i onda se naznačuju potrebne modifikacije za praktičnu primenu.

Minimalni očekivani trošak metoda pogrešne klasifikacije objašnjava se na sledeći način: neka su $f_i(x)$ gustine populacija π_i, $i=1,2,\ldots g$. (U najčešćem slučaju uzimamo $f_i(x)$ da bude multivarijaciona normalna gustina, ali ovo nije neophodno za razvoj generalne teorije). Neka je

p_i = a priori verovatnoća pripadnosti populacije π_i, $i=1,2,\ldots g$

$c(k\,|\,i)$ = trošak alokacije opservacije u π_k kada opservacija realno pripada u π_i, za $k,i=1,2,\ldots g$

Za $k=i$, $c(i\,|\,i)=0$. I neka je R_k skup svih x klasifikovanih u π_k i

$$P(k\,|\,i) = \text{P (klasifikuj opservaciju kao } \pi_k \,|\, \pi_i) = \int_{R_k} f_i(x)dx$$

za $k,i=1,2,\ldots g$ sa $P(i\,|\,i)=1-\sum_{\substack{k=1\\k\neq i}}^{g} P(k\,|\,i)$.

Očekivan uslovni trošak zbog pogrešne klasifikacije opservacije x iz π_1 u π_2, ili $\pi_3,\ldots,$ ili π_g je

$$ECM(1) = P(2\,|\,1)c(2\,|\,1) + P(3\,|\,1)c(3\,|\,1) + \ldots + P(g\,|\,1)c(g\,|\,1) = \sum_{k=2}^{g} P(k\,|\,1)c(k\,|\,1)$$

Ovaj očekivani uslovni trošak se javlja sa a priori verovatnoćom p_1 (verovatnoća za π_1).

Na sličan način, dobijaju se očekivani uslovni troškovi pogrešnih klasifikacija $ECM(2),\ldots,ECM(g)$. Ukoliko se multiplicira svaki očekivani trošak pogrešne klasifikacije sa njegovom a priori verovatnoćom i ovi proizvodi se saberu, onda dobija se zajednički očekivan trošak pogrešne klasifikacije:

$$ECM = p_1 ECM(1) + p_2 ECM(2) + \ldots + p_g ECM(g) =$$

$$= p_1 \left(\sum_{k=2}^{g} P(k \mid 1) c(k \mid 1) \right) + p_2 \left(\sum_{\substack{k=1 \\ k \neq 2}}^{g} P(k \mid 2) c(k \mid 2) \right) + \ldots + p_g \left(\sum_{k=1}^{g-1} P(k \mid g) c(k \mid g) \right) =$$

$$= \sum_{i=1}^{g} p_i \left(\sum_{\substack{k=1 \\ k \neq i}}^{g} P(k \mid i) c(k \mid 1) \right)$$

Određivanje optimalne klasifikacione procedure svodi se na izbor iz intervala $R_1, R_2, \ldots, R_g$ (koji se međusobno isključuju, a zajedno obuhvataju sve moguće vrednosti) tako da je vrednost prethodne formule minimalna. Klasifikacioni intervali koji minimiziraju ECM su definisani preko alokacije opservacije x u onu populaciju π_k, $k = 1,2,\ldots g$, za koju je najmanja vrednost izraza:

$$\sum_{\substack{i=1 \\ i \neq k}}^{g} p_i f_i(x) c(k \mid i) .$$

U situaciji kada je ova vrednost jednaka za dve različite populacije, onda se x može smestiti u bilo koju od njih.

Ako su troškovi pogrešnih klasifikacija jednaki, minimalni očekivani troškovi pravila pogrešne klasifikacije predstavljaju minimum od ukupne verovatnoće pravila pogrešne klasifikacije. Moguće je da se troškovi pogrešne klasifikacije postave tako da svi budu jednaki 1. U ovom slučaju, x bi se alocirao u onu populaciju π_k, $k = 1,2,\ldots g$, za koju je najmanja vrednost izraza:

$$\sum_{\substack{i=1 \\ i \neq k}}^{g} p_i f_i(x) .$$

To znači da je ova vrednost najmanja kada je izostavljeni izraz, $p_k f_k(x)$ najveći. Sledi, da kada su troškovi pogrešne klasifikacije jednaki, pravilo minimalnih očekivanih troškova pogrešne klasifikacije ima sledeću jednostavnu formu:

- alociraj x_0 u π_k ako je:

$$p_k f_k(x) > p_i f_i(x) \quad \text{za sve } i \neq k$$

ili, ekvivalentno,

- alociraj x_0 u π_k ako je:

$$\ln p_k f_k(x) > \ln p_i f_i(x) \quad \text{za sve } i \neq k$$

U opštom slučaju, pravilo minimalnih očekivanih troškova pogrešne klasifikacije ima tri komponente: a priori verovatnoće, troškove pogrešne klasifikacije i funkciju gustine. Ove komponente treba odrediti ili oceniti pre primene pravila.

3.5.1.Klasifikacija više normalno distribuiranih populacija

Kao poseban slučaj navešćemo:

$$f_i(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right] \quad , \quad i = 1,2\ldots,g$$

gde je multivarijaciona normalna gustina se srednjim vektorom μ_i i kovarijacionom matricom Σ_i. Ako je $c(i|i)=0, c(k|i)=1, k \neq i$ (ili ekvivalentno, troškovi pogrešne klasifikacije su jednaki), onda:

- alociraj x_0 u π_k ako je:

$$\ln p_k f_k(x) > \ln p_i f_i(x) \quad \text{za sve } i \neq k$$

- alociraj x_0 u π_k ako je:

$$\ln p_k f_k(x) = \ln p_k - \left(\frac{p}{2}\right)\ln(2\pi) - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k) = \max_i \ln p_i f_i(x)$$

Konstanta $(p/2)\ln(2\pi)$ može se ignorisati u prethodnoj formuli, jer je ista za sve populacije. Odavde se definiše kvadratni diskriminacioni skor za i-tu populaciju:

$$d_i^Q(x) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i) + \ln p_i, \quad i = 1,2\ldots,g$$

Kvadratni skor $d_i^Q(x)$ sastoji se od pondera[21] opšte varijanse $|\Sigma_i|$, a priori verovatnoće p_i, i kvadrata rastojanja opservacije x od sredina populacije μ_i.

Kada se koriste diskriminacioni skorovi, pravilo klasifikacije:

$$\ln p_k f_k(x) = \ln p_k - \left(\frac{p}{2}\right)\ln(2\pi) - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k) = \max_i \ln p_i f_i(x)$$

postaje pravilo minimuma ukupne verovatnoće pogrešne klasifikacije za normalne populacije, kada kovarijacione matrice Σ_i nisu jednake[22], i glasi:

- alociraj x_0 u π_k ako je kvadratni skor $d_i^Q(x) = $ max. vrednosti $d_1^Q(x), d_2^Q(x), \ldots, d_g^Q(x)$, gde je:

$$d_i^Q(x) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i) + \ln p_i, \quad i = 1,2\ldots,g$$

[21] Od engleske reči "contributions".
[22] TPM – od engleske reči "total probability of misclassification".

U praktičnim istraživanjima, parametri populacija μ_i i Σ_i nisu poznati, pa se ocenjuju na bazi uzoraka koji sadrže tačno klasifikovane opservacije. Tako, za uzorak iz populacije π_i imamo

$\bar{x}_i$ = vektor srednje vrednosti uzoraka

S_i = kovarijaciona matrica uzoraka

n_i = veličina uzoraka

Ocena kvadratnog diskriminacionog skora $\hat{d}_i^Q(x)$ je:

$$\hat{d}_i^Q(x) = -\frac{1}{2}\ln|S_i| - \frac{1}{2}(x-\bar{x}_i)' S_i^{-1}(x-\bar{x}_i) + \ln p_i, \qquad i = 1,2\ldots,g$$

Iz ove jednačine sledi i pravilo klasifikacije za uzorak ili pravilo minimalne ocenjene ukupne verovatnoće pogrešne klasifikacije za nekoliko normalnih populacija čije kovarijacione matrice Σ_i nisu jednake:

- alociraj x_0 u π_k ako je kvadratni skor $\hat{d}_i^Q(x)$ najveća vrednost od

$\hat{d}_1^Q(x), \hat{d}_2^Q(x), \ldots, \hat{d}_g^Q(x)$, gde

$$\hat{d}_i^Q(x) = -\frac{1}{2}\ln|S_i| - \frac{1}{2}(x-\bar{x}_i)' S_i^{-1}(x-\bar{x}_i) + \ln p_i, \qquad i = 1,2\ldots,g$$

Jednostavnija forma je moguća ako su jednake kovarijacione matrice Σ_i. Kada je $\Sigma_i = \Sigma$, za $i = 1,2\ldots,g$, diskriminacioni skor dobija oblik:

$$d_i^Q(x) = -\frac{1}{2}\ln|\Sigma| - \frac{1}{2}x'\Sigma^{-1}x + \mu_i\Sigma^{-1}x - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i + \ln p_i$$

Prva dva elementa iz izraza su isti za $d_1^Q(x), d_2^Q(x), \ldots, d_g^Q(x)$, i mogu se ignorisati. Ostali elementi sadrže konstantu $c_i = \ln p_i - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i$ i linearne kombinacije koje sadrže komponente iz x.

Linearni diskriminacionog skor se definiše na sledeći način:

$$d_i(x) = \mu_i\Sigma^{-1}x - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i + \ln p_i, \quad i = 1,2\ldots,g$$

Ocena $\hat{d}_i(x)$ linearnog diskriminacionog skora $d_i(x)$ je dobijena na osnovu združene[23] ocene Σ:

$$S_{zdr} = \frac{1}{n_1 + n_2 + \ldots + n_g - g}\left((n_1-1)S_1 + (n_2-1)S_2 + \ldots(n_g-1)S_g\right).$$

[23] Od engleske reči "pooled".

Ocena je:

$$\hat{d}_i(x) = \bar{x}_i' S_{zdr}^{-1} x - \frac{1}{2} \bar{x}_i' S_{zdr}^{-1} \bar{x}_i + \ln p_i, \quad i = 1,2\ldots, g$$

Odavde, sledi i pravilo za minimalno ocenjene ukupne verovatnoće pogrešne klasifikacije za normalne populacije sa jednakim kovarijacionim matricama:

- alociraj x_0 u π_k ako je linearni diskriminacioni skor $\hat{d}_k(x) =$ najvećoj vrednosti od $\hat{d}_1(x), \hat{d}_2(x), \ldots, \hat{d}_g(x)$, gde

$$\hat{d}_i(x) = \bar{x}_i' S_{zdr}^{-1} x - \frac{1}{2} \bar{x}_i' S_{zdr}^{-1} \bar{x}_i + \ln p_i, \quad i = 1,2\ldots, g$$

Izraz $d_i(x) = \mu_i \Sigma^{-1} x - \frac{1}{2} \mu_i' \Sigma^{-1} \mu_i + \ln p_i, \quad i = 1,2\ldots, g$ je dobra linearna funkcija za vrednosti x. Ekvivalentna forma za klasifikovanje u slučaju dve iste kovarijacione matrice može se dobiti iz formule kvadratnog diskriminacionog skora, izostavljanjem elemenata $-\frac{1}{2}\ln|\Sigma|$. Rezultat, na osnovu uzoračkih vrednosti za za nepoznate parametre populacije, može se interpretirati preko kvadratnih rastojanja:

$$D_i^2(x) = (x - \bar{x}_i)' S_{zdr}^{-1} (x - \bar{x}_i)$$

između x i vektora sredina uzoraka $\bar{x}_i$. Pravilo alokacije je sledeće:

- alociraj x u populaciji π_i koja ima najveću vrednost za $-\frac{1}{2} D_i^2(x) + \ln p_i$.

Može se zaključiti da ovo pravilo kao i pravilo pre njega dodeljuje x "najbližoj" populaciji. Ako a priori verovatnoće nisu poznate, uobičajeno je da se pretpostavi da su verovatnoće jednake, $p_1 = p_2 = \ldots = p_g = 1/g$. Opservacija se onda dodeljuje najbližoj populaciji.

Kvadratna pravila mogu biti alternativa za klasifikaciju linearne diskriminacione funkcije, samo kada je pretpostavka za normalnost ispunjena, dok pretpostavka za jednakvost kovarijacionih matrica ne mora biti ispunjena. Kvadratna pravila su osetljivija od linearnih na odstupanje od normalnog rasporeda. Ako postoji sumnja koje se pravilo treba koristiti, moguće je konstruirati dva pravila i ispitati njihove stope grešaka preko Lachenbruch-ove procedure.

3.5.2.Fišerov pristup u klasifikaciji više populacija

Fišer je predložio proširivanje svog diskriminacionog metoda za dve populacije na više populacija. Motivacija Fišerove diskriminacione analize je potreba da se dobije prikaz populacije koja sadrži samo nekoliko linernih kombinacija opservacija, kao $a_1^{'}, a_2^{'}$ i $a_3^{'}$. Njegov pristup ima nekoliko prednosti u razdvajanju više populacija, i to:

1) Prikaz g populacija preko relativno malog broja linearnih kombinacija koj ima smanjena dimenzija velikog broja manifestnih promenljivih.

2) Kreiranje dijagrama prve dve ili tri linearne kombinacije (diskriminante). Ovo omogućuje da se prikažu odnosi i moguće grupe populacija.

3) Dijagram rasturanja vrednosti uzoraka za prve dve diskriminante može prikazati nestandardne opservacije ili druge nepravilnosti u podacima.

Osnovni cilj Fišerove diskriminacione analize je razdvajanje populacije i njihova klasifikaciju. Nije neophodna da je ispunujena pretpostavka da je g populacija sa višedimenzionalnom normalnom raspodelom. Ipak, pretpostavlja se da su kovarijacione matrice populacija, dimenzija $p \times p$, jednake, odnosno $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_g = \Sigma$. Ukoliko nisu jednake, onda $P = \left[e_1, \ldots e_g\right]$ su karakteristični vektori[24] za Σ koji odgovaraju karakterističnim vrednostima[25] $\left[\lambda_1, \ldots, \lambda_g\right]$. Tada se X zamenjuje sa $P^{'}X$ koje ima kovarijacionu matricu $P^{'}\Sigma P$.

Neka $\overline{\mu}$ označava vektor srednjih vrednosti kombinacija populacija i B_μ sumu proizvoda između grupa, tako da je:

$$B_\mu = \sum_{i=1}^{g} (\mu_i - \overline{\mu})(\mu_i - \overline{\mu})^{'} \quad \text{gde je } \mu = \frac{1}{g}\sum_{i=1}^{g} \mu_i$$

Posmatra se linearna kombinacija:

$$Y = a^{'}X$$

koja ima očekivanu vrednost:

$$E(Y) = a^{'}E(X \mid \pi_i) = a^{'}\mu_i \qquad \text{za populaciju } \pi_i$$

i varijansu:

$$Var(Y) = a^{'}Cov(X)a = a^{'}\Sigma a \qquad \text{za sve populacije.}$$

Kao rezultat prethodno izloženog, očekivana vrednost $\mu_{iY} = a^{'}\mu_i$ menja se sa promenom populacije iz koje je X. Prvo se definiše zajednička sredina:

[24] Od engleske reči "eigen vectors".
[25] Od engleske reči "eigen values".

$$\bar{\mu}_Y = \frac{1}{g}\sum_{i=1}^{g}\mu_{iY} = \frac{1}{g}\sum_{i=1}^{g}a'\mu_i = a'\left(\frac{1}{g}\sum_{i=1}^{g}\mu_i\right)$$

i formira količnik:

$$\frac{\begin{pmatrix}\text{suma kvadrata rastojanja sredina}\\ \text{populacija od zajednicke sredine Y}\end{pmatrix}}{(\text{varijansa za Y})} = \frac{\sum_{i=1}^{g}(\mu_{iY}-\bar{\mu}_Y)^2}{\sigma_Y^2} = \frac{\sum_{i=1}^{g}(a'\mu_i - a'\bar{\mu})^2}{a'\Sigma a} = \frac{a'\left(\sum_{i=1}^{g}(\mu_i-\bar{\mu})(\mu_i-\bar{\mu})'\right)a}{a'\Sigma a}$$

ili

$$\frac{\sum_{i=1}^{g}(\mu_{iY}-\bar{\mu}_Y)^2}{\sigma_Y^2} = \frac{a'B_\mu a}{a'\Sigma a}.$$

Ova količnik meri **varijabilitet između grupa** u odnosu na zajednički **varijabilitet unutar grupa**. Zatim se bira vrednost a tako da se maksimizira vrednost ovog količnika.

Uobičajeno je da Σ i μ_i nisu poznate. Slučajan uzorak sa veličine n_i iz populacije π_i, $i = 1,2,\ldots,g$. Baza podataka je matrica X veličine $n_i \times p$, i njen j-ti red označava se sa x_{ij}', i ista potiče iz populacije π_i. Prvo se formira vektor sredina uzoraka:

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}x_{ij}$$

i kovarijacione matrice S_i, $i = 1,2,\ldots,g$, pa se onda formira zajednički vektor proseka:

$$\bar{x} = \frac{1}{g}\sum_{i=1}^{g}\bar{x}_i$$

koji predstavlja prosečan vektor dobijen iz individualnih proseka uzoraka.

Analogno za B_μ, definiše se B - matrica suma uzajamnih proizvoda odstupanja sredina grupa od opšte sredine:

$$B = \sum_{i=1}^{g}(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'.$$

Isto, ocena Σ se bazira na matrici variranja unutar grupa:

$$W = \sum_{i=1}^{g}(n_i - 1)S_i = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)'.$$

Sledi da $W/(n_1 + n_2 + \ldots + n_g - g) = S_{zdr}$ je ocena za Σ.

Da bi se objasnila Fišerova diskriminaciona funkcija uzoraka, neka $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_s > 0$ označava $s \le \min(g-1, p)$, karakteristične vrednosti različite od nule iz $W^{-1}B$ i $\hat{e}_1, \ldots, \hat{e}_s$ koji su odgovarajući karakteristični vektori (odrađeni tako da je $\hat{e}' S_{komb} \hat{e} = 1$). Onda je vektor koeficienta $\hat{a}$ koji maksimizira količnik:

$$\frac{\hat{a}' B \hat{a}}{\hat{a}' W \hat{a}} = \frac{\hat{a}' \left(\sum_{i=1}^{g} (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' \right) \hat{a}}{\hat{a}' \left[\sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)' \right] \hat{a}}$$

predstavljen preko vrednosti $\hat{a}_1 = \hat{e}_1$. Linearna kombinacija $\hat{a}_1' x$ zove se prva diskriminanta uzoraka. Izbor $\hat{a}_2 = \hat{e}_2$ daje drugu diskriminantu uzoraka, $\hat{a}_2' x$, i tako dalje, dobija se $\hat{a}_k x = \hat{e}_k'$, k -a diskriminanti uzoraka, $k \le s$.

Diskriminante neće imati kovarijansu nuls za svaki slučajan uzorak X_i. Umesto toga, uslov:

$$\hat{a}_i' S_{zdr} \hat{a}_k = \begin{cases} 1 & \text{ako } i = k \le s \\ 0 & \text{u suprotnom} \end{cases}$$

će biti zadovoljen. Upotreba S_{zdr} je prikladna jer se pretpostavlja da su jednake kovarijacione matrice g populacija.

Fišerove diskriminante mogu se upotrebiti i u klasifikaciji opservacija. One se određuju da bi se podatci prikazali sa manje promenljivih koje maksimalno razdvajaju populacije. I pored toga što je osnovna funkcija diskriminante separacija, diskriminante mogu obezbediti osnovu za pravilo klasifikacije. Najpre se, objašnjava povezanost preko diskriminante populacije $a_i' x$.

Ako definišimo k -tu diskriminantu:

$$Y_k = a_k' X, \quad k \le s$$

može se zaključiti da:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_s \end{bmatrix} \text{ ima vektor srednjih vrednosti } \mu_{iY} = \begin{bmatrix} \mu_{iY_1} \\ \vdots \\ \mu_{iY_s} \end{bmatrix} = \begin{bmatrix} a_1' \mu_i \\ \vdots \\ a_s' \mu_i \end{bmatrix}$$

za populaciju π_i i kovarijacionu matricu I, za sve populacije, jer komponente iz Y imaju varijanse 1 i kovarijanse 0. Odgovarajuća mera kvadratnog rastojanja između $Y = y$ i μ_{iY} je:

$$(y - \mu_{iY})'(y - \mu_{iY}) = \sum_{j=1}^{s} (y_j - \mu_{iY_i})^2$$

Dobro pravilo klasifikacije je ono koje dodeljuje y u populaciju π_k, ako je kvadratno rastojanje među y i μ_{kY} manje od kvadratnog rastojanja među y i μ_{iY}, za $i \neq k$.

Ako se koristi samo r diskriminanti za alokaciju, pravilo glasi:

- lociraj x u π_k ako:

$$\sum_{j=1}^{r} (y_j - \mu_{kY_j})^2 = \sum_{j=1}^{r} \left[a_j'(x - \mu_k)\right]^2 \leq \sum_{j=1}^{r} \left[a_j'(x - \mu_i)\right]^2 \quad \text{za sve } i \neq k$$

Bitan odnos pravila klasifikacije i "teorije normalnosti" diskriminacionih skorova je dat izrazom:

$$d_i(x) = \mu_i'\Sigma^{-1}x - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i + \ln p_i$$

ili ekvivalentno,

$$d_i(x) - \frac{1}{2}x'\Sigma^{-1}x = -\frac{1}{2}(x - \mu_i)'\Sigma^{-1}(x - \mu_i) + \ln p_i$$

dobijeno kada se doda ista konstanta $-\frac{1}{2}x'\Sigma^{-1}x$ na svako $d_i(x)$.

Neka je $y_i = a_j'x$, gde su $a_j = \Sigma^{-1/2}e_j$ i e_j karakteristični vektor $\Sigma^{-1/2}B_\mu\Sigma^{-1/2}$. Sledi da je:

$$\sum_{j=1}^{p} (y_j - \mu_{iY_j})^2 = \sum_{j=1}^{p} \left[a_j'(x - \mu_i)\right]^2 = (x - \mu_i)'\Sigma^{-1}(x - \mu_i) = -2d_i(x) + x'\Sigma^{-1}x + 2\ln p_i$$

Ako je $\lambda_1 \geq \ldots \geq \lambda_s > 0 = \lambda_{s+1} = \ldots = \lambda_p$, tada je $\sum_{j=s+1}^{p} (y_j - \mu_{iY_j})^2$ konstanta za sve populacije $i = 1,2,\ldots,g$, tako da samo prvih s diskriminanti y_j, ili $\sum_{j=1}^{s} (y_j - \mu_{iY_j})^2$ doprinosi u klasifikaciji, odnosno samo prvih s diskriminanti koriste se za klasifikaciju.

Na kraju pravilo klasifikacije na osnovu prve diskriminante, odnosno Fišerova procedura klasifikacije zasnovana na diskriminantama uzoraka glasi:

-alociraj x u π_k ako:

$$\sum_{j=1}^{r}(\hat{y}_j - \hat{y}_{kj})^2 = \sum_{j=1}^{r}\left[\hat{a}_j'(x-\bar{x}_k)\right]^2 \le \sum_{j=1}^{r}\left[\hat{a}_j'(x-\bar{x}_i)\right]^2 \qquad \text{za sve } i \neq k$$

gde je:

$$\hat{a}_j = \frac{\hat{a}'B\hat{a}}{\hat{a}'W\hat{a}} = \frac{\hat{a}'\left(\sum_{i=1}^{g}(\bar{x}_i-\bar{x})(\bar{x}_i-\bar{x})'\right)\hat{a}}{\hat{a}'\left[\sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(x_{ij}-\bar{x}_i)'\right]\hat{a}},\ \bar{y}_{kj} = \hat{a}_j'\bar{x}_k \text{ i } r \le s.$$

Do sada nismo rekli zašto su prve diskriminante značajnije od ostalih. Njihova važnost postaje očigledna ako se razmatraju njihovi ponderi. Razmatra se mera separacije:

$$\Delta_S^2 = \sum_{i=1}^{g}(\mu_i - \bar{\mu})'\Sigma^{-1}(\mu_i - \bar{\mu})$$

gde je:

$$\bar{\mu} = \frac{1}{g}\sum_{i=1}^{g}\mu_i$$

i gde je $(\mu_i - \bar{\mu})'\Sigma^{-1}(\mu_i - \bar{\mu})$ kvadratno statističko rastojanje između sredine i-te populacije i centroida $\bar{\mu}$.

Separacija pretstavljena sa Δ_S^2 može se reprodukovati pomoću diskriminacione sredine. Prva diskriminanta, $Y_1 = e_1'\Sigma^{-1/2}X$ ima sredinu $\mu_{iY_i} = e_1'\Sigma^{-1/2}\mu_i$ i kvadratno rastojanje $\sum_{i=1}^{g}(\mu_{iY_i} - \bar{\mu}_{Y_i})^2$ između μ_{iY_i} i centralne vrednosti $\bar{\mu}_{y_1} = e_1'\Sigma^{-1/2}\bar{\mu}$ je λ_1. Jer Δ_S^2 može se napisati i kao:

$$\Delta_S^2 = \lambda_1 + \lambda_2 + \ldots + \lambda_p = \sum_{i=1}^{g}(\mu_{iY} - \bar{\mu}_Y)'(\mu_{iY} - \bar{\mu}_Y) =$$

$$= \sum_{i=1}^{g}(\mu_{iY_1} - \bar{\mu}_{Y_1})^2 + \sum_{i=1}^{g}(\mu_{iY_2} - \bar{\mu}_{Y_2})^2 + \ldots + \sum_{i=1}^{g}(\mu_{iY_p} - \bar{\mu}_{Y_p})^2 .$$

Sledi da prva diskriminanta vezana za λ_1, daje najveći individualan doprinos, u meri separacije Δ_S^2. U opštem slučaju, r-ta diskriminanta, $Y_r = e_r'\Sigma^{-1/2}X$ doprinosi preko sabirka λ_r meru separacije Δ_S^2. Ako su sledećih $s-r$ karakterističnih vrednosti (kao što je prethodno naznačeno $\lambda_{s+1} = \lambda_{s+2} = \ldots \lambda_p = 0$) takve da je zbog $\lambda_{r+1} + \lambda_{r+2} + \ldots + \lambda_r$ mali u poređenju sa zbirom $\lambda_1 + \lambda_2 + \ldots + \lambda_r$, onda se poslednje

diskriminante $Y_{r+1}, Y_{r+2}, \ldots, Y_s$ mogu izostativiti, jer nemaju značajni uticaj na jačinu separacije.

3.6.Interpretacija rezultata

Kada su diskriminacione funkcije statistički značajne istraživač treba da se fokusira na formulaciju i interpretaciju dobijenih rezultata. Ovaj proces uključuje proveru diskriminacione funkcije da bi se utvrdio relativni značaj svake nezavisne promenljive u diskriminaciji između grupa. Postoje dva metoda za određivanje relativnog značaja, i to **standizovani diskriminacioni koeficienti** i **diskriminaciona opterećenja**[26] **(strukturne korelacije)**.

Tradicionalni pristup u interpretaciji diskriminacione funkcije ispituje **znak i veličinu standardizovanog diskriminacionog koeficienta** za svaku nezavisnu promenljivu koja je deo diskriminacione funkcije. Ako se izostavi znak, onda svaki koeficient pretstavlja relativni doprinos odgovarajuće promenljive u datoj funkciji. Nezavisne promenljive sa visokim vrednostima koeficijenata doprinose više za diskriminacionu moć funkcije nego promenljive sa manjim vrednostia koeficijenata. Znak ukazuje na to da li jedoprinos promenljive pozitivan ili negativan.

Diskriminaciona opterećenja su poznate i kao **strukturne korelacije**. Merenjem proste linearne korelacije između nezavisne promenljive i diskriminacione funkcije, diskriminaciona opterećenja pokazuju deo varijanse koju nezavisne promenljive dele sa diskriminacionom funkcijom. Ona takođe mere i relativni doprinos svake nezavisne promenljive diskriminacionoj funkciji.

Posebna karakteristika opterećenja je da se oni mogu izračunati za sve promenljive, nezavisno da li su ili nisu deo diskriminacione funkcije. Osnovno pitanje diskriminacionih opterećenja je koja njihova vrednost ukazuje da su promenljive biti značajni diskriminatori? Smatra se da su značajne one promenljive koje imaju opterećenja $\pm 0{,}40$ ili više. Diskriminaciona opterećenja možemo smatrati validnijom merom od diskriminacionih koeficienta zbog njihove korelacione prirode.

Ukoliko postoje dve ili više značajnih diskriminacionih funkcija, javlja se problemi njihove interpretacije: Kako da se predstavi efekat svake promenljive kroz

[26] Od engleske reči "loading".

sve funkcije. Rešenje problema se nalazi u konceptu **rotacije funkcije**, **indeksu potentnosti**, **teritorijalne mape** i **grafikonu vektora diskriminacionih opterećenja**.

Dobijene diskriminacione funkcije se mogu rotirati kako bi se varijanse redistribuirale. **Rotacijom** se zadržava originalna struktura i osnovanost diskriminacionog rešenja i omogućava lakša interpretacija funkcije. Najčešće se koristi VARIMAX kriterijum kao osnova za rotaciju.

Indeks potentnosti je relativna mera diskriminacione moći svake promenljive. Uključuje doprinos promenljive u diskriminacionoj funkciji (diskriminaciono opterećenje) i relativni doprinos funkcije u generalnom rešenju (relativna mera između funkcija na osnovu karakteristične vrednosti). Opšti indeks predstavlja zbir individualnih indeksa potentnosti za sve značajne diskriminacione funkcije. Interpretacija je ograničena jer se indeks koristi samo za rangiranje promenljivih, dok kao apsolutna mera nema nikakav značaj. Opšti indeks potentnosti se dobija kroz od dva koraka:

Prvi korak: Izračunava se vrednost potentnosti svake promenljive za sve signifikante funkcije. Diskriminaciona moć promenljive predstavljena je preko nerotiranih diskriminacijonih opterećenja. Prvo se izračunava relativna karakteristična mera za svaku signifikantu diskriminacionu funkciju kao:

$$\text{Relativna karakteristicna vrednost diskriminacione funkcije } j = \frac{\text{Karakteristicna vrednost diskriminacione funkcije } j}{\text{Zbir karakteristicnih vrednosti za sve znacajne funkcije}}$$

Vrednost potenosti za svaku promenljivu diskriminacione funkcije je:

$$\text{Vrednost potentnosti za promenljivu } i = (\text{Diskriminaciono opterecenje}_{ij})^2 \times \text{Relativna karakteristicna vrednost funkcije } j$$

Drugi korak: Izračunava se opšti indeks potentnosti za sve signifikantne funkcije. Opšti indeks potentnosti za svaku promenljivu se izračunava kao:

Opsti indeks potentosti = Zbir vrednoti potentnosti za promenljivu i za sve signifikantne diskriminacione funkcije za promenljivu

Teritorijalna mapa je grafički metod gde se svaka opservacija prestavlja grafički preko se Z vrednosti diskriminacionih funkcija. Za svaku opservaciju standardizovani diskriminacioni skor Z za prvu diskriminacionu funkciju postavlja se na X osu, a skorovi drugih diskriminacionih funkcija se predstavljaju na Y osi. Ovim načinom se lakše uočavaju razlike svake grupe kao i situacije kada se grupe preklapaju. Kada se grafički prikazuje centroid svake grupe dobija se sredina koja se koristi da bi se ocenila razdaljina svake observacije od grupnog centroida. Ovo je korisna procedura kada se ocenjuju velike vrednosti Mahalanobisove D^2 mere odstojanja koje vode do pogrešne klasifikacije. Na grafikonu se mogu predstaviti i linije koje predstavljaju kritične granice[27] diskriminacionih skorova. Svaki član grupe koji je izvan ovih granica je pogrešno klasifikovan. Ovo omogućuje da se uvidi diskriminaciona funkcija koja je najviše odgovorna za pogrešne klasifikacije.

Grafikon vektora diskriminacionih opterećenja prikazuje rotirana ili nerotirana opterećenja. Preferira se korištenje rotiranih opterećenja. Precizniji pristup prikazuje na grafikonu opterećenja i vektore za svako opterećenje i centroid grupe. Vektor je prava linija koja se postavlja od početka (centra) grafikona do koordinate diskriminacionih opterećenja ili grupnih centroida posebno za svaku promenljivu. Izduženi vektor je takva prezentacija kada dužina svakog vektora je indikator za relativni značaj svake promenljive u diskriminaciji između grupa. Procedura se sastoji od tri koraka, prvi kada se sve promenljive, bez obzira da li su značajne ili ne, prikazuju na grafikonu kao vektori. Drugi korak je izduživanje vektora. Diskriminaciono opterećenje svake promenljive povećava se množenjem diskriminacionog opterećenja sa svojom univarijacionom vrednosću F statistike. Vektori ukazuju koje imaju najveću sredinu za određenu nezavisnu promenljivu i pri tom se izdvajaju od grupa koje imaju najmanji srednji skor. Treći korak je prikazivanje centroida na grafikonu, tada se i grupni centroidi množe sa približni F vrednostima koje se odnose na svaku diskriminacionu funkciju. Ako izdužavamo opterećenja, centroidi se isto mogu izdužiti i precizno prikazati na istim grafikonu.

[27] Od engleske reči “cutting score”.

Na kraju postupka, istraživač ima prikaz grupisanja promenljivih za svaku diskriminacionu funkciju, veličinu značajnosti svake promenljive (prikazanu dužinom vektora) i profil centroida sa svaku grupu (prikazan preko blizine vektora).

3.7.Validacija rezultata

U poslednjoj fazi diskriminacione analize vrši se validaciju diskriminacionih rezultata kako bi istraživači bili sigurni da rezultati imaju kako eksternu, tako i internu validnost. Jer diskriminaciona analiza daje proprciju tačno klasifikovanih opservacija samo kod analiziranih uzoraka, pa je vrednovanje rezultata neophodni korak. Postoje dve tehnike: a) **validacione procedure**, koje sadrže kreiranje **validacionog uzorka** i **ukrštenu validaciju** i b) **profiliranje grupnih razlika**.

Prva tehnika je **validaciona procedura**. Validacija je kritičan korak u svakoj diskriminacionoj analizi jer često, a posebno kada je uzorak mali, nije moguće generalizovati rezultate (**eksterna validnost**). Najpoznatija procedura dobijanja eksterne validnosti je ocena učešća tačno klasifikovanih opservacija. Validacija se može dobiti ili preko izdvojenog uzorka[28] ili preko korišćenja procedure koje ponovo procesiraju uzorak za analizu. Eksterna validnost postoji kada je učešće tačno klasifikovanih opservacija veće od standarda određenog za granicu koji tačnog predviđanja.

Najčešča validacija učešća tačno klasifikovanih opservacija se vrši preko posebno izabranog uzorka koji se još zove i uzorak za proveru (validaciju). Cilj korišćenja validacionog uzorka je ocenivanje koliko diskriminaciona funkcija diskriminira uzorak opservacija koji nije korišten u dobijanju diskriminacione funkcije. Ovaj proces uključuje utvrđivanje diskriminacionih funkcija za opservacije analiziranog uzorka i dalje njihovo apliciranje na uzorke za proveru (validaciju). Opravdanje za deobu uzoraka na deo za analizu i deo za proveru (validaciju) je utvrđivanje nepristrasnih ocena i tačnost predviđavanja diskriminacione funkcije, ukoliko se opservacije koje se koriste u određivanju klasifikacione matrice koriste i u izračunavanju funkcija jer klasifikaciona tačnost bi bila veća od validne tačnosti ukoliko bi se koristio na uzorak za ocenivanje.

[28] Od engleske reči “holdout sample".

Istraživači ističu da bi pouzdanost validacije diskriminacione funkcije bila veća ukoliko bi se ova procedura ponovila nekoliko puta. Nasuprot slučajne deobe uzoraka na deo za analizu i deo za proveru (validaciju), istraživač treba da ponovi deljenje nekoliko puta, i svaki put da testira validnost diskriminacione funkcije preko utvrđene klasifikacione matrice i učešća tačno klasifikovanih opservacija. U tom slučaju treba odrediti prosečno učešće tačno klasifikovanih opservacija kako bi se dobila jedna vrednost.

Ukrštena validacija je pristup za ostvarivanje eksterne validnosti koji koristi više poduzoraka kreiranih od ukupnog uzorka. Najčešće korišćen pristup je metoda noža na sklapanje[29]. Ukrštena validacija se bazira na principu "izostavi jedan". Metod ocenjuje $k-1$ poduzorak, eliminišući po jednu opservaciju iz uzorka koji se sastoji od k opservacija. Izračunava se diskriminaciona funkcija za svaki poduzorak. Pripadnost opservacije grupi iz koje je eliminisana ocenjuje se preko diskriminacione funkcije ocenjene na osnovu iz ostalih opservacija. Nakon što se dobiju predviđivanja za pripadnost grupi, jedne po jedne, konstruiše se klasifikaciona matrica i izračunava se učešće tačno klasifikovanih opservacija. Ukrštena validacija je jako osetljiva na veličinu uzorka. Predlaže se da se ona koristi kada je najmanja grupa bar tri puta veća od broja nezavisnih promenljivih, dok najveći broj istraživača predlaže da bude pet puta veća. Ali, ukrštena validacija je jedina moguča validacija čak i kada je uzorak mali, da bi se razdvoijo na deo za proveru (validaciju) i deo za analizu. Njena upotreba je sve veća od početka korištenja kompjuterskih programa.

Profiliranje grupnih razlika je druga validaciona tehnika koja profilira grupe preko nezavisnih promenljivih da bi se osigurala njihova povezanost sa konceptualne osnove u formulaciju originalnog modela. Nakon što istraživač identifikuje nezavisne promenljive koje najviše doprinose diskriminaciji grupe, sledi profiliranje karakteristika grupa na osnovu grupnih sredina ili centroida. Ovaj profil omogućava da istraživač shvati karakter svake grupe u zavisnosti od nezavisnih promenljiva.

[29] Od engleske reči "jackknife".

3.8.Primena diskriminacione analize u rešavanju empiriskih problema

3.8.1.Primena diskriminacione analize u slučaju dve grupe za klasifikaciju zemalja članica Evropske Unije i zemalja koje nisu članice Evropske Unije u odnosu na njihove karakteristike

Cilj ove analize je sprovesti diskriminaciju zemalja članica Europske Unije i zemalja koje nisu članice Europske Unije na osnovu sledećih promenljivih: *Strane direktne investicije*, *Bruto domaći proizvod*, *Rast bruto domaćeg proizvoda*, *Inflacija*, *Korisnici mobilne telefonije, Populacija i Površina zemlje*. Osim poslednje dve promenjlive, ostale promenljive ukazuju na ekonomski rast i razvoj jedne zemlje, što predstavlja bitan preduslov za ulazak u Evropsku Uniju. Diskriminaciona analiza bi trebalo i da ukaže koja od ovih promenljiva najviše doprinosi u razdvajanju grupa, i kao takva je najbitna promenljiva za članstvo u Evropskoj Uniji.

Nakon sprovedene diskriminacije, na osnovu dobijenih diskriminacionih koeficienta, vrši se klasifikacija zemalja. Ova klasifikacija treba da ukaže koje su zemlje pravilno klasificirane, a koje zemlje nisu, kao i da utvrdi koje su zemlje sa najvećim potencijalom za članstvo u Europskoj Uniji.

Analizirano je 48 zemalja na osnovu sedam nezavisnih kvanitativnih promenljivih i jedne zavisne kategorijske promenljive, *Članstvo u Europskoj Uniji*. Atributivna promenljiva ima dve kategorije, pa su formirane dve grupe, 1 - zemlje koje su članice i 0 - zemlje koje nisu članice Europske Unije, koje jasno reflektuju razlike nezavisnih promenljivih.

Podaci su dobijeni iz baze podataka oficijalne statistike Svetske Banke za 2007 godinu. Diskriminacionom analizom ocenuje se jedna diskriminaciona funkcija (broj grupa – 1= 2 – 1=1).

Kako je **veličina uzoraka** veoma značajna, a posebno **veličina uzoraka po grupama**, svaka analizirana grupa ima više od dvadeset zemalja, odnosno prva grupa broji 22 zemlje, a druga 26. Grupe su približno iste veličine, pa ne postoji razlog za neproporcionalno veće šanse u klasifikaciji. Zbog nedostataka podataka, dve zemlje su isključene iz analize.

Sledeći korak je ispitivanje ispunjenosti uslova za primenu diskriminacione analize. Prvo se podaci skeniraju kako bi se utvrdilo prisustvo **nestandardnih zemalja**. Za ovo smo koristili Mahalanobisovo D^2 rastojanje. Najmanja vrednost Mahalanobisovog rastojanja 0,65 utvrđena je za Makedoniju, a najveća 43,89 za

Rusiju. Ovo ukazuje da bi Rusiju trebalo isključiti iz analize kao nestandardnu zemlju. Pre odstranivanja Rusije iz analize, sproveden je Kolmogorov - Smirnov test normalnosti za nezavisne promenljive. S obzirom da SPSS ne omogućava ispitivanje multivarijacione normalnosti, ispitana je normalnost posebno za svaku promenljivu. Rezultati **Kolmogorov – Smirnov-og testa** ukazuje da nezavisne promenljive ne slede normalan raspored.

Zbog postojanja nestandardne opservacije i promenljivih koje ne slede normalan raspored, predlaže se **logaritamska transformacija za sve promenljive**, osim za promenljivu *Rasta bruto domaćeg proizvoda* za koju je najbolja transformacija **kvadratnog korena**. Nakon transformacija, Mahalanobisovo rastojanje kreće se od 0,82 za Hrvatsku do 19,14 za Luksemburg, što znači da ne postoje značajne nestandardne opservacije. Testovi normalnosti ukazuju da osim promenljivih, *Inflacija* i *Broj korisnika mobilne telefonije*, sve ostale promenljive slede normalni raspored, nakon transformacije.

Na ovaj način je poboljšana baza podataka za diskriminacionu analizu.

Zbog lakše preglednosti za nezavisne promenljive su uveden skraćenice: *SDI – Strane direktne investije u dolarima*, *neto priliv*, *BDP - Bruto domaći proizvod po tekovnim cenam, u dolarima*, *BDP% - Godišni rast bruto domaćeg proizvoda u procentima*, *INF – Inflacija ili deflator bruto domaćeg proizvoda*, *godišni u procentima*, *MOB – Broj korisnika mobilne telefonije (na 100 žitelja)*, *POP – Ukupno stanovništvo*, *POV – Površina zemlje u kvadratnim kilometrima*.

Rezultati dobijeni upotrebom softvera SPSS prikazani su u Tabeli 3.1.

Podaci iz Tabele 3.1. ukazuju na potencijalni problem. Za sve promenljive, veće vrednosti aritmetičke sredine su povezane sa većim vrednostima standardne devijacije. Korelacija ova dva pokazatelja je isto jaka i iznosi 0,881, i ostaje ista i posle tranformacije promenljivih.

Da bi se ocenio doprinos svake promenljive u modelu, korišćeni su testovi jednakosti sredina grupa, diskriminacionih koeficijenata i strukturna matrica, a njihovi rezultati prikazani su tabelarno.

Tabela 3.1. Osnovni statistički pokazatelji za analizirane karakteristike po kategorijama zemalja

Grupa zemalja	Promenljiva	Aritmetička sredina	Standardna devijacija	Validne opservacije	
				Neponderirane	Ponderirane
0	SDI	9,34	0,74	20	20
	GDP	10,55	0,78	20	20
	GDP%	2,67	0,57	20	20
	INF	0,91	0,37	20	20
	MOB	76,65	33,80	20	20
	POP	6,86	0,59	20	20
	POV	5,24	0,73	20	20
1	SDI	10,19	0,67	26	26
	GDP	11,37	0,67	26	26
	GDP%	2,05	0,61	26	26
	INF	0,52	0,31	26	26
	MOB	117,27	16,27	26	26
	POP	6,95	0,59	26	26
	POV	4,96	0,56	26	26
Ukupno	SDI	9,82	0,81	46	46
	GDP	11,01	0,82	46	46
	GDP%	2,32	0,66	46	46
	INF	0,69	0,39	46	46
	MOB	99,61	32,31	46	46
	POP	6,91	0,58	46	46
	POV	5,08	0,65	46	46

Izvor: Rezultati dobijeni primenom SPSS – a

Rezultati testa jednakosti sredina grupa (Tabele 3.2.) ukazuju na diskriminacioni potencijal svake nezavisne promenljive pre dobijanja konačnog modela. Ovi rezultati su dobijeni primenom jednofaktorske analize varijanse za nezavisnu promenlju pri čemu je faktor zavisno promenljiva *Članstvo u EU*. Ako je nivo signifikantosti iznad 0,10, smatra se da promenljiva nije značajna za model.

Tabela 3.2. Ocenivanje doprinosa svake nezavisne promenljive na bazi testa jednakvosti sredina grupa

Promenliva	Vilksova Lamda	F	Stepeni slobode 1	Stepeni slobode 2	Signifikantnost
SDI	0,73	16,67	1	44	0,00
GDP	0,75	14,73	1	44	0,00
GDP%	0,78	12,75	1	44	0,00
INF	0,75	14,57	1	44	0,00
MOB	0,60	28,97	1	44	0,00
POP	0,99	0,27	1	44	0,60
POV	0,96	2,04	1	44	0,16

Izvor: Rezultati dobijeni primenom SPSS – a

U modelu, promenljive *Ukupno stanovništvo* i *Površina zemlje* nisu značajne.

Vilksova Lambda je još jedna mera za potencijal promenljive. Male vrednosti ovog koeficienta ukazuju da je ta promenljiva dobar diskriminator grupa.

Rezultati iz tabele sugerišu da je promenljiva *Broj korisnike mobilne telefonije* najbolji diskriminator, sledi promenljiva *Strane direktne investicije*, pa promenljiva *Bruto domaći proizvod*, promenljiva *Inflacija* i na kraju je promenljiva *Rast bruto domaćeg proizvoda*.

Tabela 3.3. Združene matrice za nezavisne promenljive[30] (a)

Matrica	Promenljiva	SDI	GDP	GDP%	INF	MOB	POP	POV
Kovarijaciona	SDI	0,49	0,40	-0,19	-0,10	7,36	0,21	0,16
	GDP	0,40	0,51	-0,20	-0,10	6,21	0,32	0,29
	GDP%	-0,19	-0,20	0,35	0,09	-4,46	-0,06	-0,01
	INF	-0,10	-0,10	0,09	0,12	-2,30	0,01	0,06
	MOB	7,36	6,21	-4,46	-2,30	643,72	-0,35	-0,31
	POP	0,21	0,32	-0,06	0,01	-0,35	0,35	0,29
	POV	0,16	0,29	-0,01	0,06	-0,31	0,29	0,41
Korelaciona	SDI	1,00	0,81	-0,45	-0,40	0,41	0,50	0,35
	GDP	0,81	1,00	-0,47	-0,41	0,34	0,76	0,63
	GDP%	-0,45	-0,47	1,00	0,43	-0,30	-0,18	-0,01
	INF	-0,40	-0,41	0,43	1,00	-0,27	0,05	0,27
	MOB	0,41	0,34	-0,30	-0,27	1,00	-0,02	-0,02
	POP	0,50	0,76	-0,18	0,05	-0,02	1,00	0,78
	POV	0,35	0,63	-0,01	0,27	-0,02	0,78	1,00

(a) Kovarijaciona matrica ima 44 stepena slobode.

Izvor: Rezultati dobijeni primenom SPSS – a

Tabela 3.3. prikazuje kovarijacionu i korelacionu matricu dobijenu združivanjem odgovarajućih elementa grupa. Elementi kombinirane matrice se dobijaju kao prosek iz različitih matrica za sve grupe kovarijanse i varijanse. U kovarijacionoj matrici varijanse su na glavnoj dijagonali, a ostale pozicije matrice prikazuju varijanse. Ako postoji jaka korelacija [(0,75;1) ili (-0,75;-1)], između nekoliko promenljivih preporučuje novi izbor promenljivih koji bi dao bolje rezultate.

Korelaciona matrica između nezavisnih promenljivih ukazuje da najjača korelacija postoji između *Bruto domaćeg proizvoda* i *Stranih direktnih investicija* i iznosi 0,81, pa bi se moglo zaključiti da je ova korelacija dovoljno jaka i signifikantna da bi se uzela u obzir. Da bi bili sugurni, uporedićemo razlike između strukturne matrice i diskriminacionih koeficienta.

[30] Od engleske reči "pooled within groups matrices".

Jedna od neophodnih pretpostavki diskriminacione analize je **jednakost kovarijacionih matrica** između grupa ili kategorija zavisnih promenljiva.

U Tabeli 3.4. su prikazane vrednosti varijanse i kovarijanse u grupama. Varijanse se nalaze na glavnoj dijagonali, a na ostalim pozicijama su kovarijanse. Ako uporedimo kovarijanse između promenljivih za prvu i drugu grupu, kovarijanse za promenljivu *Korisnici mobilne telefonije* se najviše razlikuju među grupama. Ostale nisu značajne.

Tabela 3.4. Kovarijacione matrice za grupe zemalja (a)

Grupe zemalja	Promenljiva	SDI	GDP	GDP%	INF	MOB	POP	POV
0	SDI	0,55	0,52	-0,12	-0,08	17,34	0,22	0,27
	GDP	0,52	0,60	-0,12	-0,08	16,96	0,29	0,35
	GDP%	-0,12	-0,12	0,32	0,09	-9,32	0,06	0,09
	INF	-0,08	-0,08	0,09	0,14	-5,37	0,08	0,12
	MOB	17,34	16,96	-9,32	-5,37	1142,24	2,04	2,09
	POP	0,22	0,29	0,06	0,08	2,04	0,35	0,32
	POV	0,27	0,35	0,09	0,12	2,09	0,32	0,54
1	SDI	0,45	0,32	-0,23	-0,10	-0,22	0,20	0,07
	GDP	0,32	0,45	-0,25	-0,11	-1,96	0,35	0,25
	GDP%	-0,23	-0,25	0,37	0,08	-0,77	-0,16	-0,08
	INF	-0,10	-0,11	0,08	0,10	0,03	-0,04	0,01
	MOB	-0,22	-1,96	-0,77	0,03	264,85	-2,16	-2,14
	POP	0,20	0,35	-0,16	-0,04	-2,16	0,35	0,27
	POV	0,07	0,25	-0,08	0,01	-2,14	0,27	0,31
Ukupno	SDI	0,66	0,57	-0,32	-0,18	15,89	0,22	0,10
	GDP	0,57	0,67	-0,32	-0,18	14,42	0,33	0,23
	GDP%	-0,32	-0,32	0,44	0,15	-10,77	-0,07	0,04
	INF	-0,18	-0,18	0,15	0,15	-6,21	0,00	0,08
	MOB	15,89	14,42	-10,77	-6,21	1043,89	0,59	-3,07
	POP	0,22	0,33	-0,07	0,00	0,59	0,34	0,28
	POV	0,10	0,23	0,04	0,08	-3,07	0,28	0,42

(a) Kovarijaciona matrica (Ukupno) ima 45 stepena slobode.

Izvor: Rezultati dobijeni primenom SPSS – a

Boks M test (Tabela 3.5.) testira nultu hipotezu i ujedno pretpostavku za jednakvost kovarijacionih matrica populacije. Signifikantost ove statistike se bazira na F transformaciji. Na osnovu rezultata Boksove M statistike u našem primeru (p - vrednost = 0,04>0,01), prihvata se nulta hipoteza da su kovarijacione matrice jednakve nivou značajnosti 0,01. Uobičajeno, ovaj test je signifikantan, ili prihvata se alternativna hipoteza da su kovarijacione matrice različite, kada grupe imaju veliki broj opservacija ili kada pretpostavka o multivarijacionoj normalnosti nije ispunjena.

Tabela 3.5. Rezultat provere homogenosti kovarijacione matrice Boks M testom

Boks M test		52,06
F	Aproksimativno	1,53
	Stepen slobode 1	28
	Stepen slobode 2	5824,99
	Signifikantnost	0,04

Izvor: Rezultati dobijeni primenom SPSS – a

Rang i prirodni logaritam prikazanih determinanti su dobijeni iz kovarijacione matrice grupa (Tabela 3.6.). U modelu sa više grupa, determinanta prirodnog logaritma ukazuje kovarijaciona matrica koje grupe se najviše razlikuje. Determinanta prirodnog logaritma je proizvod karakterističe vrednosti svoje kovarijacione matrice unutar grupa. Implicitna pretpostavka je da su svi odnosi promenljivih linearni. Rang prikazuje najveći broj linearno nezavisnih redova ili kolona u primeru.

Tabela 3.6. Determinante prirodnog logaritma

Grupa zemalja	Rang	Determinanta prirodnog logaritma
0	7	-5,31
1	7	-7,05
Kombinirane unutar grupa	7	-5,12

Izvor: Rezultati dobijeni primenom SPSS – a

Determinante prirodnog logaritma su mere varijabilnosti grupe. Varijabilnije grupe imaju veće vrednosti determinante prirodnog logaritma. Velike vrednosti determinanta između grupa ukazuju da grupe imaju različite kovarijacione matrice, ali to nije slučaj u ovom primeru.

Konačni zaključak u vezi pretpostavki diskriminacione analize je da su promenljive večinom međusobno nezavisne, da postoji multivarijaciona normalnost nezavisnih promenljivih i da su kovarijacione matrice homogene na nivou značajnosti 0,01. Zbog ovoga se može zaključiti da bi rezultati dobijeni iz diskriminacione analize bili adekvatni.

Prikladnosti modela ocenjena je preko kanoničke diskriminacione funkcije (Tabela 3.7.). Kao dopunjenje mera koje ukazuju na individualni doprinos nezavisnih promenljivih u diskriminacionom modelu, procedura diskriminacione analize daje i

podatke o karakterističnim vrednostima i Vilksovoj Lambdi (Tabelu 3.8.) da bi se utvrdilo koliko diskriminacioni model u celini odgovara bazi podataka.

Tabela 3.7. Karakteristične vrednosti

Funkcija	Sopstvena vrednost	% od varijanse	Kumulativni %	Kanonička korelacija
1	1,16	100,00	100,00	0,73

Izvor: Rezultati dobijeni primenom SPSS – a

Karakteristični koreni daju informaciju o relativnoj efikasnosti svake diskriminacione funkcije. Kada postoje samo dve grupe kao u ovom slučaju, kanonična korelacija je najkorisna mera iz tabele i ekvivalentna je Pirsonovom koeficientu korelacije između diskriminacionih skorova i predviđene vrednosti zavisne promenljive (pripadnost grupe 0 ili 1 nakon sprovedene diskriminacione analize).

Tabela 3.8. Vilksova Lambda

Test funkcije	Vilksova lambda	χ^2	Stepeni slobode	Signifikantnost
1	0,46	31,15	7	0,00

Izvor: Rezultati dobijeni primenom SPSS – a

Vilksova Lamda je mera koja pokazuje koliko precizno svaka funkcija razdvaja opservacije u grupe. Male vrednosti Vilksove Lamde ukazuju na veću diskriminacionu sposobnost funkcije. Ova se statistika kreće između 0 i 1. Vrednost bliža 0 ukazuje da su sredine grupa različite, dok vrednost bliža 1 ukazuje da sredine grupa nisu različite.

χ^2 transformacije Vilksove lamde se koriste u kombinaciji sa stepenima slobode da bi se utvrdila signifikantost. χ^2 statistika, testira hipotezu da su sredine funkcije jednake između grupa. Niski novoi signifikantosti (ispod od 0,10) ukazuju da se sredine grupa razlikuju, dok vrednosti veće od 0,10 ukazuju da se sredine ne razlikuju. Niski novoi signifikantosti ukazuju da diskriminaciona funkcija dobro separiše grupe, kao što je primer u ovom slučaju.

Tabela 3.9. Koeficijenti standardizovane kanoničke diskriminacione funkcije

Koeficijenti	Funkcija 1
SDI	-0,23
GDP	1,67
GDP%	0,04
INF	0,44
MOB	0,38
POP	-0,23
POV	-1,11

Izvor: Rezultati dobijeni primenom SPSS – a

Kada se promenljive ocenjuju različitim mernim skalama, veličina nestandardizovanih koeficienta omogućava malu indikaciju za relativni doprinos promenljive u celoj diskriminaciji. Standardizacija koeficienata omogućuje poredjenje promenljivih merenih različitim mernim skalama. Koeficienti sa visokim apsolutnim vrednostima korespondiraju promenljivim sa većom sposobnošću značajne diskriminacije. U ovom slučaju to je *Bruto domaći proizvod*, zatim sledi *Površina zemlje*, koje nije značajna, zatim *Inflacija*, *Broj korisnka mobilne telefonije*, *Strane direktne investicije*, *Ukupan broj stanovnika* i na kraju promenljiva *Rast bruto domaćeg proizvoda* (Tabela 3.9.).

Tabela 3.10. Koeficijenti matrice strukture

Koeficijenti	Funkcija 1
MOB	0,75
SDI	0,57
BDP	0,54
INF	-0,54
GDP%	-0,50
POV	-0,20
POP	0,07

Izvor: Rezultati dobijeni primenom SPSS – a

Strukturna matrica (Tabela 3.10.) prikazuje korelacije između nezavisnih promenljivih i kanoničke diskriminacione funkcije. Redosled u strukturnoj matrici je isti kao i redosled u testu za jednakost sredina grupa, ali se razlikuje od redosleda standardizovanih koeficijenata.

Ova se razlika najverovatnije javlja zbog kolinearnosti između promenljive s*trane direktne investicije* i promenljive *bruto domaći proizvod*, naznačene u korelacionoj matrici. Strukturna matrica nije pod uticajem kolinearnosti, nekoliko prvih promenljivih najadekvatnije vrše diskriminaciju između zemalja koje nisu i zemalje koje su članice EU.

Tabela 3.11. Koeficienti kanoničke diskriminacione funkcije

Promenljive	Funkcija
	1
SDI	-0,33
GDP	2,33
GDP%	0,07
INF	1,29
MOB	0,02
POP	-0,39
POV	-1,73
Konstanta	-13,42

Izvor: Rezultati dobijeni primenom SPSS – a

Koeficienti kanoničke funkcije (Tabela 3.11.) se koriste da bi se izračunao kanonički skor promenljive za svaku zemlju.

Tabela 3.12. Funkcija za centroide grupa

Grupa zemalja	Funkcija
	1
0	-1,20
1	0,92

Izvor: Rezultati dobijeni primenom SPSS – a

U Tabeli 3.12. su prikazane nestandardizirane kanoničke diskriminacione funkcije ocenjene za sredine grupa. Prikazana je sredina kanoničke promenljive po grupama. Za prvu kanoničku promenljivu, prosečan diskriminacioni skor ili skor kanonične promenljive za zemlje koje nisu članice EU iznosi -1,2, a za zemlje koje su članice iznosi 0,923.

U Tabeli 3.13 prikazane su a priori verovatnoće za pripadnost u grupi. A priori verovatnoća je ocena verovatnoće da li bi opservacije pripale u određenu grupu u situaciji kada druga informacija nije dostupna. Ako nije drugačije naznačeno, sve a

priori verovatnoće su jednakve i njihov zbir je 1. A priori verovatnoće se koriste za klasifikaciju opservacije.

Tabela 3.13. A priori verovatnoće grupa

Grupa zemalja	A priori verovatnoća	Observacije uključene u analizu	
		Ponderirane	Neponderirane
0	0,44	20	20
1	0,57	26	26
Ukupno	1,00	46	46

Izvor: Rezultati dobijeni primenom SPSS – a

Klasifikacione funkcije se koriste da bi se opservacije razvrstale u grupe. Za svaku grupu postoji posebna funkcija. Za svaku opservaciju, klasifikacioni skor se izračunava za svaku funkciju posebno. Diskriminacioni model raspoređuje opservacije u grupu koja ima ima najveći klasifikacioni skor.

Tabela 3.14. Koeficiente klasifikacione funkcije

Promenljiva	Grupa zemalja	
	0	1
SDI	-5,11	-5,81
GDP	111,63	116,57
GDP%	32,93	33,09
INF	98,05	100,79
MOB	-0,37	-0,33
POP	-30,80	-31,63
POV	-56,45	-60,13
Konstanta	-386,90	-414,83

Izvor: Rezultati dobijeni primenom SPSS – a

Kao što se može uvideti iz Tabele 3.14 za svaku grupu zemalja imamo posebnu diskriminacionu funkciju.

Tabela 3.15 (a). Pripadnost opština klasifikacionim grupama

	Redni broj	Zemlja	Originalna grupa	Najveća grupa					Druga najveća grupa			Disk. skorovi
				Predviđena grupa	P(D>d | G=g) p	P(D>d | G=g) ss	P(G=g | D=d)	Kvadratno Mahalanobisovo rastojanje do centroida	Grupa	P(G=g | D=d)	Kvadratno Mahalanobisovo rastojanje do centroida	Funkcija 1
Originalne	1	Albanija	0	0	0,91	1	0,90	0,01	1	0,10	5,02	-1,32
	2	Ermenija	0	0	0,77	1	0,93	0,08	1	0,07	5,81	-1,49
	3	Austrija	1	1	0,86	1	0,95	0,03	0	0,05	5,31	1,11
	4	Azerbejđan	0	0	0,55	1	0,67	0,36	1	0,33	2,31	-0,60
	5	Belorusija	0	0	0,76	1	0,79	0,09	1	0,21	3,30	-0,90
	6	Belgija	1	1	0,48	1	0,98	0,49	0	0,02	7,97	1,62
	7	BIH	0	0	0,86	1	0,91	0,03	1	0,09	5,29	-1,38
	8	Bugarija	1	1	0,26	1	0,53	1,27	0	0,47	1,00	-0,20
	9	Hrvatska	0	1(**)	0,49	1	0,74	0,47	0	0,26	2,06	0,24
	10	Kipar	1	1	0,93	1	0,91	0,01	0	0,09	4,14	0,83
	11	Češka	1	1	0,87	1	0,95	0,03	0	0,05	5,25	1,09
	12	Danska	1	1	0,52	1	0,98	0,41	0	0,02	7,64	1,56
	13	Estonija	1	1	0,83	1	0,89	0,04	0	0,11	3,66	0,71
	14	Finska	1	1	0,37	1	0,65	0,79	0	0,35	1,52	0,03
	15	Francija	1	1	0,75	1	0,86	0,10	0	0,14	3,26	0,61
	16	Gruzija	0	0	0,58	1	0,96	0,31	1	0,04	7,16	-1,75
	17	Germanija	1	1	0,43	1	0,99	0,64	0	0,02	8,53	1,72
	18	Grčka	1	1	0,90	1	0,94	0,01	0	0,06	5,03	1,04
	19	Mađarska	1	1	0,60	1	0,80	0,28	0	0,20	2,54	0,40
	20	Island	0	0	0,61	1	0,71	0,26	1	0,29	2,61	-0,69
	21	Irska	1	1	0,96	1	0,93	0,00	0	0,07	4,74	0,98
	22	Italija	1	1	0,31	1	0,99	1,05	0	0,01	9,91	1,95
	23	Kazakstan	0	0	0,38	1	0,98	0,77	1	0,02	8,99	-2,08
	24	Kurgistan	0	0	0,02	1	1,00	5,07	1	0,00	19,13	-3,45
	25	Latvija	1	1	0,58	1	0,79	0,31	0	0,21	2,47	0,37
	26	Litvanija	1	1	0,98	1	0,92	0,00	0	0,08	4,40	0,90
	27	Luksembur	1	1	0,17	1	1,00	1,86	0	0,00	12,15	2,29
	28	Makedonija	0	0	0,50	1	0,63	0,46	1	0,37	2,08	-0,52
	29	Moldavija	0	0	0,53	1	0,97	0,39	1	0,04	7,56	-1,83
	30	Montenegro	0	1(**)	0,28	1	0,55	1,19	0	0,45	1,07	-0,17
	31	Holandija	1	1	0,47	1	0,98	0,52	0	0,02	8,07	1,64
	32	Norveška	0	1(**)	0,61	1	0,81	0,26	0	0,19	2,59	0,41
	33	Polska	1	1	0,56	1	0,78	0,35	0	0,22	2,35	0,33
	34	Portugalija	1	1	0,90	1	0,94	0,02	0	0,06	5,03	1,04
	35	Romanija	1	1	0,44	1	0,71	0,59	0	0,29	1,83	0,15
	36	Rusija	0	0	0,92	1	0,86	0,01	1	0,15	4,09	-1,10
	37	Srbija	0	1(**)	0,30	1	0,58	1,08	0	0,42	1,17	-0,12
	38	Slovačka	1	1	0,30	1	0,58	1,07	0	0,42	1,18	-0,11
	39	Slovenija	1	1	0,96	1	0,92	0,00	0	0,08	4,27	0,87
	40	Španija	1	1	0,91	1	0,91	0,01	0	0,09	4,05	0,81
	41	Švedska	1	1	0,53	1	0,76	0,40	0	0,24	2,22	0,29
	42	Švajcarska	0	1(**)	0,80	1	0,96	0,06	0	0,05	5,64	1,18
	43	Tađikistan	0	0	0,07	1	1,00	3,20	1	0,00	15,29	-2,99
	44	Turcija	0	1(**)	0,34	1	0,62	0,91	0	0,38	1,37	-0,03
	45	Turkmenistan	0	0	0,01	1	1,00	6,11	1	0,00	21,11	-3,67
	46	Ukraina	0	1(**)	0,34	1	0,62	0,91	0	0,38	1,37	-0,03
	47	Velika Brit.	1	1	0,30	1	0,99	1,09	0	0,01	10,01	1,96
	48	Uzbekistan	0	0	0,20	1	0,99	1,65	1	0,01	11,61	-2,49

Tabela 3.15 (b). Pripadnost opština klasifikacionim grupama

	Refni broj	Zemlja	Originalna grupa	Predvidena grupa	Najveća grupa				Druga najveća grupa		
					P(D>d | G=g)		P(G=g | D=d)	Kvadratno Mahalanobisovo rastojanje do centroida	Grupa	P(G=g | D=d)	Kvadratno Mahalanobisovo rastojanje do centroida
					p	ss					
Unakrsno validirane	1	Albanija	0	0	0,42	7	0,87	7,11	2	0,13	11,39
	2	Ermenija	0	0	0,17	7	0,90	10,27	2	0,10	15,14
	3	Austrija	1	1	1,00	7	0,94	0,76	1	0,06	5,88
	4	Azerbejđan	0	1(**)	0,00	7	1,00	384,70	1	0,00	451,35
	5	Belorusija	0	0	0,89	7	0,76	2,91	2	0,24	5,69
	6	Belgija	1	1	0,56	7	0,98	5,82	1	0,02	13,12
	7	BIH	0	0	0,98	7	0,90	1,65	2	0,10	6,66
	8	Bugarija	1	0(**)	0,30	7	0,67	8,42	2	0,33	10,39
	9	Hrvatska	0	1(**)	0,99	7	0,79	1,22	1	0,21	3,30
	10	Kipar	1	1	0,76	7	0,89	4,14	1	0,11	7,83
	11	Češka	1	1	0,98	7	0,94	1,59	1	0,06	6,60
	12	Danska	1	1	0,88	7	0,98	3,06	1	0,02	10,11
	13	Estonija	1	1	0,23	7	0,83	9,28	1	0,17	11,94
	14	Finska	1	1	0,55	7	0,55	5,88	1	0,45	5,76
	15	Francija	1	1	0,33	7	0,80	8,08	1	0,20	10,33
	16	Gruzija	0	0	0,78	7	0,95	3,98	2	0,05	10,53
	17	Germanija	1	1	0,62	7	0,98	5,34	1	0,02	13,12
	18	Grčka	1	1	0,31	7	0,92	8,30	1	0,08	12,67
	19	Mađarska	1	1	0,07	7	0,64	12,97	1	0,36	13,61
	20	Island	0	1(**)	0,00	7	0,76	27,03	1	0,24	28,84
	21	Irska	1	1	0,88	7	0,92	3,04	1	0,08	7,46
	22	Italija	1	1	0,19	7	0,99	9,96	1	0,01	18,94
	23	Kazakstan	0	0	0,20	7	0,98	9,76	2	0,02	17,72
	24	Kurgistan	0	0	0,17	7	1,00	10,32	2	0,00	25,84
	25	Latvija	1	1	0,01	7	0,54	18,72	1	0,46	18,50
	26	Litvanija	1	1	0,21	7	0,88	9,66	1	0,12	13,20
	27	Luksemburg	1	1	0,00	7	1,00	33,60	1	0,00	45,71
	28	Makedonija	0	0	0,54	7	0,52	5,98	2	0,48	6,67
	29	Moldavija	0	0	0,03	7	0,95	15,77	2	0,05	22,20
	30	Montenegro	0	1(**)	0,08	7	0,84	12,88	1	0,16	15,62
	31	Holandija	1	1	0,56	7	0,98	5,86	1	0,02	13,24
	32	Norveška	0	1(**)	0,13	7	0,97	11,22	1	0,03	17,37
	33	Polska	1	1	0,81	7	0,73	3,70	1	0,27	5,21
	34	Portugalija	1	1	0,74	7	0,93	4,34	1	0,07	8,96
	35	Romanija	1	1	0,81	7	0,65	3,72	1	0,35	4,44
	36	Rusija	0	0	0,00	7	0,62	23,15	2	0,38	24,66
	37	Srbija	0	1(**)	0,80	7	0,67	3,79	1	0,33	4,66
	38	Slovačka	1	0(**)	0,00	7	0,85	21,08	2	0,15	25,11
	39	Slovenija	1	1	0,34	7	0,88	7,97	1	0,12	11,50
	40	Španija	1	1	0,94	7	0,89	2,33	1	0,11	6,08
	41	Švedska	1	1	0,63	7	0,70	5,27	1	0,31	6,39
	42	Švajcarska	0	1(**)	0,79	7	0,99	3,90	1	0,01	12,54
	43	Tađikistan	0	0	0,37	7	1,00	7,58	2	0,00	20,29
	44	Turcija	0	1(**)	0,69	7	0,74	4,75	1	0,26	6,33
	45	Turkmenistan	0	0	0,01	7	1,00	18,11	2	0,00	36,01
	46	Ukraina	0	1(**)	0,17	7	0,85	10,36	1	0,15	13,34
	47	Velika Britanija	1	1	0,59	7	0,99	5,60	1	0,01	14,57
	48	Uzbekistan	0	0	0,10	7	0,99	12,16	2	0,01	22,36

Za originalne podatke, kvadratno Mahalanobisovo rastojanje se bazira na kanoničnoj funkciji. Za podatke unakrsne validacije, kvadratno Mahalanobisovo rastojanje se bazra i na opservacijama .

*** Pogrešno klasifikovana zemlja*

Izvor: Rezultati dobijeni primenom SPSS – a

Za svaku opservaciju izračunata je vrednost obe diskriminacione funkcije. Model diskriminacione analize klasifikuje opservaciju u onu grupu (kategoriju) koja ima veću vrednost diskriminacione funkcije.

Tabela 3.15 (a). prikazuje originalnu grupu, predviđenu grupu, posteriornu verovatnoću, kvadratno Mahalanobisovo rastojanje do centroida i diskriminacione skorove.

Za svaku opservaciju, moguće je uporediti pripadnost originalnoj i predviđenoj grupi. Model predviđa pripadnost određenoj grupi na osnovu aposteriornih verovatnoća, odnosno zemlja pripada grupi koja ima veću aposteriornu verovatnoću. Verovatnoća označena sa $P(G = g \mid D = d)$ u najvećoj grupi je posteriorna verovatnoća za pripadnost predviđenoj grupi. Ista ova verovatnoća u drugoj po veličini najvećoj grupi prikazuje za pripadnosti sledećoj najverovatnijoj grupi.

Verovatnoća označena sa $P(D > d \mid G = g)$ u delu najveće grupe, često se zove i uslovna verovatnoća i predstavlja verovatnoću za pripadnost najverovatnijjoj grupi na bazi opserviranog skora.

Opservacije koje imaju velike vredosti Mahalanobisovog rastojanja od sredine grupa mogu da se smatrati nestandardnim opservacije.

Tabela 3.16. Rezultati klasifikacije

		Grupe zemalja	Predviđena pripadnost u grupi		Ukupno
			0	1	
Originalne vrednosti	Broj	0	15,00	7,00	22,00
		1	0,00	26,00	26,00
	%	0	68,20	31,80	100,00
		1	0,00	100,00	100,00
Unakrsno validirane vrednosti (a)	Broj	0	13,00	9,00	22,00
		1	2,00	24,00	26,00
	%	0	59,10	40,90	100,00
		1	7,70	92,30	100,00

(a) Unakrsna validacija se vrši samo za one zemlje koje su uključene u analizu. U unakrsnoj validaciji, svaka zemlja se klasifikuje preko funkcije izvedene na osnovu svih zemalja, osim jedne izostavljene.

Izvor: Rezultati dobijeni primenom SPSS – a

Diskriminacioni skorovi se izračunavaju množenjem nestandardizovanih diskriminacionih koeficienta sa vrednostima nezavisne promenljive, a zatim se ovi proizvodi sumiraju i dodaje se konstanta. Za svaku diskriminacionu funkciju postoji

po jedan skor. Sredina skora za sve kombinovane opservacije je 0, dok kombinovana varijansa unutar grupa je 1.

Tabela 3.16. zadrži informacije o uspešnosti klasifikacije na bazi analiziranog uzorka: Prikazan je broj i procenat pogrešno klasifikovanih opservacija. U ovom primeru 41 zemlja (15+26) ili 85,42% je tačno klasifikovano, dok je pogrešno klasifikovano samo 7 zemalja (7+0) ili 14,58%, i to su sve zemlje koje nisu članice EU, a svrstane su u grupu 1, članica.

Ispod originalne vrednosti navedeni su rezultati za unakrsno validirane vrednosti. Uobičajeno je da originalne vrednosti daju dosta optimističke rezultate, dok unakrsno validirane vrednosti nastoje da reše ovaj problem. U unakrsno validirane vrednosti, svaka analizirana zemlja je klasifikovana preko funkcija dobijenih od svih zemalja osim, ove jedne. U primeru 37 zemalja (13+24) ili 77,08% je tačno klasifikovano. Ako je procenat za tačnu klasifikaciju unakrsno validirane vrednosti značajno manji od procenta tačne klasifikacije za originalne vrednosti, onda je moguće da ima previše nezavisnih promenljivih u modelu.

Na osnovu navedenih činjenica vidimo da model preciznije klasifikuje zemlje članice EU, i sa originalnim i sa unakrsno validiranim vrednostima. Razlog za ovakav rezultat, leži u činjenici da je Boksova M statistika nesignifikanta, odnosno da su kovarijacione matrice jednake na nivou značajnosti 0,01. Ne sprovodi se još jedna analiza gde se koristi kovarijaciona matrica za različite grupe.

Konačni rezultati potvrđuju da su sve zemlje članice Europske Unije tačno klasificirane. S druge strane, Hrvatska, Crna Gora, Norveška, Srbija, Švajcarska, Turska i Ukraina su zemlje koje nisu članice Evropske Unije, ali su u model klasifikovane kao članice Evropske Unije, ili se tretiraju kao pogrešno klasificirane opservacije. Model je dobar, jer su sve pretpostavke ispunjene i zemje su tačno klasificirane prema modelu diskriminacione analize, mada su se pojavile pogrešno klasifikovane zemlje.

Razlozi zbog kojih ove zemlje nisu članice Evropske Unije su sledeći: Norveška i Švajcarska nisu članice zbog neekonomskih razloga, jer ove zemlje ispunjavaju sve kriterijume za ulaz u Evropsku Uniju. Razlog zbog kojeg su Hrvatska, Crna Gora, Srbija, Turska i Ukraina predstavljene kao zemlje članice Evropske Unije, je što ove zemlje zadovoljavaju ekonomske kriterijume izražene preko promenljivih u analizi, kao što je visoki priliv stranih direktnih investicija i visok broj korisnika mobilne

telefonije kao jedan od indikatora razvoja. Ove zemlje zbog različitih političkih i drugih razloga još nisu članice Evropske Unije. I Rusija ima visoke vrednosti ovih promenljivih, ali se u analizi izdvaja kao nestandardna opservacija.

Promenljive koje najviše doprinose diskriminaciji su *Broj korisnika mobilne telefonije*, *Strane direktne investicije* i *Veličina bruto domaćeg razvoja*. Visoka stopa penetracije mobilne telefonije je indikator za razvoj jedne zemlje, visoki priliv stranih direktnih investicija i visoka vrednost bruto domaćeg proizvoda su ključne promenljive za članstvo u Evropskoj uniji i potencijalne zemlje kandidati i zemlje koju su aplicirale za članstvo trebaju da se fokusirati na njih.

3.8.2.Primena diskriminacione analize u slučaju više grupa za klasifikaciju opština Makedonije prema njihovim karakteristikama

Cilj ove analize je da se izvrši diskriminacija opština Makedonije prema kategorijsoj promenljivoj *Razvijenost* u tri grupe, dobijene preko metoda *k* - sredina klaster analize. Prva grupa obuhvata velike i razvijene opštine, druga grupa manje razvijene opštine i treća grupa najmanje i najnerazvijenije opštine u kojima dominiraju poljoprivredna domaćinstva. Analiza koristi sedam nezavisnih promenljivih: *Broj prodavnica u maloprodaji (MAL)*, *Ukupan broj stanovnika (STA), Ukupan broj domaćinstava (DOM), Broj obrazovanih stanovnika (OBR), Ekonomski aktivno stanovnistvo (EKO), Ukupan broj zaposlenih lica (ZAP)* i *Ukupan broj nezaposlenih lica (NEZ)*. Sve nezavisne promenljive su indikatori ekonomskog i demografskog razvoja makedonskih opština.

Diskriminaciona analiza ima dva cilja. Prvi cilj je da se odredi, grafički i algebarski, koje nezavisne promenljive najbolje razdvajaju opštine. Drugi cilj, klasifikaciona analiza nastoji da sortira opštine u tri definisane grupe. Naglasak je na formulisanju pravila koje se može koristiti, za alokaciju opštine na nabolji način u prethodno definisane grupame. Ukoliko diskriminaciona analiza daje dobre rezultate, to znači da je i podela opštine prema razvijenosti dobra i da se ista može koristiti pri kreiranju ekonomske politike zemlje za unapređivanje razvoja nerazvijenih i manje razvijenih opština.

Za validaciju diskriminacione analize, preferira se deljenje uzoraka u dva manja poduzorka, gde se jedan koristi za ocenu diskriminacione funkcije, a drugi za

validacione ciljeve. Uzorak sadrži 82 makedonske opštine i njega delimo na dva jednakva uzoraka.

Podaci za analizu su dobijeni iz oficijalne statistike popisa spoveđenog u 2002 od strane Državnog zavoda za statistiku Republike Makedonije.

Diskriminacionom analizom treba oceniti dve diskriminacione funkcije (broj grupa – 1 = 3 – 1 = 2).

Prema ranije izvršenoj klasterizaciji, prva grupa sadrži 14 opština, druga 22 opštine i treča grupa sadrži 46 opština. Jasno je da grupe imaju različite veličine i da jedino prva grupa ima manje od dvadeset opština, ali i pored toga, smatramo da je uzorak povoljan za analizu.

Prvo ispitujemo pretpostavke za primenu diskriminacione analize. Ukoliko izračunamo Mahalanobisovo D^2 rastojanje za navedene promenljive, a da prethodno ne izvršimo transformaciju, najmanja vrednost odstojanja iznosi 0,59 za opštinu Sopište, a najveća 42,54 za opštinu Tetovo. Ako se ne sprovodi transformacija , trebalo bi izbaciti iz analize kao **nestandardne opservacije** sledeće opštine: Tetovo, Karpoš, Gostivar i Prilep. Zato, dalje ispitujemo normalnost nezavisnih promenljivih. Sa obzirom da SPSS ne omogućava ispitivanje multivarijacione normalnosti, ispitujemo pojedinačnu normalnost nezavisnih promenljivih. **Kolmogorov – Smirnov-ljiv** test ukazuje da nezavisne promenljive ne slede normalan raspored.

Zbog postojanja nestandardne opservacije i zbog činjenice da promenljive ne slede normalan raspored, predlaže se logaritamska transformacija promenljivih. Nakon transformacije, Mahalanobisovo rastojanje varira od 0,96 za opštinu Probištip do 24,06 za opštinu Konče, šta znači da u ovoj fazi ne odbacujemo opštine kao nestandardne opservacije. Takođe, testovi normalnosti ukazuju da sve nezavisne promenljive slede normalni raspored, nakon transformacije.

Na ovaj način poboljšana je baza podatka za diskriminacionu analizu uz pomoč softvera SPSS.

Pokazatelji za grupe (Tabela 3.17.) ukazuju da ne postoji potencijalni problem. Za sve promenljive, veće vrednosti aritmetičke sredine su povezane sa manjim vrednostima standardne devijacije. To potvrđuje i negativna vrednost koeficijenta korelacije.

Tabela 3.17. Statistički pokazatelji za grupe opština – Provera za postojanje korelacije između aritmetičke sredine i varijanse

Grupe opština prema razvijenost	Promenljiva	Aritmetička sredina	Standardna devijacija	Validne opservacije	
				Neponderirane	Ponderirane
1	MAL	2,78	0,23	7	7
	STA	4,83	0,09	7	7
	DOM	4,29	0,08	7	7
	OBR	4,74	0,08	7	7
	EKO	4,41	0,12	7	7
	ZAP	4,22	0,17	7	7
	NEZ	3,93	0,16	7	7
2	MAL	2,25	0,33	12	12
	STA	4,43	0,13	12	12
	DOM	3,84	0,15	12	12
	OBR	4,31	0,13	12	12
	EKO	3,91	0,22	12	12
	ZAP	3,59	0,37	12	12
	NEZ	3,55	0,18	12	12
3	MAL	1,66	0,41	22	22
	STA	3,83	0,31	22	22
	DOM	3,29	0,29	22	22
	OBR	3,73	0,30	22	22
	EKO	3,37	0,34	22	22
	ZAP	3,11	0,38	22	22
	NEZ	2,98	0,33	22	22
Ukupno	MAL	2,02	0,56	41	41
	STA	4,18	0,47	41	41
	DOM	3,62	0,45	41	41
	OBR	4,07	0,46	41	41
	EKO	3,71	0,49	41	41
	ZAP	3,44	0,54	41	41
	NEZ	3,31	0,46	41	41

Izvor: Rezultati dobijeni primenom SPSS – a

Potencijal svake nezavisne promenljive pre nego što se dobije konačni model kvantifikovan je na osnovi testa jednakosti sredina grupa (Tabela 3.18.)

Tabela 3.18. Rezultati testa jednakosti sredina grupa

Promenljiva	Vilksova Lambda	F	Stepeni slobode 1	Stepeni slobode 2	Signifikantnost
MAL	0,40	28,40	2	38	0,00
STA	0,27	52,71	2	38	0,00
DOM	0,25	57,15	2	38	0,00
OBR	0,25	55,85	2	38	0,00
EKO	0,31	41,44	2	38	0,00
ZAP	0,40	28,69	2	38	0,00
NEZ	0,33	39,28	2	38	0,00

Izvor: Rezultati dobijeni primenom SPSS – a

Ovim testom ocenjen je doprinos svake nezavisne promenljive na osnovu rezultata jednofaktorske analize varijanse za nezavisnu promenljivu pri čemu je faktor, zavisna promenljiva, *Razvijenost* ($p < 0{,}10$).

U ovom modelu, sve promenljive su statistički značajne, a na to ukazuje i Vilksova Lambda.

Iz tabele se vidi da su promenljiva *Ukupan broj domaćinstva* i promenljiva *Broj obrazovanih stanovnika* najbolji diskriminatori grupa. Sledi promenljiva *ukupan broj stanovnika*, onda promenljiva *ekonomski aktivno stanovništo*, pa promenljiva *ukupan broj nezaposlenih lica*. Promenljive *broj prodavnica u maloprodaji* i *ukupan broj zaposlenih lica* najmanje doprinose diskriminaciji grupa.

Tabela 3.19. Združene kovarijaciona i korelaciona matrica (a)

Matrica	Prome-nljiva	MAL	STA	DOM	OBR	EKO	ZAP	NEZ
Kovari-jaciona	MAL	0,13	0,06	0,06	0,06	0,08	0,10	0,06
	STA	0,06	0,06	0,06	0,06	0,06	0,06	0,06
	DOM	0,06	0,06	0,05	0,05	0,06	0,06	0,06
	OBR	0,06	0,06	0,05	0,06	0,06	0,06	0,06
	EKO	0,08	0,06	0,06	0,06	0,08	0,09	0,06
	ZAP	0,10	0,06	0,06	0,06	0,09	0,12	0,05
	NEZ	0,06	0,06	0,06	0,06	0,06	0,05	0,07
Korela-ciona	MAL	1,00	0,66	0,74	0,71	0,82	0,81	0,63
	STA	0,66	1,00	0,96	0,99	0,87	0,67	0,91
	DOM	0,74	0,96	1,00	0,98	0,93	0,77	0,88
	OBR	0,71	0,99	0,98	1,00	0,91	0,73	0,90
	EKO	0,82	0,87	0,93	0,91	1,00	0,92	0,83
	ZAP	0,81	0,67	0,77	0,73	0,92	1,00	0,57
	NEZ	0,63	0,91	0,88	0,90	0,83	0,57	1,00

(a) Kovarijaciona matrica ima 38 stepena slobode.

Izvor: Rezultati dobijeni primenom SPSS – a

U ovom slučaju praktično je nemoguće izbeći korelaciju između promenljivih, jer je jasno da je ukupan broj stanovnika u korelacii sa ukupnim brojem domaćinstva ili ukupnim brojem obrazovanih stanovnika. Jasno je da korelaciona matrica dobijena združivanja korelacione matrice za grupe pokazuje jake korelacije (preko $|0{,}75|$) između nezavisnih promenljivih (Tabela 3.19.).

U Tabeli 3.20. prikazani su rezultati Boks M testa kojim se testira nulta hipoteza o jednakvosti kovarijacionih matrica populacije. Na osnovu rezultata ovog testa (p -

vrednost = 0,001<0,01), odbacuje se nulta hipotezu da su kovarijacione matrice naših grupa jednakve na nivou značajnosti 0,01.

Tabela 3.20. Provera homogenosti kovarijacionih matrica Boks M testom

Boks M test		77,80
F	Aproksimativno	2,00
	Stepen slobode 1	28
	Stepen slobode 2	1823,91
	Signifikantnost	0,001

Izvor: Rezultati dobijeni primenom SPSS – a

S obzirom da nisu jednake kovarijacione matrice sprovodimo diskriminacionu analizu još jednom, tako što u delu klasifikacije, umesto združene kovarijacione matrice grupa koristimo kovarijacionu matricu zasebne grupe. Dobijeni rezultati klasifikacije i Boksov M testa su prikazani u daljem tekstu.

Tabela 3.21. Determinante prirodnog logaritma

Grupa opštine prema razvijenosti	Rang	Determinanta prirodnog logaritma
1	(a)	(b)
2	7	-39,48
3	7	-36,26
Kombinirane unutar grupa	7	-36,14

Rang i prirodni logaritam prikazanih determinanti su iz kovarijacione matrice grupe.

(a) Rang < 7.

(b) Mali broj opština da bi matrice bile nesingularne.

Izvor: Rezultati dobijeni primenom SPSS – a

Determinante prirodnog logaritma (Tabela 3.21.) su mere varijabilnosti grupe. Varijabilnije grupe imaju veće vrednosti determinante prirodnog logaritma. Velike vrednosti determinanta između grupa ukazuju da grupe imaju različite kovarijacione matrice, što je slučaj u ovom primeru.

Konačni zaključak u vezi pretpostavki diskriminacione analize je da promenljive većinom promenljivih nisu nezavisne, da postoji multivarijaciona normalnost nezavisnih promenljivih i da su kovarijacione matrice heterogene.

Tabela 3.22. Karakteristične vrednosti

Diskriminaciona funkcija	Karakteristična vrednost	% od varijanse	Kumulativni %	Kanonička korelacija
1	3,54(a)	95,60	95,60	0,883
2	0,17(a)	4,40	100,00	0,376

(a) Prve dve kanoničke diskriminacione funkcije su korišćene u analizi.

Izvor: Rezultati dobijeni primenom SPSS – a

Tabela 3.22. daje informacije o relativnoj efikasnosti svake diskriminacione funkcije. Kao što se može utvrditi na osnovu rezultata iz tabele, skoro sav varijabilitet koji je objašnjen modelom, objašnjen je na osnovu prve diskriminacione funkcije (95,6%). Druga diskriminaciona funkcija obuhvata samo 4,4% od varijabiliteta koji je objašnjen modelom.

Na osnovu broja nezavisnih promenljivih i broja kategorija (grupa) zavisnih promenljivih potrebno je oceniti dve diskriminacione funkcije. Na osnovu njihovih karakterističnih vrednosti, iz dalje analize možemo isključiti drugu diskriminacionu funkciju.

Tabela 3.23. Vilksova Lambda

Test funkcije	Vilksova Lambda	χ^2	Stepeni slobod	Nivo značajnosti
1 kroz 2	0,189	58,319	14	0,000
2	0,859	5,335	6	0,502

Izvor: Rezultati dobijeni primenom SPSS – a

Vilksova Lambda i nivo njene značajnosti (Tabela 3.23.) ukazuje da diskriminaciona funkcija, i to prva, dobro separatiše grupe.

Finalni zaključak je da samo prva diskriminaciona funkcija dobro razdvaja grupe i da je ona jedina signifikantna funkcija modela.

Standardizovani koeficienti kanoničke diskriminacione funkcije (Tabela 3.24.) omogućavaju poređenje promenljive merene različitim mernim skalama sa visokim apsolutnim vrednostima korespondiraju promenljivi sa značajnom sposobnošću diskriminacije. U ovom slučaju to su broj obrazovanih stanovnika i ukupan broj stanovnika.

Tabela 3.24. Koeficienti standardizirane kanoničke diskriminacione funkcije

Promenljiva	Funkcija	
	1	2
MAL	-0,16	0,88
STA	-6,69	10,59
DOM	-0,48	1,14
OBR	9,14	-11,85
EKO	-1,37	3,59
ZAP	0,07	-2,65
NEZ	0,23	-1,68

Izvor: Rezultati dobijeni primenom SPSS – a

Kao što se vidi iz Tabele 3.25., sve promenljive su pozitivno i značajno korelisane sa prvom diskriminacionom funkcijom. Druga diskriminaciona funkcija nije značajno korelisana sa promenljivim, a ujedno ova diskriminaciona funkcija nije statistički značajna.

Zvezdica (*) označava značajne koeficijente korelacije između date nezavisne promenljive i diskriminacione funkcije. U okviru svake diskriminacione funkcije ovi markirani koeficijenti se uređuju apsolutnoj vrednosti. Sve promenljive su korelisane sa prvom diskriminacionom funkcijom, a to ukazuje da su sve nezavisne promenljive značajne za ovaj model diskriminacione analize.

Tabela 3.25. Strukturna matrica

Promenljiva	Funkcija	
	1	2
MAL	0,92(*)	0,13
STA	0,91(*)	0,23
DOM	0,88(*)	0,31
OBR	0,78(*)	-0,03
EKO	0,76(*)	0,30
ZAP	0,65(*)	0,02
NEZ	0,65(*)	-0,31

Kombinirane korelacije unutar grupa između nezavisne promenljive i standardizovane kanonične diskriminante.

Promenljive su rangirane na osnovu apsoulutne veličine korelacije sa funkcijom.

** Najveća apsolutna korelacija između svake promenljive i dve diskriminacione funkcije.*

Izvor: Rezultati dobijeni primenom SPSS – a

Koeficienti kanoničke promenljive (Tabele 3.26.) koriste za izračunavanje kanoničkih skorova za svaku opštinu.

Tabela 3.26. Koeficienti kanoničke diskriminacione funkcije

Promenljive	Funkcija	
	1	2
MAL	-0,45	2,42
STA	-27,24	43,11
DOM	-2,09	4,91
OBR	38,50	-49,96
EKO	-4,90	12,86
ZAP	0,21	-7,61
NEZ	0,86	-6,24
Konstanta	-19,91	-0,13

Izvor: Rezultati dobijeni primenom SPSS – a

U Tabeli 3.27. su prikazane nestandardizovane kanoničke diskriminacione funkcije ocenjene za sredine grupa. Prikazana je sredina kanonične promenljive za grupe. Sredine unutar grupa su izračunate za svaku kanoničku promenljivu. Za prvu kanoničku promenljivu, prosečan diskriminacioni skor ili skor kanoničke promenljive za zemlje koje pripadaju prvoj grupi po *Razvijenosti* iznosi 3,21, za zemlje druge grupe iznosi 0,95 i za zemlje koje spadaju u treću grupu prema *Razvijenosti* -1,54.

Tabela 3.27. Funkcija za centroide grupa

Grupe opština prema razvijenosti	Funkcija	
	1	2
1	3,21	-0,51
2	0,95	0,57
3	-1,54	-0,15

Izvor: Rezultati dobijeni primenom SPSS – a

Združena kovarijaciona matrica unutar grupa kanoničkih diskriminacionih funkcija predstavlja jediničnu matricu[31] po definiciji (Tabela 3.28).

Ispunjivanje pretpostavke za jednakost kovarijacione matrice je od izuzetne važnosti za analizu i dobijene konačne rezultate, iz tog razloga sprovodimo diskriminacionu analizu još jednom, gde u delu klasifikacije, umesto da se koristi kovarijaciona matrica unutar grupa, koristi se kovarijaciona matrica zasebnih grupa.

[31] Iz engleske reči "identity matrix".

Tabela 3.28. Grupne kovarijanse kanoničke diskriminacione funkcije

Grupe opština prema razvijenosti	Funkcija	1	2
1	1	0,18	0,01
	2	0,01	0,35
2	1	0,65	-0,48
	2	-0,48	1,34
3	1	1,42	0,25
	2	0,25	1,01

Izvor: Rezultati dobijeni primenom SPSS – a

Dobijeni rezultati Boks M testa (Tabela 3.29.) koji testira nultu hipotezu i ujedno pretpostavku za jednakost kovarijacionih matrica populacije ukazuje na sledeće: Boksova M statistika u našem primeru (p -vrednost = 0,038>0,01), ukazuje da treba prihvatiti nultu hipotezu da su kovarijacione matrice jednake na nivou značajnosti 0,01.

Tabela 3.29. Boks M test za jednakvost kovarijacione matrice kanoničke diskriminacione funkcije

Boks M test		14,89
F	Aproksimativno	2,23
	Stepen slobode1	6
	Stepen slobode 2	3337,04
	Signifikantnost	0,038

Izvor: Rezultati dobijeni primenom SPSS – a

U ovoj klasifikaciji testira se jednakost kovarijacionih matrica gde se koriste kanoničke diskriminacione funkcije, a ne originalne podaci. Sada možemo da kažemo da je pretpostavka o homogenosti kovarijacionih matrica kanoničke diskriminacione funkcije ispunjena.

S obzirom na to da su najvažnije pretpostavke uspunjene (pretpostavka za homogenost kovarijacionih matrica i multivarijaciona normalnost) možemo zaključiti da bi analiza dala validne rezultate. Rang i prirodni logaritam (Tabela 3.30.) prikazanih determinanta su oni iz kovarijacione matrice grupe kanoničke diskriminacione funkcije.

Kako su vrednosti determinanti prirodnog logaritma su niske, zaključujemo da grupe imaju malu varijabilnost.

Tabela 3.30. Determinante prirodnog logaritma

Grupe opština prema razvijenosti	Rang	Determinanta prirodnog logaritma
1	2	-2,76
2	2	-0,44
3	2	0,31
Matrica istovetnosti[32]	2	0,00

Izvor: Rezultati dobijeni primenom SPSS – a

Tabela 3.31. A priori verovatnoće grupa za pripadnost opštine grupi

Grupe opština prema razvijenosti	A priori verovatnoća	Opštine uključene u analizu	
		Ponderirane	Neponderirane
1	0,33	7	7
2	0,33	12	12
3	0,33	22	22
Ukupno	1,00	41	41

Izvor: Rezultati dobijeni primenom SPSS – a

Tabela 3.32. Koeficienti klasifikacione funkcije

Promenljiva	Razvijenost		
	1	2	3
MAL	-78,02	-74,39	-75,03
STA	-1669,83	-1561,40	-1524,88
DOM	-510,13	-500,07	-498,43
OBR	2435,82	2294,47	2234,98
EKO	292,51	317,55	320,43
ZAP	-225,22	-233,94	-228,96
NEZ	-128,18	-136,89	-134,51
Konstanta	-450,82	-401,22	-352,32

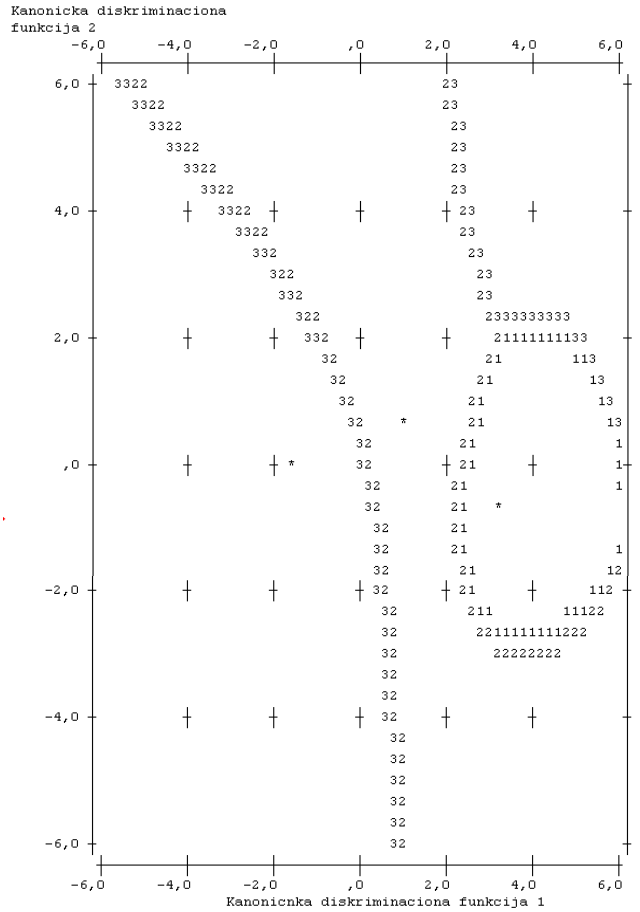
Izvor: Rezultati dobijeni primenom SPSS – a

Kao što se može videti iz Tabele 3.32., za svaku kategoriju zavisne promenljive imamo posebnu diskriminacionu funkciju. Za svaku opštinu izračunata je vrednost obe diskriminacione funkcije. Prema modelu diskriminacione analize opštine treba klasifikovati opservaciju u onu grupu (kategoriju) koja ima veću vrednost diskriminacione funkcije.

Mape prikazana na Slici 3.4. od izuzetne je koristi za proučavanja međusobnog odnosa između grupa i diskriminacionih funkcija.

[32] Iz engleske reči "identity matrix".

Slika 3.4. Teritorijalna mapa



** Označava grupni centroid, 1 označava prvu grupu, 2 označava drugu grupu i 3 označava treću grupu.*

Izvor: Rezultati dobijeni primenom SPSS – a

Imajući u vidu i rezultate matrice strukture, možemo interpretirati odnos između nezavisnih promenljivih i grupa prikazani Slikom 3.4.

Prva diskriminaciona funkcija, koja je prikazana na horizontalnoj osi, razdvaja grupu 1 (najrazvijene opštine) od grupe 3 (najmanje razvijene opštine). Upravo zbog toga što su sve promenljive izrazito visoko pozitivno korelisane sa prvom diskriminacionom funkcijom, možemo sa sigurnošću pretpostaviti da najrazvijenije opštine, imaju *Visok broj prodavnica u maloprodaji*, *Ukupan broj stanovnika, Ukupan broj domaćinstava, Ukupan broj ekonomski aktivnog stanovništva, Ukupan broj zaposlenih lica* i *Ukupan broj nezaposlenih lica*. Posledne dve promenljive imaju najmanju korelaciju u mogu se uzeti i aproksimativno.

Tabela 3.33 (a). Pripadnost opština klasifikacionim grupama

Redni broj	Opština	Originalna grupa	Najveća grupa					Druga najveća grupa			Disk. skorovi
			Predvidena grupa	P(D>d | G=g)		P(G=g | D=d)	Kvadratno Mahalanobisovo rastojanje do centroida	Grupa	P(G=g | D=d)	Kvadratno Mahalanobisovo rastojanje do centroida	Funkcija 1
				p	ss						
1	Butel	2	2	0,54	2	0,97	1,24	3	0,02	8,54	1,74
2	Gazi Baba	1	1	0,78	2	0,99	0,51	2	0,01	8,11	3,24
3	Gorče Petrov	2	2	0,30	2	0,69	2,38	1	0,30	6,35	2,14
4	Karpoš	1	1	0,98	2	0,99	0,03	2	0,01	7,77	3,17
5	Kisela Voda	1	1	0,45	2	0,98	1,61	2	0,03	6,61	3,01
6	Saraj	2	2	0,90	2	0,96	0,22	3	0,04	5,65	1,28
7	Čair	1	1	0,55	2	0,98	1,20	2	0,02	6,35	2,88
8	Šuto Orizari	2	2	0,44	2	0,69	1,63	3	0,31	2,50	-0,07
9	Aračinovo	3	3	0,02	2	0,85	7,78	2	0,15	11,95	-1,85
10	Berovo	3	3	0,49	2	0,64	1,41	2	0,36	3,32	-0,12
11	Bitola	1	1	0,15	2	1,00	3,86	2	0,00	15,41	4,02
12	Bogdanci	3	3	0,96	2	1,00	0,09	2	0,00	11,82	-1,21
13	Bogovinje	2	2	0,20	2	0,96	3,24	3	0,04	8,63	1,32
14	Bosilovo	3	3	0,58	2	0,94	1,08	2	0,06	7,37	-0,46
15	Brvenica	3	3	0,54	2	0,68	1,23	2	0,32	3,49	-0,31
16	Valandovo	3	3	0,72	2	0,97	0,66	2	0,03	8,30	-0,66
17	Vasilevo	3	3	0,93	2	1,00	0,14	2	0,00	13,68	-1,22
18	Vevcani	3	3	0,34	2	1,00	2,16	2	0,00	43,34	-3,27
19	Veles	1	1	0,53	2	0,95	1,28	2	0,05	4,90	2,73
20	Vinica	2	2	0,50	2	0,71	1,40	3	0,29	2,45	0,33
21	Vranešnica	3	3	0,08	2	1,00	5,12	2	0,00	57,46	-3,00
22	Vrapčište	2	2	0,94	2	0,95	0,12	3	0,05	5,33	1,20
23	Gevgelija	2	2	0,74	2	0,89	0,61	3	0,11	4,06	0,82
24	Gostivar	1	1	0,17	2	1,00	3,50	2	0,00	12,68	3,41
25	Gradsko	3	3	0,60	2	1,00	1,02	2	0,00	33,68	-2,73
26	Debar	2	2	0,33	2	0,72	2,25	3	0,28	3,35	-0,27
27	Debarca	3	3	0,27	2	1,00	2,62	2	0,00	15,33	-0,75
28	Delčevo	2	2	0,36	2	0,68	2,06	3	0,32	2,81	0,38
29	D. Kapija	3	3	0,81	2	1,00	0,43	2	0,00	27,10	-2,13
30	Demir Hisar	3	3	0,74	2	0,98	0,60	2	0,02	9,60	-0,76
31	Dojran	3	3	0,20	2	1,00	3,27	2	0,00	38,23	-3,32
32	Dolneni	3	2(**)	0,17	2	0,52	3,50	3	0,48	2,93	-0,56
33	Drugovo	3	3	0,61	2	1,00	0,99	2	0,00	33,39	-2,39
34	Želino	2	2	0,09	2	0,97	4,90	3	0,03	11,19	0,38
35	Zajas	3	2(**)	0,37	2	0,66	2,02	3	0,34	2,60	-0,19
36	Zelenikovo	3	3	0,46	2	1,00	1,57	2	0,00	41,38	-2,89
37	Zrnovci	3	3	0,56	2	1,00	1,15	2	0,00	34,43	-2,80
38	Ilinden	3	3	0,26	2	0,74	2,69	2	0,26	5,57	0,04
39	Jegunovce	3	3	0,59	2	0,85	1,04	2	0,15	5,31	-0,34
40	Kavadarci	2	2	0,38	2	0,81	1,96	1	0,19	7,21	2,07
41	Karbinci	3	3	0,49	2	1,00	1,42	2	0,00	34,35	-2,88

Izvor: Rezultati dobijeni primenom SPSS – a

Tabela 3.33 (b). Pripadnost opština klasifikacionim grupama

	Redni broj	Opština	Originalna grupa	Najveća grupa					Druga najveća grupa			Disk. skorovi
				Predvidena grupa	P(D>d | G=g)		P(G=g | D=d)	Kvadratno Mahalanobisovo rastojanje do centroida	Grupa	P(G=g | D=d)	Kvadratno Mahalanobisovo rastojanje do centroida	Funkcija 1
					p	ss						
Originalne	42	Kičevo	Negrupsane	2	0,94	2	0,95	0,12	3	0,05	5,23	1,16
	43	Konče	Negrupsane	3	0,02	2	1,00	7,65	2	0,00	39,27	-3,70
	44	Kočani	Negrupsane	2	0,56	2	0,96	1,16	1	0,02	11,22	1,80
	45	Kratovo	Negrupsane	3	0,37	2	0,94	2,02	2	0,07	8,10	-0,30
	46	Kriva Palanka	Negrupsane	2	0,55	2	0,90	1,21	3	0,10	4,84	0,91
	47	Krivogaštani	Negrupsane	3	0,93	2	1,00	0,14	2	0,00	20,78	-1,76
	48	Kruševo	Negrupsane	3	0,86	2	0,99	0,31	2	0,01	10,20	-0,92
	49	Kumanovo	Negrupsane	1	0,28	2	1,00	2,53	2	0,00	14,78	3,80
	50	Lipkovo	Negrupsane	2	0,12	2	0,97	4,24	3	0,03	10,15	0,16
	51	Lozovo	Negrupsane	3	0,55	2	1,00	1,20	2	0,00	37,35	-2,68
	52	Mavr. i Rostuša	Negrupsane	3	0,41	2	0,91	1,77	2	0,09	7,13	-1,09
	53	Mak. Kamenica	Negrupsane	3	0,68	2	1,00	0,78	2	0,00	12,94	-1,59
	54	Mak. Brod	Negrupsane	3	0,98	2	1,00	0,03	2	0,00	13,62	-1,38
	55	Mogila	Negrupsane	3	0,35	2	1,00	2,08	2	0,00	20,36	-1,19
	56	Negotino	Negrupsane	2	0,34	2	0,59	2,19	3	0,41	2,16	0,20
	57	Novaci	Negrupsane	3	0,42	2	1,00	1,76	2	0,00	30,75	-1,96
	58	Novo Selo	Negrupsane	3	0,28	2	0,94	2,58	2	0,07	8,67	-0,24
	59	Oslomej	Negrupsane	3	0,53	2	0,71	1,28	2	0,29	3,83	-0,45
	60	Ohrid	Negrupsane	1	0,53	2	0,96	1,28	2	0,04	5,14	2,75
	61	Petrovec	Negrupsane	3	0,89	2	1,00	0,23	2	0,00	15,38	-1,29
	62	Pehčevo	Negrupsane	3	0,94	2	1,00	0,14	2	0,00	21,58	-1,97
	63	Plašnica	Negrupsane	3	0,00	2	1,00	16,46	2	0,00	29,99	-3,45
	64	Prilep	Negrupsane	1	0,86	2	1,00	0,31	2	0,00	10,15	3,45
	65	Probištip	Negrupsane	2	0,24	2	0,72	2,88	3	0,28	4,00	0,54
	66	Radoviš	Negrupsane	2	0,85	2	0,88	0,32	3	0,12	3,48	0,49
	67	Rankovce	Negrupsane	3	0,89	2	1,00	0,25	2	0,00	21,90	-2,06
	68	Resen	Negrupsane	2	0,38	2	0,79	1,92	3	0,21	3,81	0,61
	69	Rosoman	Negrupsane	3	0,73	2	1,00	0,63	2	0,00	28,81	-2,47
	70	Sveti Nikole	Negrupsane	2	0,28	2	0,77	2,58	3	0,23	4,26	0,62
	71	Sopište	Negrupsane	3	0,84	2	1,00	0,34	2	0,00	26,34	-2,22
	72	Staro Nagorican.	Negrupsane	3	0,52	2	1,00	1,32	2	0,00	19,22	-1,26
	73	Struga	Negrupsane	2	0,04	2	0,83	6,35	1	0,12	12,51	2,41
	74	Strumica	Negrupsane	1	0,11	2	0,63	4,44	2	0,36	3,25	2,37
	75	Studeničani	Negrupsane	3	0,32	2	0,50	2,27	2	0,50	3,05	-0,43
	76	Tearce	Negrupsane	2	0,78	2	0,97	0,49	3	0,03	6,42	1,41
	77	Tetovo	Negrupsane	1	0,03	2	0,98	7,24	2	0,02	13,34	3,27
	78	Centar Župa	Negrupsane	3	0,01	2	0,81	10,33	2	0,19	13,94	-2,07
	79	Časka	Negrupsane	3	0,51	2	1,00	1,34	2	0,00	13,06	-1,69
	80	Česinovo	Negrupsane	3	0,54	2	1,00	1,25	2	0,00	13,96	-0,90
	81	Čučer - Sandovo	Negrupsane	3	0,85	2	1,00	0,33	2	0,00	18,78	-1,50
	82	Štip	Negrupsane	1	0,27	2	0,86	2,59	2	0,14	3,84	2,53

*** Pogrešno klasifikovana opština*

Izvor: Rezultati dobijeni primenom SPSS – a

Druga diskriminaciona funkcija ne razdvaja grupe, što je jasno jer ova funkcija nije bila statistički značajna.

Ako na mapi uočimo poziciju centroida svake grupe (označeni su sa zvezdicom *), možemo zaključiti da su sva tri centroida relativno blizu jedan drugom, i da razdvajanja između grupa nisu tako jasno izražena.

Prvi deo Tabele 3.32. odnosi se na opštine iz uzorka za analizu., dok druga tabela prikazuje pripadnost opština uzorku za proveru (validaciju). Tabele prikazuju originalnu grupu, predviđenu grupu, posteriorne verovatnoće, kvadratno Mahalanobisovo rastojanje do centroida i diskriminacione skorove. Za svaku opštinu, moguće je uporediti pripadnost originalnoj grupi i predviđenoj grupi. Model predviđa pripadnost određenoj grupi preko posteriorne verovatnoće, odnosno opština pripada grupi koja ima najveću posteriornu verovatnoću, određena na osnovu diskriminacionih skorova. Verovatnoća označena sa $P(G = g \mid D = d)$ u najvećoj grupi je posteriorna verovatnoća za pripadnost predviđenoj grupi.

Verovatnoća oznaćena sa $P(D > d \mid G = g)$ u delu najveće grupe, često se zove i uslovna verovatnoća i predstavlja verovatnoću za pripadnost u najverovatniju grupu na bazi opserviranog skora.

Opštine koje imaju velike vredosti Mahalanobisovog rastojanja od sredine grupa mogu da se uzmu kao nestandardne opservacije.

Diskriminacioni skorovi se izračunavaju preko množenja nestandardiziranih diskriminacionih koeficienta sa vrednostima nezavisne promenljive, onda se ovi proizvodi sabiraju i dodaje se konstanta. Za svaku diskriminacionu funkciju postoji po jedan skor. Sredina skora za sve kombinovane opštine je 0, dok združena varijansa unutar grupa je 1.

Iz prvog dela tabele se jasno može videti koja opština je tačno klasificirana, a koja nije. Od ukupno 41 opštine, samo 2, Dolneni i Zajas su pogrešno klasifikovane. Oboje su u trećoj grupi, a trebale bi biti u drugoj grupi. Znači da je tačno klasifikovano 39 opština ili 95,1%.

Na osnovu drugog dela tabele, pod pretpostavkom da ne znamo kojoj grupi pripadaju opštine, na osnovi diskriminacione analize i izabranih sedam nezavisnih ekonomskih i demografskih promenljivih, možemo utvrditi pripadnost određenoj grupi. Jer ove 41 opštine predstavljaju validacioni uzorak, u analizu nije uključena zavisna promenljiva *Razvijenost.* Ova promenljiva je dobijena prethodno na bazi k-

sredina klaster analize, i poznata nam je originalna pripadnost opština. Ukoliko uporedimo dobijene rezultate i rezultate za validacioni uzorak, može se zaključiti da je samo jedna opština Struga pogrešno klasifikovana, odnosno nalazi se u grupi 1, a na osnovu analize trebalo bi da bude u grupi 2. To znači da je u validacionom uzorku 41 opština ili 97,6% tačno klasifikovane.

Od ukupno 82 opštine tačno je klasifikovano 79 opština ili 96,34%. Ovo ukazuje da je dobar diskriminacioni model.

Tabela 3.34. Rezultati klasifikacije (a)

Frekvencije	Uzorak	Originalna pripadnost grupe opština		Predviđena pripadnost grupe opština		
				1	2	3
Apsolutne	Za analizu	1	7,00	7,00	0,00	0,00
		2	12,00	0,00	12,00	0,00
		3	22,00	0,00	2,00	20,00
	Za testiranje		41,00	6,00	11,00	24,00
Relativne	Za analizu	1	100,00	100,00	0,00	0,00
		2	100,00	0,00	100,00	0,00
		3	100,00	0,00	9,10	90,90
	Za testiranje		100,00	14,60	26,80	58,50

(a) 95,1% od originalne grupisane opštine je tačno klasifikovano.

Izvor: Rezultati dobijeni primenom SPSS – a

Mera uspeha klasifikacije korišćenog uzorka je prikazana preko broja i procenta pogrešno klasifikovanih opština samo u uzorku za analizu (Tabela 3.34.). Deo negrupisanih opština se odnosi na opštine iz uzoraka za proveru (validacije).

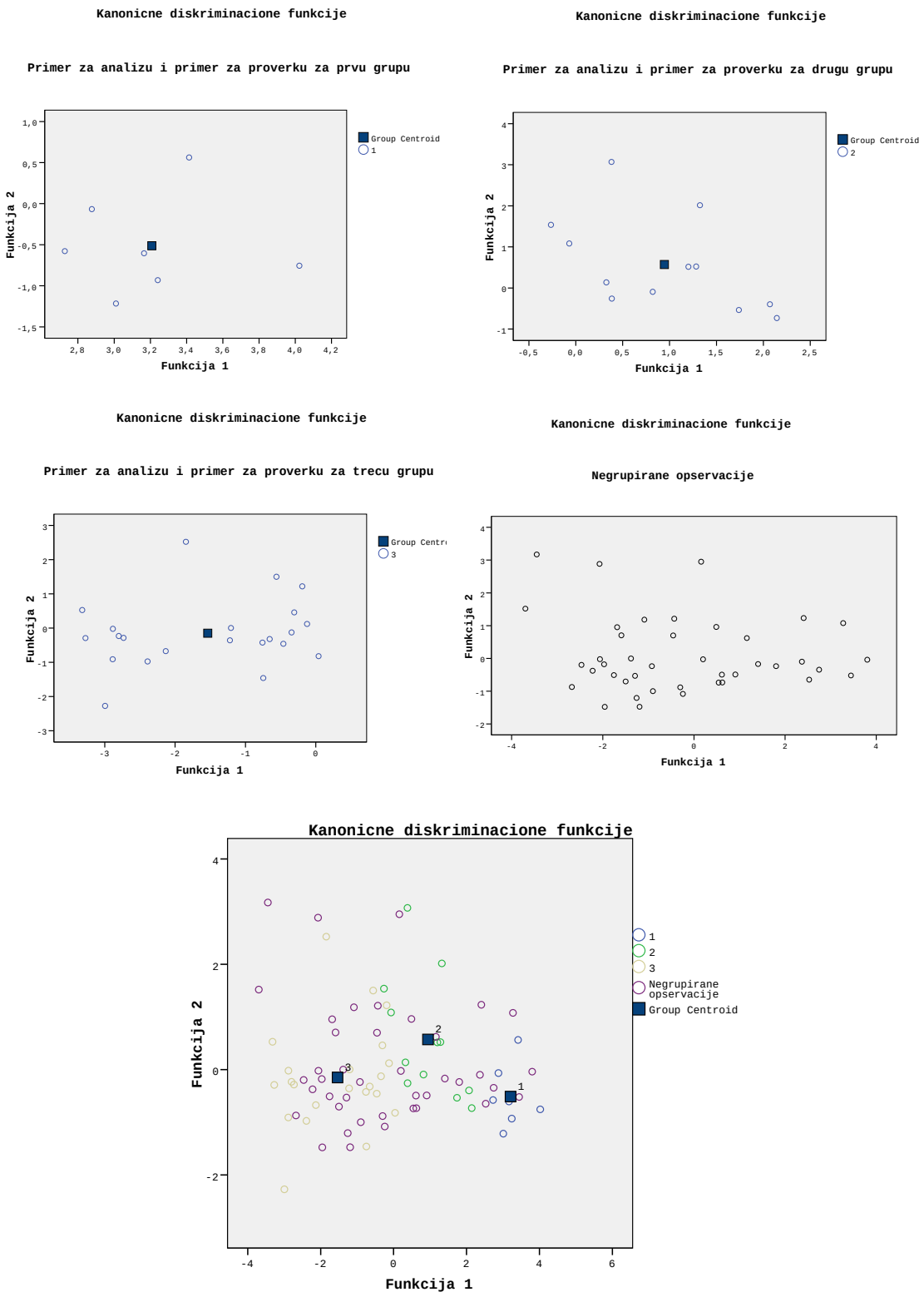
Kao što je prethodno rečeno, uzorak za analizu obuhvata 41 opština od kojih su 2 pogrešno klasifikovane, ili 39 opština (7+12+20) ili 95,1% su tačno klasifikovane. Pogrešno klasifikovane opštine dolaze iz grupe 3, a po analizi trebalo bi da budu u grupu 2.

Tabela 3.34. sadrži samo informacije za prvi uzorak, uzorak za analizu, jer za drugi uzorak nema originalne vrednosti pripadnosti da bi mogla da se proveri tačnost klasifikacije.

Grafikon kanonične diskriminacione funkcije grupa (Slika 3.5.) predstavljaju dvodimenzionalni prikaz, gde je prva diskriminaciona funkcija prikazana na horizontalnoj osi, dok je druga diskriminaciona funkcija prikazujana na vertikalnoj osi. Vrednost opservacija je prikazana preko vrednosti diskriminacionih skorova za

prvu i drugu funkciju. Treba naznačiti da rezultate koristimo samo orientaciono, jer druga diskriminaciona funkcija nije statistički značajna.

Slika 3.5. Grafikon kanonične diskriminacione funkcije grupa



Na prvom grafikonu je prikazano 7 opština iz prve grupe – najrazvijene zemlje, samo iz uzoraka za analizu. Vidimo da su dve opštine više udaljene od centroida grupe. I na drugom i na trećem grafikonu koji se odnose na podatke iz druge i treće grupe jasno se vidi da su opštine relativno udaljene od centroida.

Na Grafikonu 3.5. negrupisane opštine prikazuju opštine iz uzoraka za proveru (validaciju). Na poslednjem grafikonu su prikazane sve opštine, i one iz uzorka za analizu i one iz uzorka za validaciju, kao i centroidi dobijeni iz uzorka za analizu. Može se videti tendencija grupisanja negrupiranih opština u jednu od tri grupe i prema njihovim centroidima.

Na osnovu izloženog može se zaključiti da diskriminaciona analiza uspešno klasifikuje opštine Republike Makedonije, da sve promenljive značajno doprinose u diskriminaciji opština, a posebnu ulogu imaju promenljive *Ukupan broj domaćinstava* i promenljiva *Broj obrazovanih stanovnika*. To su promenljive koje su najbolji diskriminator grupa.

Diskriminaciona analiza daje korisne rezultate za kreiranje različitih razvojnih ekonomskih strategija države koje bi trebalo da omoguće ramnomerni ekonomski i demografski razvoj opština u Makedoniji. *Obrazovanje* je bitna promenljiva koja bi trebalo da se uzme u razmatranje, kao i *Visoki broj domaćinstava*.

"The world is full of obvious things which nobody by any chance ever observes.".

Sherlock Holmes in "The Hound of Baskervilles"

4.Analiza klasifikacionog stabla - Uvod

4.1.Analiza klasifikacionog stabla

Analiza klasifikacionog stabla kreira stabla klasifikovanja i odlučivanja, na osnovu kojih postoji mogućnost da se bolje identifikuju grupe, otkriju veze između grupa, i predvide buduća dešavanja. Upotrebom analize stabala klasifikovanja i odlučivanja donosi se zaključci o segmentaciji, stratifikaciji, prognozi, i redukciji podataka; zatim o proveravanju promenljivih, identifikovanju interakcija, spajanju kategorija; kao i kategorisanje (diskretizacija) kvantitativnih promenljivih.

Klasifikaciono stablo se koristi da se predvidi pripadnost opservacija ili objekta kategorijskima diskretnih zavisnih promenljivih na osnovu njihove vrednosti za jedne ili više nezavisnih promenljivih.

Cilj klasifikacionog stabla je da objasni zašto opservacije pripadaju određenoj grupi kategorijske zavisne promenljive ili da predvide pripadnost jednoj od dve ili više grupa kategorijske zavisne promenljive. Po svome cilju, analiza klasifikacionog stabla je dosta slična diskriminacionoj analizi, klaster analizi, neparametarskoj statistici i nelinearnim estimacijama. Fleksibilnost klasifikacionog stabla čini ovu tehniku atraktivnijom, ali ovo ne znači da bi trebalo isključiti ostale tradicionalne metode. Istina, kada su ispunjene stroge teoretske pretpostavke tradicionalnih metoda, preporučuje se njihovo korišćenje. Međutim, kada se tradicionalni metodi ne mogu koristiti, analiza klasifikaciong stabla je nenadmašljiva kao istraživačka tehnika, po mišljenju mnogih istraživača.

Upotreba klasifikacionog stabla nema široku primenu u oblastima verovatnoće i statističkog prepoznavanja šema[33]. Ali klasifikaciono stablo ima široku upotrebu u primenjenim oblastima medicine, kompjuterske nauke, botanike i psihologije. Analiza klasifikacionih stabla je grafički metod, koji je lakši za interpretiraju od analitičkih metoda. Analiza klasifikacionog stabla se može uspešno primeniti u ekonomiji i biznisu, a naročito veliku primenu ima u oblastima marketinga, istraživanja tržišta i analize kreditnog rizika.

Posebna karakteristika analize klasifikacionog stabla je **fleksibilnost**. Jedan od načina da se prikaže fleksibilnost klasifikacionog stabla je njegova sposobnost da ispituje efekat svake nezavisne promenljive posebno, a ne da ispituje efekt za sve promenljive odjedanput. Postoji nekoliko načina preko kojih se može ukazati na

[33] Od engleske reči "statistical pattern recognition".

fleksibilnost ove analize u odnosu na tradicionalne analize. Klasifikaciono stablo se može dobiti za kategorijske nezavisne promenljive, kvantitativne nezavisne promenljive ili kombinaciju ova dva tipa nezavisnih promenljivih.

Tradicinalna diskriminaciona analiza traži da se nezavisne promenljive mere na intervalnoj skali. Klasifikaciono stablo koje deli promenljive jednu po jednu i predstavlja ih na ordinalnoj skali ima takvu karakteristiku da i pored transformacije nezavisne promenljive (koja može sačuvati redosled vrednosti promenljive) dobiće iste rezultate, odnosno iste vrednosti za predviđenu pripadnost opservacije. To znači, da klasifikaciono drvo jedino traži da se nezavisne promenljive predstave bar na ordinalnoj skali i da pretpostavke oko merne skale nezavisnih promenljivih nisu tako stroge.

Klasifikaciono stable nije ograničeno samo na podelu nezavisno promenljivih jedne po jedne. Kada se kvantitativne promenljive mere na intervalnoj skali, izračunavaju se podele linearne kombinacije, koje su slične podele u diskriminacionoj analizi. Ipak, podele linearne kombinacije u klasifikacionom stablu značajno se razlikuju od onih u linearnoj diskriminacionoj analizi. Kod diskriminacione analize broj diskriminacionih funkcija je manji od broja nezavisnih promenljivih ili je jednak broju grupa kategorijske promenljive minus jedan. Ovo nije slučaj kod klasifikacionog stabla. To znači da može da se dobije onoliko linearnih kombinacija koliko ima nezavisnih promenljivih, a može imati samo dve kategorije iz zavisno promenljiva. Kada bi se koristila diskriminaciona analiza, u ovom slučaju, imali bi samo jednu linearnu funkciju i ne bi se iskoristile značajne informacije iz ostalih nezavisnih promenljivih.

Pristup implementiran za kreiranje klasifikacionog stabla za podelu linearne kombinacije može se koristiti i kao metod analize za konstrukciju klasifikacionog stabla sa podelu samo jedne promenljive[34]. U stvari, podela za jednu promenljivu je posebni slučaj deobe linearne kombinacije ako imamo deobu linearne kombinacije gde koeficijenti za kreiranje pondersane funkcije su nula, osim za jednu nezavisnu promenljivu, jer skorovi dobijeni ponderisanom funkcijom zavise samo od skorova nezavisnih promenljivih sa koeficientima različitim od nule, tako da dobijena deoba bi bila deoba za jednu promenljivu.

[34] Iz engleske reči “univariate split”.

Analiza klasifikacionog stabla sadrži četiri različite **metode rasta klasifikacionog stabla**:

1) **CHAID (Chi-squared automatic interaction detection – Hi kvadrat za automatsku detekciju interakcije)**,

2) **Iscrpni CHAID**,

3) **CART (classification and regression tree - klasifikaciono i regresiono stablo)**,

4) **QUEST (quick, unbiased, efficient statistical tree – brzo, nepristrasno, efikasno statističko stablo)**.

U analizi nad datom bazom podataka, u mogućnosti smo da upotrebimo, sva četiri različita algoritma i da se odlučimo za onaj koji najbolje modelira date podatke.

4.2.Istraživački dizajn u analizi klasifikacionog stabla

Procedura klasifikacionog stabla. Ona klasifikuje objekte ili opservacije u grupe ili predviđa vrednosti zavisne promenljive na osnovu vrednosti nezavisnih promenllivih. Početni korak u analizu klasifikacionog stabla je **definisanje cilja analize**, jer se analiza može koristiti u rešavanju različitih ekonomskih problema kao i problema iz biznis sektora. Ciljevi analize mogu da budu:

1) segmentacija – identifikacija opservacija koje imaju najveću verovatnoću da pripadaju određenoj grupi;
2) stratifikacija – raspoređivanje opservacije u jednu od nekoliko kategorija;
3) predviđanje – formulasanje pravila i njihova upotreba predviđanje novih opservacija;
4) redukcija podataka i skeniranje promenljivih – izbor najkorisnijih nezavisnih promenljivih iz skupa promenljivih za model koji se kreira;
5) identifikacija interakcije – identifikacija odnosa koji postoje između podgrupa i njihova specifikacija u modelu i
6) spajanje kategorija i transformacija kvantitativnih promenljive u kategorijsku promenljivu.

U analizi istraživač treba da naznači koja promenljiva je zavisna. A koliko i koje promenljive su nezavisne. Prema mernoj skali nezavisne i zavisne promenljive mogu biti: nominalne, ordinalne ili intervalna.

Validacija analize klasifikacionog stabla omogućava da se oceni koliko dobro dobijena struktura stabla može da se generalizira na populaciju. Postoje dva metoda za validaciju: **ukrštena validacija** i **validacija podeljenog uzorka**.

Ukrštena validacija deli uzorak u dva pod-uzoraka. Nakon toga generiraju se dva modela stabla, pri čemu se u prvi model uključuju sve opservacije iz prvog pod - uzoraka, a u drugi model se uključuju sve opservacije iz drugog pod-uzorka. Za svako stablo, rizik pogrešne klasifikacije se ocenjuje kada se dobijeno stablo iz jednog pod-uzoraka, na primer prvog, primeni na drugi pod-uzorak, i obrnuto. U SPSS-u može se koristiti maksimum 25 pod-uzoraka. Ako postoji više pod-uzoraka, manji je broj isključenih vrednosti za svaki model. Ocena rizika ukrštene validacije za konačno stablo se izračunava kako prosek rizika ostalih stabala.

Validacija podeljenog uzoraka koristi uzorak za analizu i uzorak za proveru. Uzorak za analizu može se označiti kao procenat od ukupnog uzoraka ili preko promenljive koja deli ukupni uzorak na uzorak za analizu (kod promenljive 1) i uzorak za proveru (kod promenljive 2). Rezultati mogu da se prikažu odvojeno za uzorak za analizu i za uzorak za proveru. Ukoliko je ukupni uzorak mali, onda validaciju podeljenog uzoraka treba pažljivo koristiti. Mali uzorci mogu da daju slabe modele, jer ne postoji dovoljno opservacija u određenim kategorijama da bi se izgradilo stablo.

4.3.Pretpostavke analize klasifikacionog stabla

Analiza klasifikacionog stabla zasniva se nekoliko pretpostavki koje bi trebale da se ispune da bi dobijeni rezultati analize bili relevantni. Procedura pretpostavlja da je odgovarajući **nivo merenja** pridodat svim promenljivama u analizi, odnosno za kategorijske zavisne promenljive, opis kategorija mora biti definisan za sve kategorije koje treba uključiti u analizu.

Nivo merenja ima uticaj na proračune u analizi klasifikacionog stabla, zato bi sve promenljive trebale da imaju odgovarajući nivo merenja.

Procedura klasifikacionog stabla pretpostavlja da za kategorijske zavisne promenljive, mora biti jasno definisana svaka kategorija. Neke procedure nisu moguče ukoliko bar dve kategorije zavisne promenljive nemaju opis. Ako bar dve

kategorije atributivne zavisne promenljive imaju opis, sve opservacije koje spadaju u kategorije koje nemaju opis biće isključene iz analize.

4.4.Osnovne metode formisanja klasifikacionog stabla (algoritmi analize klasifikacionog stabla)

Analiza klasifikacionog stabla sadrži četiri različite **metode formiranja klasifikacionog stabla** ili poznatiji kao **algoritmi**:

1) **CHAID**[35] ili Hi-kvadrat za automatsku detekciju interakcije: po ovom algoritmu u svakom koraku se bira nezavisna promenljiva koja ima najaču interakciju sa zavisnom promenljivom. Kategorije svake nezavisne promenljive se spajaju ako se ne razlikuju značajno u odnosu na zavisnu promenljivu.

2) **Iscrpni CHAID**: modifikacija CHAID algoritma tako da se ispituju i istražuju sve moguće deobe po svakoj nezavisnoj promenljivoj.

3) **CART**[36] ili klasifikaciono i regresiono stablo: klasifikaciono stablo binarnog tipa koje razdvaja objekte i formira precizne homogene podgrupe.

4) **QUEST**[37] ili brzo, nepristrasno, efikasno statističko stablo: To je statistički algoritam koji selektuje nepristrasno promenljive za model, i koji kreira precizno stablo (binarnog tipa) brzo i efikasno.

Svaki algoritam se može upotrebiti za analizu određene baze podataka. Između ova četiri modela može se izabrati model koji je najbolji za podatke.

4.4.1.CHAID algoritam i iscrpni CHAID algoritam

Ovaj algoritam se sastoji od tri koraka: **spajanje**, **razdvajanje** i **zaustavljanje**. Stablo raste sa ponavljavanjem ova tri koraka za svaku granu, a počinje se od prve grane[38].

Da bi objasnili algoritam i njegovu funkciju u kreiranju klasifikacionog stabla, najprije čemo definisati oznake:

[35] CHAID – od engleske reči "chi - squared automatic interaction detection".
[36] CART – od engleske reči "classification and regression tree".
[37] QUEST – od engleske reči "quick, unbiased, efficient statistical tree".
[38] Od engleske reči – "root node".

- Y označava zavisnu promenljivu koja može biti kategorijska ordinalna ili nominalna, ili kvantitativna. Ako Y je kategorijska promenljiva sa J kategorija, onda kategorija uzima vrednost $C = (1,\ldots,J)$.

- $X_m, m = 1,\ldots,M$ označava skup svih nezavisnih promenljivih koje mogu da budu kvantitativne ili kategorijske.

- $\hbar = \{x_n, y_n\}_{n=1}^{N}$ označava ukupni uzorak.

- w_n označava opservacioni ponder za opservaciju n.

- f_n označava frekvencioni ponder za opservaciju n.

CHAID algoritam prihvata samo kategorijske nezavisne promenljive. Zato, ukoliko su promenljive numeričke, transformišu se u rangove pred početak upotrebe algoritma.

Prvi korak algoritma je **spajanje**. Za svaku nezavsnu promenljivu X spajaju se kategorije koje se najmanje razlikuju. Svaka konačna kategorija iz X imaće jednu granu–dete[39] ako se X koristi da bi se razgranala grana. U korak spajanja izračunava se prilagođena p - vrednost koja se koristi i u koraku deljenja.

1) Ako X ima samo jednu kategoriju, zaustavlja se i uzima se da prilagođena p - vrednost iznosi 1.

2) Ako X ima 2 kategorije, produžava se sa korakom 8.

3) U suprotnom, pronalazi se par kategorija iz X (za ordinalne nezavisne promenljive par čine su dve susedne kategorije, dok za nominalne nezavisne promenljive su bilo koje dve kategorije) koji se najmanje razlikuje (najsličniji par). Najsličniji par je par čija statistika testa ima najveću p-vrednost u odnosu na zavisnu promenljivu Y.

4) Za par koji ima najveću p-vrednost, proverava se da li je p-vrednost veća od naznačenog nivoa spajanja, $\alpha_{spajanje}$ (alfa_spajanje). Ako je ovaj uslov ispunjen, par se spaja u jednu kategoriju i dalje se formira novi skup kategorija promenljive X. Ako nije, onda se produžava sa korakom 7.

5) (Izborno) Ako se novo formirana spojena kategorija sastoji od tri ili više orginalnih kategorija, traži se najbolje binarno razdvajanje u spojenu kategoriju čija je p - vrednost najmanja. Binarno razdvajanje se sprovodi ako njegova p-vrednost nije veća od nivoa alfa, $\alpha_{razdvajanje-spajanje}$ (alfa_razdvajanje - spajanje).

[39] Od engleske reči – "child node".

6) Vraćanje na korak 2.

7) (Izborno) Kategorija koja ima mali broj opservacija (određuje se preko naznačenog minimuma veličine segmenata od strane istraživača) se spaja sa najsličnijom kategorijom određenom preko najveće p-vrednosti.

8) Izračunava se prilagođena p-vrednost za spojene kategorije preko Bonferonijevog prilagođenja.

Drugi korak je **deljenje**. Najbolje razdvajanje za svaku nezavisnu promenljivu se nalazi u koraku spajanja. Korak deljenja bira nezavisnu promenljivu koja će najbolje razdvojiti grane. Izbor se vrši poređenjem prilagođenih p-vrednosti za svaku nezavisnu promenljivu. Prilagođena p-vrednost se dobija u fazi spajanja:

1) Izabere se nezavisna promenljiva koja ima najmanju prilagođenu p-vrednost (koja je najznačajnija).
2) Ako je prilagođena p-vrednost manja ili jednaka naznačenom alfa nivou, $\alpha_{razdvajanje}$ (alfa_razdvajanje), onda se grana deli pomoću ove nezavisne promenljive. U suprotnom, grana se ne deli i smatra se kao poslednja grana.

Poslednji korak je **zaustavljanje**. Ovaj korak proverava da li proces razgraničavanja stabla treba biti zaustavljen u saglasnosti sa sledećim pravilama zaustavljanja:

1) Ako grana postane čista, odnosno, ako sve opservacije u grani imaju identičnu vrednost za zavisnu promenljivu, grana se ne deli.
2) Ako sve opservacije u grani imaju identične vrednosti za svaku nezavisnu promenljivu, grana se ne deli.
3) Ako stablo dostigne naznačenu granicu dubine stabla koju naznači istraživač, proces razvoja stabla se zaustavlja.
4) Ako je veličina grana manja od minimalne granične veličine grana koje naznači istraživač, grana se ne deli.
5) Ako razdvajanje grane rezultira novom granom–dete, koja je manja od naznačene minimalne granice veličine grane-dete koje naznači istraživač, grana–dete imaće mali broj opservacija i spojiće se sa najsličnijom granom–dete izmerenom preko p-vrednosti. Ipak, ako je broj grana–dete 1, grana se ne deli.

Iscrpni CHAID koraci razdvajanja i zaustavljanja su isti kao u CHAID algoritmu. Korak spajanja koristi iscrpnu proceduru traženja da bi spojio bilo koji sličan par i procedura nastavlja sve dok ne ostane samo jedan par. Kako i CHAID algoritma, samo kategorijske nezavisne promenljive su dopuštene, dok se kvantitativne nezavisne promenljive transformišu u ordinalne pre započinjanja korišćenja algoritma.

Korak spajanja se sprovodi u sledećih 9 faza:

1) Ako X ima samo jednu kategoriju pretpostavlja se da prilagođena p-vrednost iznosi 1.
2) Pretpostavlja se $indeks = 0$. Izračunava se p-vrednost na osnovu skupa kategorija iz X. p-vrednost jednaka je $p(indeks) = p(0)$.
3) U suprotnom, pronalazi se par kategorija iz X, čija je razlika najmanje značajna (najsličniji). Najsličniji par je par čija statistika testa daje najveću p-vrednost u odnosu na zavisnu promenljivu Y.
4) Par koji ima najveću p-vrednost spaja se u jednu kategoriju.
5) (Izborno) Ako se formirana kategorija sastoji od tri ili više originalnih kategorija, traži se najbolje binarno razdvajanje za spojenu kategoriju, čija je p-vrednost najmanja. Ako je ova p-vrednost veća od one iz spojenih kategorija formiranih spajanjem u prethodnom koraku, sprovodi se binarno razdvajane za tu spojenu kategoriju.
6) Ažurira se $indeks = indeks + 1$, izračunava se p-vrednost za osnovni skup kategorija X. Označava se $p(indeks)$ kao p-vrednost.
7) Ponavljaju se faze 3 do 6 dok ne ostanu samo dve kategorije. Preko indeksa nalazi se skup kategorija koji ima najmanju vrednost za $p(indeks)$.
8) (Izborno) Kategorija koja ima mali broj opservacija (određuje se preko minimalnog segmenta obrađenog od strane istraživača) spaja se sa najsličnijom kategorijom izmerenom preko najveće p-vrednosti.
9) Prilagođena p-vrednost se izračunava za spojene kategorije preko Bonferonijevog prilagođenja.

Za razliku od CHAID algoritma, ovde nije potrebno odrediti nivo alfa, osim u koraku razdvajanja.

Određivanje p-vrednosti

Izračuvanje (neprilagođenih) p-vrednosti u prethodnim algoritmima zavisi od tipa zavisne promenljive.

Korak spajanja i u CHAID i u iscprnom CHAID algoritmu nekad traži p-vrednost za par kategorija X, a nekad je potrebno da ima p-vrednost za sve kategorije X. Kada je potrebna p-vrednost za par kategorija X, onda je samo deo podataka iz grane relevantan. Ako sa D označimo relevantne podatke i pretpostavimo da u D ima I kategorija iz X i J kategorija iz Y (ako je kategorijska promenljiva Y), izračunavanje p-vrednosti za podatke iz D prikazano je u nastavku.

Kvantitativna zavisna promenljiva. Ako je zavisna promenjiva Y je kvantitativna, sprovodi se F test analize varijanse, kojom se proverava da li su iste srednje vrednosti za Y za različite kategorije X. Test analize varijanse zasniva se na F-statistiku iz gde se dobija p-vrednost kao:

$$F = \frac{\sum_{i=1}^{I}\sum_{n\in D} w_n f_n I(x_n = i)(\bar{y}_i - \bar{y})^2 /(I-1)}{\sum_{i=1}^{I}\sum_{n\in D} w_n f_n I(x_n = i)(y_n - \bar{y}_i)^2 /(N_f - 1)}$$

$$p = \Pr(F(I-1, N_f - I) > F)$$

gde su:

$$\bar{y}_i = \frac{\sum_{n\in D} w_n f_n y_n I(x_n = i)}{\sum_{n\in D} w_n f_n I(x_n = i)}, \qquad \bar{y} = \frac{\sum_{n\in D} w_n f_n y_n}{\sum_{n\in D} w_n f_n}, \qquad N_f = \sum_{n\in D} f_n ,$$

i gde se $F(I-1, N_f - I)$ slučajna promenljiva koja ima F-raspored sa I i $N_f - I$ stepeni slobode.

Nominalna zavisna promenljiva. Ako je zavisna promenljiva Y nominalna kategorička promenljiva, testira se nulta hipoteza o nezavisnosti između X i Y. Kako bi se sproveo test, formira se tabela kontingencije u kojoj su kategorije za Y date u kolonama, a kategorije nezavisne promenljive X u redovima. Očekivane frekvencije za ćelije se ocenjuju za nultu hipotezu. Originalne frekvencije ćelija odgovarajuče i očekivane frekvencije ćelija se koriste da bi se izračunala Pearsonova χ^2 statistika ili statistika stope najveće verovatnoće. p-vrednost se izračunava na osnovu jedne od ove dve statistike.

Pearsonova χ^2 statistika je definisana izrazom:

$$\chi^2 = \sum_{j=1}^{J}\sum_{i=1}^{I}\frac{(n_{ij}-\hat{m}_{ij})^2}{\hat{m}_{ij}}.$$

Statistika stope najveće verovatnoće se dobija preko formule:

$$G^2 = 2\sum_{j}^{J}\sum_{i=1}^{I} n_{ij}\ln(n_{ij}/\hat{m}_{ij})$$

gde su $n_{ij} = \sum_{n\in D} f_n I(x_n = i \wedge y_n = j)$ originalne frekvencije ćelija i $\hat{m}_{ij}$ ocenjene takozvane očekivane frekvencije ćelija $(x_n = i, y_n = j)$ iz modela nezavisnosti. p-vrednost se izračunava preko $p = \Pr(\chi_d^2 > \chi^2)$ za Pearsonov χ^2 test ili $p = \Pr(\chi_d^2 > G^2)$ za test najveće verodostojnosti, gde χ_d^2 ima χ^2 raspored sa $d = (J-1)(I-1)$ stepeni slobode.

Ocena očekivane frekvencije ćelije bez pondera za opservacije je:

$$\hat{m}_{ij} = \frac{n_{i.}n_{.j}}{n_{..}}$$

gde su:

$$n_{i.} = \sum_{j=1}^{J_t} n_{ij}\ , \quad n_{.j} = \sum_{i=1}^{I_t} n_{ij}\ , \quad n_{..} = \sum_{j=1}^{J_t}\sum_{i=1}^{I_t} n_{ij}\ .$$

Ako su naznačeni ponderi za opservacije, očekivana frekvencija ćelije za testiranje nulte hipoteze nezavisnosti jednaka je:

$$m_{ij} = \overline{w}_{ij}^{-1}\alpha_i\beta_j$$

gde α_i i β_j su parametri koje treba oceniti i gde su:

$$\overline{w}_{ij} = \frac{w_{ij}}{n_{ij}},\ w_{ij} = \sum_{n\in D} w_n f_n I(x = i \wedge y_n = j).$$

Ocene parametra $\hat{\alpha}_i$, $\hat{\beta}_j$ i $\hat{m}_{ij}$ se dobijaju sledećim iterativnim postupkom:

1) $k = 0$, $\alpha_i^{(0)} = \beta_j^{(0)} = 1$, $m_{ij}^{(0)} = \overline{w}_{ij}^{-1}$

2) $\alpha_i^{(k+1)} = \dfrac{n_{i\cdot}}{\sum_j \overline{w}_{ij}^{-1}\beta_j^{(k)}} = \alpha_i^{(k)}\dfrac{n_{i.}}{\sum_j m_{ij}^{(k)}}$

3) $\beta_j^{(k+1)} = \dfrac{n_{.j}}{\sum_i \overline{w}_{ij}^{-1}\alpha_i^{(k+1)}}$

4) $m_{ij}^{(k+1)} = \overline{w}_{ij}^{-1} \alpha_i^{(k+1)} \beta_j^{(k+1)}$

5) Ако је $\max_{i,j} \left| m_{ij}^{(k+1)} - m_{ij}^{(k)} \right| < \varepsilon$, zaustavlja se postupak i rezultati $\alpha_i^{(k+1)}$, $\beta_j^{(k+1)}$ i $m_{ij}^{(k+1)}$ predstavljaju konačne ocene $\hat{\alpha}_i$, $\hat{\beta}_j$, $\hat{m}_{ij}$. U suprotnom slučaju, $k = k+1$, ide se nazad na fazu 2.

Ordinalna zavisna promenljiva. Ako je zavisna promenljiva Y kategorička i to ordinalna, nulta hipoteza o nezavisnosti između X i Y se testira nasuprot modelu efekta redova (gde su redovi kategorije za X, a kolone su kategorije za Y). Dva skupa očekivanih frekvencija ćelija, $\hat{m}_{ij}$ (hipoteza za nezavisnost) i $\hat{\hat{m}}_{ij}$ (hipoteza da podaci slede model efekta redova) se ocenjiuju. Statistika stope najveće verodostojnosti i p-vrednost su:

$$H^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} \hat{\hat{m}}_{ij} \ln(\hat{\hat{m}}_{ij} / \hat{m}_{ij})$$

$$p = \Pr(\chi^2_{I-1} > H^2).$$

Ocena očekivane frekvencije ćelije za model efekta redova: U modelu efekta redova, potrebni su skorovi za kategorije promenljive Y. Po pravilu, rang[40] kategorije promenljive Y se uzima njen skor. Istraživač može da naznači svoj skup skorova. Skorovi se postavljaju na početak stabla i ne mjenjaju se. Ako je s_j skor za kategoriju j promenljive Y, $j = 1,\ldots,J$, očekivana frekvencija ćelije za model efekta redova se dobija kao:

$$m_{ij} = \overline{w}_{ij}^{-1} \alpha_i \beta_j \gamma_i^{(s_j - \overline{s})}$$

gde je

$$\overline{s} = \sum_{j=1}^{J} w_{.j} s_j \Big/ \sum_{j=1}^{J} w_{.j}$$

i gde su $w_{.j} = \sum_i w_{ij}$, α_i, β_j i γ_i nepoznati parametri koji se ocenjuju.

Ocene parametra $\hat{\hat{\alpha}}_i$, $\hat{\hat{\beta}}_j$, $\hat{\hat{\gamma}}_i$ i iz njih $\hat{\hat{m}}_{ij}$ su dobijaju sledećom iterativnom procedurom:

1) $k = 0$, $\alpha_i^{(0)} = \beta_j^{(0)} = \gamma_i^{(0)} = 1$, $m_{ij}^{(0)} = \overline{w}_{ij}^{-1}$

[40] Od engleske reči “order”.

2) $\alpha_i^{(k+1)} = \dfrac{n_{.j}}{\sum_j \overline{w}_{ij}^{-1} \beta_j^{(k)} (y_i^{(k)})^{(s_j - \bar{s})}} = \alpha_i^{(k)} \dfrac{n_{i.}}{\sum_j m_{ij}^{(k)}}$

3) $\beta_j^{(k+1)} = \dfrac{n_{.j}}{\sum_i \overline{w}_{ij}^{-1} \alpha_i^{(k+1)} (\gamma_i^{(k)})^{(s_j - \bar{s})}}$

4) $m_{ij}^{*} = \overline{w}_{ij}^{-1} \alpha_i^{(k+1)} \beta_j^{(k+1)} (\gamma_i^{(k)})^{(s_j - \bar{s})}$, $G_i = 1 + \dfrac{\sum_j (s_j - \bar{s})(n_{ij} - m_{ij}^{*})}{\sum_j (s_j - \bar{s})^2 m_{ij}^{*}}$

5) $\gamma_i^{(k+1)} = \begin{cases} \gamma_i^{(k)} G_i & G_i > 0 \\ \gamma_i^{(k)} & \text{u suprotnom} \end{cases}$

6) $m_{ij}^{(k+1)} = \overline{w}_{ij}^{-1} \alpha_i^{(k+1)} \beta_j^{(k+1)} (\gamma_i^{(k+1)})^{(s_j - \bar{s})}$

7) Ако $\max_{i,j} \left| m_{ij}^{(k+1)} - m_{ij}^{(k)} \right| < \varepsilon$, algoritam se zaustavlja i rezultati $\alpha_i^{(k+1)}$, $\beta_j^{(k+1)}$, $\gamma_i^{(k+1)}$ i $m_{ij}^{(k+1)}$ predstavljaju konačne ocene $\hat{\hat{\alpha}}_i$, $\hat{\hat{\beta}}_j$, $\hat{\hat{\gamma}}_i$, $\hat{\hat{m}}_{ij}$. U suprotnom slučaju, $k = k + 1$, ide se nazad na fazu 2.

Bonferoni prilagođavanje

Prilagođena p-vrednost se izračunava kada se p-vrednost pomnoži sa Bonferonijevim multiplikatorom. Bonferoniev multiplikator se koristi za višestruke testove.

CHAID. Primenjuje se ako se pretpostavi da nezavisna promenljiva ima I kategorija, koje se posle koraka spajanja smanjuju na r kategorije. Bonferoniev multiplikator B je broj mogućih načina spajanja I kategorija u r kategorija. Za $r = I$, $B = 1$. Za $2 \le r \le I$, koristi se sledeća jednačina:

$$\mathrm{B} = \begin{cases} \dbinom{\mathrm{I}-1}{\mathrm{r}-1} & \text{Ordinalna nezavisna promenljiva} \\ \sum_{\mathrm{v}=0}^{\mathrm{r}-1} (-1)^{\mathrm{v}} \dfrac{(\mathrm{r}-\mathrm{v})^{\mathrm{I}}}{\mathrm{v}!(\mathrm{r}-\mathrm{v})!} & \text{Nominalna nezavisna promenljiva} \\ \dbinom{\mathrm{I}-2}{\mathrm{r}-2} + \mathrm{r}\dbinom{\mathrm{I}-2}{\mathrm{r}-1} & \text{Ordinalna nezavisna promenljiva sa nedostasujuca kategorija} \end{cases}$$

Iscrpni CHAID spaja po dve kategorije iterativno, dok ne ostanu samo dve kategorije. Bonferoniev multiplikator B je zbir broja mogućih načina za spajanje dve kategorije u svakoj iteraciji:

$$B = \begin{cases} \dfrac{I(I-1)}{2} & \text{Ordinalna nezavisna promenljiva} \\ \dfrac{I(I^2-1)}{2} & \text{Nominalna nezavisna promenljiva} \\ \dfrac{I(I-1)}{2} & \text{Ordinalna nezavisna promenljiva sa nedostasujućim kategorijama} \end{cases}$$

Nedostajuće vrednosti

Navešćemo sledeća pravila:

- ako nedostaje vrednost zavisne promenljive za opservaciju, ta promenljiva se neće koristiti u analizi.

- ako nedostaju vrednosti svih nezavisnih promenljiva, opservacija će se izostaviti.

- ako nedostaje ponder opservacije, opservacija se izostavi.

- ako frekvencioni ponder nedostaje, opservacija se izostavi.

U suprotnom slučaju, nedostajuće vrednosti se tretiraju kao kategorija nezavisne promenljive. Za ordinalne nezavisne promenljive, algoritam najpre generira "najbolji" skup kategorija koristeći sve raspoložive podatke. Posle toga, algoritam identifikuje kategoriju koja je najsličnija nedostajućim kategorijama. Na kraju, algoritam odlučuje da li da spoji nedostajuću kategoriju sa njoj njenu najsličnom kategorijom ili da je zadrži kao zasebnu kategoriju. Dve p-vrednosti se izračunavaju, jedna za skup kategorija formiranih spajanjem nedostajućih kategorija sa njima najsličnijim kategorijama, i druga za skup kategorija formiranih kad se nedostajuća kategorija doda kao zasebna kategorija. Preduzima se ona akcija koja daje najmanju p-vrednost.

Za nominalne nezavisne promenljive, nedostajuća kategorija se tretira isto kao i ostale kategorije u analizi.

4.4.2CART algoritam

CART algoritam ili klasifikaciono i regresiono stablo je binarno stablo odlučivanja, koje se konstruiše deljenjem grane na dve grane–dece, i koje se ponavlja, i gde prva grana sadrži sve opservacije iz uzoraka.

Neke od oznaka algoritma se ponavljaju isto kao i u slučaju CHAID algoritma, dok su ostale specifične samo za CART algoritam. Koriste se sledeče oznake:

– Y označava zavisnu promenljivu ili (ciljnu) promenljivu koja može biti kategorijska ili kvantitativna. Ako Y je kategorijska promenljiva sa J kategorija, onda kategorija uzima vrednost $C = (1,\ldots,J)$.

- $\mathrm{X_m}$, $\mathrm{m} = 1,\ldots,\mathrm{M}$ označava skup svih nezavisnih promenljivih koje mogu biti kategorijske ordinalne ili nominalne, ili kvantitativne.
- $\hbar = \{x_n, y_n\}_{n=1}^{N}$ označava ceo uzorak.
- $\hbar(t)$ označava uzorak koji je sadržan u grani t.
- w_n označava opservacioni ponder za opservaciju n.
- f_n označava frekvencioni ponder za opservaciju n.
- $\pi(j), j = 1,\ldots,J$ označava a priori verovatnoću za $Y = j$, $j = 1,\ldots,J$.
- $p(j,t)$, $j = 1,\ldots,J$ označava verovatnoću opservacije iz grupe j i grane t.
- $p(t)$ označava verovatnoću opservacije u grani t.
- $p(j \mid t)$, $j = 1,\ldots,J$ označava verovatnoću opservacije iz grupe j ukoliko ona spada u granu t.
- $C(i \mid j)$ označava trošak zbog pogrešne klasifikacije opservacije u kategoriju i kada opservacija spada u kategoriju j. Jasno je da $C(j \mid j) = 0$.

Proces razvijanja stabla ima osnovnu ideju - da se od svih mogućih razdvajanja grana izabere jedno razdvajanje grane, tako da dobijene grane–deca budu "najčistije". U ovom algoritmu se koriste samo univarijaciona razdvajanja. To znači da svako razdvajanje zavisi od vrednosti samo jedne nezavisne promenljive. Ako je X kategorijska nominalna promenljiva sastavljena od I kategorija, onda postoji 2^{I-1} mogućih razdvajanja za ovu nezavisnu promenljivu. Ako je X kategorijska ordinalna ili kvantitativna promenljiva sa K različitih vrednosti, onda postoji $K-1$ različitih razdvajanja za promenljivu X.

Stablo se razvija tako što se počinje od početne grane i sprovode se sledeći koraci koji se ponavljaju za svaku novu granu:

1. Za svaku nezavisnu promenljivu traži se najbolje razdvajanje. Za svaku kvantitativnu ili ordinalnu promenljivu, sortiraju se njene vrednosti od najmanje do najveće. Kada je promenljiva uređena, razmatra se svaka

vrednost od početka da bi se utvrdila tačka razdvajanja za svaku vrednost (tačka razdvajanja označava se sa v, ako važi $x \le v$, onda opservacija ide u levu granu–dete, u suprotnom ide u desnu, kako bi se odredila najbolja. Najbolja tačka razdvajanja je ona koja najviše maksimizira kriterijum razdvajanja kada se grana deli u saglasnosti sa ovim kriterijumom.

2. Za svaku nominalnu nezavisnu promenljivu, ispituje se svaki mogući podskup kategorija (naziva se A, i ako je $x \in A$, opservacija ide u levu granu–dete, u suprotnom ide u desnu) da bi se odredilo najbolje razdvajanje.
3. Traži se najboje razdvajanje grana. Iz najboljih razdvajanja grane iz koraka 1, odabire se ono koje maksimizira kriterijum razdvajanja.
4. Grana se razdvaja koristeći najbolje razdvajanje iz koraka 2, ako pravila zaustavljanja nisu zadovoljena.

Kriterijum razdvajanja i mere nečistoće[41]

Za granu t, odabire se najbolje razdvajanje s da bi ono maksimiziralo kriterijum razdvajanja $\Delta i(s,t)$. Kada može da se definiše mera nečistoće za granu, kriterijum razdvajanja korespondira sa smanjivanjem nečistoće. Kod rezultata u SPSS-u, $\Delta I(s,t) = p(t)\Delta i(s,t)$ predstavlja poboljšanje.

Kategorička zavisna promenljiva. Ako je Y kategorička promenljiva, postoje tri kriterijuma razdvajana: **Đini**, **Tvoing** i **podređeni Tvoing kriterijum**.

Za granu t neke verovatnoće $p(j,t)$, $p(t)$ i $p(j \mid t)$ se ocenjuju sa:

$$p(j,t) = \frac{\pi(j) N_{w,j}(t)}{N_{w,j}}$$

$$p(t) = \sum_j p(j,t)$$

$$p(j \mid t) = \frac{p(j,t)}{p(t)} = \frac{p(j,t)}{\sum_j p(j,t)}$$

gde su:

$$N_{w,j} = \sum_{n \in h} w_n f_n I(y_n = j)$$

[41] Od engleske reči "impurity measures".

$$N_{w,j}(t) = \sum_{n \in h(t)} w_n f_n I(y_n = j)$$

gde je $I(a = b)$ funkcija indikatora koja ima vrednost 1 kada $a = b$, a 0 u suprotnom slučaju.

Đini mera nečistoće za granu t je definisana kao:

$$i(t) = \sum_{i,j} C(i \mid j) p(i \mid t) p(j \mid t) .$$

Điniev kriterijum razdvajanja smanjivanje nečistoće definiše se kao:

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

gde su p_L i p_R verovatnoće da opservacija bude grupisana u levu granu–dete t_L i u desnu granu–dete t_L. One su ocenjene kao:

$$p_L = p(t_L) / p(t)$$

$$p_R = p(t_R) / p(t) .$$

Kada istraživač uključi i troškove, promenjene a priori verovatnoće mogu da se koriste kako bi se zamenile originalne a priori verovatnoće. Kada se koriste promenjene a priori verovatnoće, pristupa se problemu kao da ne postoje troškovi. Promenjena a priori verovatnoća se definiše kao:

$$\pi'(j) = \frac{C(j)\pi(j)}{\sum_j C(j)\pi(j)} \text{ gde } C(j) = \sum_i C(i \mid j) .$$

Tvoing kriterijum dobija se kao:

$$\Delta i(s,t) = p_L p_R \left[\sum_j \left| p(j \mid t_L) - p(j \mid t_R) \right| \right]^2 .$$

Podređeni Tvoing kriterijum se koristi kada Y je kategorijska ordinalna promenljiva. Njegov algoritam je:

1) Najpre se odvaja kategorija $C = \{1, \ldots, J\}$, iz Y preko dve super kategorije C_1 i $C_2 = C - C_1$, tako da C_1 ima formu $C_1 = \{1, \ldots, j_1\}$, $j_1 = 1, \ldots, J-1$.

2) Koristi se mera 2-kategorije $i(t) = p(C_1 \mid t) p(C_2 \mid t)$ da bi se dobilo razdvajanje $s^*(C_1)$ koje maksimizira $\Delta i(s,t)$:

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) = p_L p_R \left[\sum_{j \in C_1} \{ p(j \mid t_L) - p(j \mid t_R) \} \right]^2 .$$

3) Nalazi se super klasa C_1^* iz C_1 koja maksimizira $\Delta i(s^*(C_1), t)$.

Kvantitativna zavisna promenljiva. Kada Y je kvantitativna promenljiva, kriterijum razdvajanja $\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$ koristi za meru nečistoće najmanjih kvadrata standardne devijacije:

$$i(t) = \frac{\sum_{n \in h(t)} w_n f_n (y_n - \bar{y}(t))^2}{\sum_{n \in h(t)} w_n f_n}$$

gde su:

$$p_L = N_w(t_L) / N_w(t), \quad p_R = N_w(t_R) / N_w(t), \quad N_w(t) = \sum_{n \in h(t)} w_n f_n$$

$$\bar{y}(t) = \frac{\sum_{n \in h(t)} w_n f_n y_n}{N_w(t)}$$

Pravila zaustavljanja

Pravila zaustavljanja kontrolišu da li proces razvijanja stabla treba da se prekine ili ne. Koriste se sledeća pravila zaustavljanja:

1) Ako grana postane čista, odnosno, ako sve opservacije u grani imaju identične vrednosti za zavisnu promenljivu, onda se grana ne deli.
2) Ako sve opservacije u grani imaju identične vrednosti za svaku nezavisnu promenljivu, grana se ne deli.
3) Ako dubina stabla dostigne maksimalnu dubinu određenu od strane istraživača, proces razvijanja stabla se stopira.
4) Ako je veličina grane manja od minimalne veličine grane naznačene od strane istraživača, grana se ne deli.
5) Ako razdvajanje grane rezultira sa novom granom–dete koja je manja od naznačene minimalne granice veličine grane–dete koju odredi istraživač, grana se ne deli.
6) Ako za najbolje razdvajanje s^* za granu t, poboljšanje $\Delta I(s^*, t) = p(t)\Delta i(s^*, t)$ je manje od minimuma poboljšanja određenog od strane istraživača, grana se ne deli.

Surogat razdvajanje[42]

Za razdvajanje $X^* \le s^*$, njegovo surogat razdvajanje je razdvajanje koje koristi drugu nezavisnu promenljivu X, $X \le s_X$ (ili $X > s_X$) tako da je ovo razdvajanje nasličnije prethodnom razdvajanju i ima pozitivnu meru asocijacije za predviđanje[43]. Može biti više surogat razdvajanja. Ukoliko je veća pozitivna mera asocijacije za predviđanje, u toliko je bolje surogat razdvajanje .

Mera asocijacije za predviđanje: Ako je $\hbar_{X^* \cap X}$ skup opservacija iz analiziranog uzorka koji ne sadrži ne-nedostajuće vrednosti iz X i X^* i ako $p(s^* \approx s_X \mid t)$ je verovatnoća da se zadrži opservacija u $\hbar_{X^* \cap X}(t)$ u istoj grani i istovremeneo s^* i s_X, i $\tilde{s}_X$ je razdvajanje sa maksimalnom verovatnoćom $p(s^* \approx \tilde{s}_X \mid t) = \max_{s_X}(p(s^* \approx s_X \mid t))$, onda mera asocijacije za predviđanje $\lambda(s^* \approx \tilde{s}_X \mid t)$ između s^* i $\tilde{s}_X$ za granu t:

$$\lambda(s^* \approx \tilde{s}_X \mid t) = \frac{\min(p_L, p_R) - (1 - p(s^* \approx \tilde{s}_X \mid t))}{\min(p_L, p_R)}$$

gde je p_L (respektivno i p_R) relativna verovatnoća da najbolje razdvajanje s^* za granu t usmerena opservaciju bez nedostajuće vrednostu iz X^* u levu granu–dete, $p_L = p(t_L)/p(t)$ i $p_R = p(t_R)/p(t)$ respektivno. I gde je:

$$p(s^* \approx s_X \mid t) = \begin{cases} \sum_j \dfrac{\pi(j) N_{w,j}(s^* \approx s_X, t)}{N_{w,j}(X^* \cap X)} & \text{ako Y je kategoricka promenljiva} \\ \dfrac{N_w(s^* \approx s_X, t)}{N_w(X^* \cap X)} & \text{ako Y je kvantitativna promenljiva} \end{cases}$$

i gde su:

$$N_w(X^* \cap X) = \sum_{n \in \hbar_{X^* \cap X}} w_n f_n, \quad N_w(X^* \cap X, t) = \sum_{n \in \hbar_{X^* \cap X}(t)} w_n f_n$$

$$N_w(s^* \approx s_X, t) = \sum_{n \in \hbar_{X^* \cap X}(t)} w_n f_n I(n : s^* \approx s_X)$$

$$N_{w,j}(X^* \cap X) = \sum_{n \in \hbar_{X^* \cap X}} w_n f_n I(y_n = j), \quad N_{w,j}(X^* \cap X) = \sum_{n \in \hbar_{X^* \cap X}(t)} w_n f_n I(y_n = j)$$

$$N_{w,j}(s^* \approx s_X, t) = \sum_{n \in \hbar_{X^* \cap X}(t)} w_n f_n I(y_n = j) I(n : s^* \approx s_X)$$

[42] Iz engleske reči "surrogate split".
[43] Iz engleske reči "positive predictive measure of association".

i $I(n:s^* \approx s_X)$ je funkcija indikatora koja ima vrednost 1 kada razdvajanja s^* i s_X upućuju opservaciju u istu granu–dete, a 0 je u suprotnom.

Nedostajuće vrednosti

Pravila nedostajuće vrednosti su ista kod CART algoritma kako i kod CHAID i iscrpnog CHAID algoritama.

Metod surogat razdvajanja može se primeniti na nedostajuće vrednosti u nezavisnoj promenljivoj. Pretpostavlja se da je $X^* < s^*$ najbolje razdvajanje u grani. Ako vrednost X^* nedostaje za opservacije, najbolje surogat razdvajanje (između svih ne-nedostajućih vrednosti povezanih su sa surogat razdvajanjima) koristi se da se odluči u kojoj granu–dete treba da ide. Ako ne postoje surogat razdvajanja ili ako nedostaju sve nezavisne promenljive povezane sa surogat razdvajanjem, koristi se pravilo većine.

4.4.3.QUEST algoritam

Razvijanje klasifikacionog stabla može se izvršiti i preko **QUEST algoritma**. QUEST algoritam je skraćen naziv za brzo, nepristrasno, efikasno statističko stablo. Prestavlja klasifikacioni algoritam za strukturu stabla koji kreira binarno stablo odlučivanja.

Oznake u QUEST algoritmu su slične i u CART algoritmu:

- Y označava zavisnu ili ciljnu promenljivu. Ova promenljiva mora biti kategorijska. Ako je Y kategorijska promenljiva sa J kategorija, onda kategorija uzima vrednost $C = (1,\ldots,J)$.
- $\mathrm{X_m}$, $\mathrm{m} = 1,\ldots,\mathrm{M}$ označava skup svih nezavisnih promenljivih koje mogu biti kvantitativne (gde spadaju i ordinalne kategorijske promenljive) ili kategorijske nominalne promenljive.
- $\hbar = \{x_n, y_n\}_{n=1}^{N}$ označava ceo uzorak.
- $\hbar(t)$ označava uzorak koji je sadržan u grani t.
- f_n označava frekvencioni ponder za opservaciju n.
- N_f označava ukupni broj opservacija u uzorku za analizu, $N_f = \sum_{n \in \hbar} f_n$
- $N_{f,j}$ označava ukupni broj opservacija u uzorku za analizu za kategoriju j, $N_{f,j} = \sum_{n \in \hbar} f_n I(y_n = j)$

- $N_f(t)$ označava ukupni broj opservacija u uzorku za analizu u grani t, $N_f(t) = \sum_{n \in \hbar(t)} f_n$
- $N_{f,j}(t)$ označava ukupni broj opservacija u uzorku za analizu u kategoriji j i u grani t, $N_{f,j}(t) = \sum_{n \in \hbar(t)} f_n I(y_n = j)$
- $\pi(j), j = 1,\ldots,J$ označava a priori verovatnoću za $Y = j$, $j = 1,\ldots,J$.
- $p(j,t)$, $j = 1,\ldots,J$ označava verovatnoću opservacije iz grupe j i grane t.
- $p(j \mid t)$, $j = 1,\ldots,J$ označava verovatnoću opservacije iz grupe j ukoliko ona spada u granu t.
- $C(i \mid j)$ označava trošak zbog pogrešne klasifikacije opservacije u kategoriju i kada opservacija spada u kategoriju j. Jasno je da $C(j \mid j) = 0$.

Proces razvijanja stabla sastoji se od izbora nezavisne promenljive po kojoj će se sprovoditi razdvajanje, izbora tačke razdvajanja za izabranu nezavisnu promenljivu, i zaustavljanja. U ovom algoritmu, samo univarijaciona razdvajanja su predmet analize.

Izbor nezavisne promenljive po kojoj se sprovodi razdvajanje

Izbor nezavisne promenljive po kojoj se sprovodi razdvajanje vrši se u sledećim koracima:

1. Za svaku kvantitativnu promenljivu X sprovodi se analiza varijanse, koja testira da li sve različite kategorije zavisne promenljive Y imaju iste sredine za nezavisnu promenljivu X, i za primenjena F - statistika i izračunava se p-vrednost. Za svaku kategorijsku zavisnu promenljivu, sprovodi se i Pearson-ov χ^2 test za ispitivanje zavisnosti između Y i X, i izračunava se p-vrednost χ^2 statistike.
2. Traži se nezavisna promenljiva koja ima najmanju p-vrednost i označava se X^*.
3. Ako je najmanja p-vrednost manja od α / M, gde $\alpha \in (0,1)$ i predstavlja naznačeni nivo značajnosti od strane istraživača a M ukupni broj nezavisnih promenljivih, nezavisna promenljiva X^* se uzima kao

promenljiva za razdvajanje u grani. Ako to nije slučaj, ide se na korak 4.

4.Ako je najmanja p-vrednost veća ili jednakva α / M, onda se za svaku nezavisnu promenljivu X, izračunava Levenova F statistika koja se dobija na osnovu apsolutne razlike između promenljive X i sredine njene kategorije, kako bi se testiralo da li su iste varijanse X za različite kategorije , i izračunava se p-vrednost za svaki test:

- traži se nezavisna promenljiva koja ima najmanju p-vrednost i označava se X^{**},
- ako je najmanja p-vrednost manja od $\alpha /(M + M_1)$, gde je M_1 broj nezavisnih promenljivih, nezavisna promenljiva X^{**} se bira kao promenljiva za razdvajanje u grani. U suprotnom, grana se ne deli.

Analiza varijanse

Pretpostavimo za granu t, da postoji J_t kategorija zavisne promenljive Y. F-statistika za nezavisnu kvantitativnu promenljivu X se dobija kao:

$$F_X = \frac{\sum_{j=1}^{J_t} N_{f,j}(t)(\bar{x}^{(j)}(t) - \bar{x}(t))^2 /(J_t - 1)}{\sum_{n \in \hbar(t)} f_n (x_n - \bar{x}^{(y_n)}(t)^2 \Big/ (N_f(t) - J_t)}$$

gde su:

$$\bar{x}^{(j)}(t) = \frac{\sum_{n \in \hbar(t)} f_n x_n I(y_n = 1)}{N_{f,j}(t)}, \quad \bar{x}(t) = \frac{\sum_{n \in \hbar(t)} f_n x_n}{N_f(t)}.$$

p-vrednost F-statistike dobija se iz relacije:

$$p_X = \Pr(F(J_t - 1, N_f(t) - J_t) > F_X)$$

gde $F(J_t - 1, N_f(t) - J_t)$ ima F-raspored sa $J_t - 1$ i $N_f(t) - J_t$ stepeni slobode.

Pearsonov χ^2 test - Ako se pretpostavi da za granu t ima J_t kategorija nezavisne promenljive Y, Pearsonova χ^2 statistika za kategorijsku nezavisnu promenljivu X sa I_t je predstavljena kao:

$$X^2 = \sum_{j=1}^{J_t} \sum_{i=1}^{I_t} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$$

gde su:

$$n_{ij} = \sum_{n \in \hbar(t)} f_n I(y_n = j \wedge x_n = i), \qquad \hat{m}_{ij} = \frac{n_{i.} n_{.j}}{n_{..}}$$

$$n_{i.} = \sum_{j=1}^{J_t} n_{ij}, \quad n_{.j} = \sum_{i=1}^{I_t} n_{ij}, \quad n_{..} = \sum_{j=1}^{J_t} \sum_{i=1}^{I_t} n_{ij}$$

i gde je $I(y_n = j \wedge x_n = i) = 1$ ako opservacija n ima $y_n = j$ i $x_n = i$; a 0 je u suprotnom.

p-vrednost se dobija preko $p_X = \Pr(\chi_d^2 > X^2)$ pri čemu χ_d^2 ima χ^2 raspored sa $d = (J_t - 1)(I_t - 1)$ stepeni slobode.

Levenov F test

Za kvantitativnu promenljivu X, izračunava se $z_n = \left| x_n - \bar{x}^{(y_n)}(t) \right|$. Levenova F statistika za nezavisnu promenljivu X je F statistika iz analize varijanse za z_n.

Izbor tačke razdvajanja

Ako je X kvantitativna nezavisna promenljiva, određuje se tačka razdvajanja d tako da je $X \le d$. Ako je X nominalna kategorijska promenljiva, određuje se podskup K iz skupa svih vrednosti X-a za razdvajanje $X \in K$.

Postupak razdvajanja na osnovu kvantitativne nezavisne promenljive:

1) Grupišu se kategorije zavisne promenljive Y u dve super klase. Ako promenljiva Y ima samo dva klase, ide se na korak 2. U suprotnom, izračunavaju se sredine promenljive X za uzorak za svaku kategoriju promenljivu Y. Ako su sredine svih kategorija jednake, kategorija koja ima najveći broj opservacija je super kategorija A, a ostale kategorije su super kategorije B. Ako nisu sve sredine kategorija identične, primenjuje se metod k-sredina klaster analize sa inicijalnim centroidama klastera definisanim sa dve ekstremne sredine kategorija, kako bi se kategorije promenljive Y podelile u dve super klase: A i B. $\bar{x}_A$ i s_A^2 označavaju sredinu i varijansu za super klasu A, $\bar{x}_B$ i s_B^2 označavaju sredinu i varijansu za super klasu B.
2) Ako je $\min(s_A^2, s_B^2) = 0$, uređuju se dve super klase prema njihovim vrednosti varijansi, uređenje se vrši preko rastućeg niza i varijanse se označavaju $s_1^2 \le s_2^2$, i njihove sredine sa $\bar{x}_1, \bar{x}_2$. Neka je ε mnogo mali pozitivni broj, na primer $\varepsilon = 10^{-12}$.

3) Ako je $\bar{x}_1 < \bar{x}_2$, tada je $d = \bar{x}_1(1+\varepsilon)$. U suprotnom važi $d = \bar{x}_1(1-\varepsilon)$.

4) Ako je $\min(s_A^2, s_B^2) \neq 0$, koristi se kvadratna diskriminaciona analiza da bi se utvrdila tačka razdvajanja d. Kvadratna diskriminaciona analiza pretpostavlja da promenljiva X ima normalnom raspored u svakoj super klasi sa izračunatom sredinom i varijansom uzorka. Tačka razdvajanja je u okviru početne grane na osnovu čega je verovatnoća $\Pr(x, A \mid t) = \Pr(x, B \mid t)$ za granu t, gde je

$$\Pr(x, A \mid t) = P(x, A \mid t)P(A \mid t) = P(A \mid t)\frac{1}{\sqrt{2\pi s_A^2}}\exp\left\{-\frac{(x-\bar{x}_A)^2}{2s_A^2}\right\}$$

sa

$$p(A \mid t) = \sum_{j \in A} p(j \mid t) = \sum_{j \in A} \frac{p(j,t)}{\sum_j p(j,t)}, \quad p(j,t) = \frac{\pi(j)N_{f,j}(t)}{N_{f,j}}.$$

Rešenje za $P(x, A \mid t) = P(x, B \mid t)$ je jednako rešenju sledeće kvadratne jednačine

$$ax^2 + bx + c = 0$$

gde su:

$$a = s_A^2 - s_B^2, \quad b = 2(\bar{x}_A s_B^2 - \bar{x}_B s_A^2), \quad c = \bar{x}_B^2 s_A^2 - \bar{x}_A^2 s_B^2 + 2s_A^2 s_B^2 \log\frac{p(A \mid t)s_B}{p(B \mid t)s_A}.$$

Ako je samo jedna realna početna grana, ta grana se uzima za tačku razdvajanja, u slučaju da ona daje dve neprazne grane. Ako postoje dve realne početne grane, bira se ona koja je bliža sredini $\bar{x}_A$, u slučaju da ona daje dve neprazne grane. U suprotnom, koristi se sredina $(\bar{x}_A + \bar{x}_B)/2$ kao tačka razdvajanja.

U koraku 3, potrebna je distribucija a priori verovatnoća za zavisnu promenljivu. Kada se uključe i troškovi određeni od istraživača, promenjene a priori verovatnoće mogu da se koriste kao zamena a priori verovatnoća. Promenjena a priori verovatnoća se definiše kao:

$$\pi'(j) = \frac{C(j)\pi(j)}{\sum_j C(j)\pi(j)}, \text{ gde je } C(j) = \sum_i C(i \mid j).$$

Postupak razdvajanja na osnovu nominalne nezavisne promenljive – Ako je izabrana nezavisna promenljiva X nominalna i sadrži više od dve kategorije (ako je X binarna, tačka razdvajanja je jasna), QUEST je prvo transformiše u kvantitativnu promenljivu (označava se sa ξ) dodeljivanjem najvećih diskriminacionih koeficienata

kategorijama nezavisne promenljive. Zatim, QUEST aplicira algoritam za selekciju tačke razdvajanja za novu nezavisnu promenljivu ξ kako bi se utvrdila tačka razdvajanja.

Transformacija kategoriske promenljive u kvantitativnu promenljivu se vrši tako što se pretpostavlja da je X nominalna kategorijska nezavisna promenljiva, koja uzima vrednosti iz skupa $\{b_1,\ldots,b_I\}$. Transformacija promenljive X u kvantitativnu promenljivu ξ vrši se tako što se maksimizira količnik zbira kvadrata između kategorija i zbira kvadrata unutar kategorija za ξ (misli se na kategorije zavisne promenljive). Preciznije, postupak se sprovodi u sledećim fazama:

- Transformiše se svaka vrednost x iz X iz $\hbar$ u I-dimenzionalni veštački[44] vektor:

$$v=(v_1,\ldots,v_I)', \text{ gde } v_i=\begin{cases}1 & x=b_i\\ 0 & \text{u suprotnom}\end{cases}$$

- Izračunava se ukupna sredina i sredina kategorije j za vektor v:

$$\bar{v}=\frac{\sum_{n\in\hbar} f_n v_n}{N_f},\quad \bar{v}^{(j)}=\frac{\sum_{n\in\hbar} f_n v_n I(y_n=j)}{N_{f,j}}$$

- Izračunavaju se sledeće matrice dimenzijama $I\times I$:

$$B=\sum_{j=1}^{J} N_{f,j}(\bar{v}^{(j)}-\bar{v})(\bar{v}^{(j)}-\bar{v})',\quad T=\sum_{n\in\hbar} f_n(v_n-\bar{v})(v_n-\bar{v})'$$

- Sprovodi se dekompozicija jedne vrednosti[45] za T da bi se dobilo $T=QDQ'$, gde je Q je ortogonalna matrica dimenzija $I\times I$, $D=diag(d_1,\ldots,d_I)$ tako da je $d_1\geq\ldots\geq d_I\geq 0$. Neka je $D^{1/2}=diag(d_1^*,\ldots,d_I^*)$ gde je $d_i^*=d_i^{-1/2}$ ako je $d_i>0$, i nula u suprotnom, sprovodi se dekompozicija jedne vrednosti za $D^{-1/2}Q'BQD^{-1/2}$, kako bi se dobio njen karakteristični vektor a koji je povezan sa najvećom karakterističnom vrednošću.

- Najveća diskriminaciona koordinatna vektora v je proekcija:

$$\xi=a'D^{-1/2}Q'v.$$

Originalni QUEST algoritam transformiše kategorijsku nezavisnu promenljivu u kvantitativnu nezavisnu promenljivu u grani koja se bira na osnovu raspoloživih

[44] Iz engleske reči "dummy".
[45] Od engleske reči "single value decomposition".

podatka za tu granu. SPSS implementacija QUEST algoritma sprovodi transformaciju samo jedan put i to na samom početku na osnovu podataka iz celog uzoraka.

Zaustavljanje

Pravila zaustavljanja su ista kao i kod CART algoritma.

Nedostajuće vrednosti

Pravila nedostajuće vrednosti su ista kako i kod CART algoritma, CHAID i iscrpnog CHAID algoritma.

4.5.Dodeljivanje (asignacija) i ocena rizika

U ovom delu rada objašnjava se kako se kategorija ili vrednost dodeljuje jednoj grani i jednoj opservaciji, kao i tri **metode za ocenu rizika**: **metod ponovne zamene**, **metod testa uzorka** i **metod ukrštene validacije**. Dobijene informacije mogu da se primene na sva četiri algoritma. Pri tom, polazi se od pretpostavke da je stablo uspešno razvijeno preko uzorka za analizu prema bilo kojem od prethodno navedenih algoritama. Oznake imaju značenje kao u delu o CART algoritmu.

Dodeljivanje

Kada je stablo razvijeno, **dodeljivanje** (ili još poznato kako akcija ili odlučivanje) za svaku granu se vrši na osnovu uzoraka za analizu. Da se predvidi vrednost zavisne promenljive za jednu opservaciju, najpre se traži kojoj konačnoj grani ta opservacija, a onda se koristi dodeljivanje toj konačnoj grani za predviđavanje.

Dodeljivanje grani se sprovodi na sledeći način: za bilo koju granu t, neka je d_t dodeljivanje naznačeno za granu t:

$$d_t = \begin{cases} j^*(t) & Y \text{ je kategoricka promenljiva} \\ \bar{y}(t) & Y \text{ je kvantitativna promenljiva} \end{cases}$$

$$j^*(t) = \arg\min_{\mathrm{i}} \sum_j C(i \mid j)\, p(j \mid t)$$

$$\bar{y}(t) = \frac{1}{N_w(t)} \sum_{n \in h(t)} w_n f_n y_n$$

gde su:

$$p(j \mid t) = \frac{p(j,t)}{\sum_j p(j,t)}, \quad p(j,t) = \pi(j) \frac{N_{w,j}(t)}{N_{w,j}},$$

$$N_w = \sum_{n \in \hbar} w_n f_n \,, \quad N_{w,j} = \sum_{n \in \hbar} w_n f_n I(y_n = j) \,,$$

$$N_w(t) = \sum_{n \in \hbar(t)} w_n f_n \,, \quad N_{w,j}(t) = \sum_{n \in \hbar(t)} w_n f_n I(y_n = j) \,.$$

Ako ima više od jedne kategorije j koja dostiže minimum, bira se $j^*(t)$ koje je najmanje, tako da je j za koje $N_{f,j}(t) = \sum_{n \in \hbar(t)} f_n I(y_n = j)$ veće od 0, ili apsolutno najmanje ako je $N_{f,j}(t)$ nula kod svih njih.

Za CHAID i iscrpni CHAID u jednačini se koristi $\pi(j) = N_{w,j} / N_w$.

Za opservaciju vektora nezavisne promenljive x, dodeljivanje ili predviđanje $d_T(x)$ u stablu T je:

$$d_t = \begin{cases} j^*(t(x)) & Y \text{ je kategoricka promenljiva} \\ \bar{y}(t(x)) & Y \text{ je kvantitativna promenljiva} \end{cases}$$

gde je $t(x)$ konačna grana kojoj pripada opservacija.

Ocena rizika

Na početku treba naglasiti da ponder opservacije nije uključen pri oceni rizika, i pored toga što je uključen u procesu razvijanja stabla i dodeljivanja opservacija.

Funkcija gubitka $L(y,a)$ je funkcija u kojoj je y je faktička vrednost za Y i a je izbarano dodeljivanje. Nadalje, se koriste sledeće funkcije gubitka:

$$L(y,a) = \begin{cases} C(a \mid y) & Y \text{ je kategoricka promenljiva} \\ (y-a)^2 & Y \text{ je kvantitativna promenljiva} \end{cases}$$

Ocena rizika stabla T koristi se kada se pretpostavlja da je stablo T razvijeno i kada su dodeljivanja određena za svaku granu. Neka $\widetilde{T}$ označava skup konačnih grana stabla i neka D prestavlja skup podataka koji se koristi da bi se izračunao rizik. Ako se opservacije premeste od D u T, tada $D(t)$ označava skup svih opservacija koje spadaju u granu t. Rizik stabla dobijen za podatke iz D je ocenjen preko:

$$R(T \mid D) = \begin{cases} \sum_j \pi(j) \bar{L}_j & Y \text{ je kategoricka promenljiva} \\ \bar{L} & Y \text{ je kvantitativna promenljiva} \end{cases} =$$

$$= \begin{cases} \bar{L} & Y \text{ je kategoricka promenljiva, M1} \\ \sum_j \pi(j) \bar{L}_j & Y \text{ je kategoricka promenljiva, M2} \\ \bar{L} & Y \text{ je kvantitativna promenljiva} \end{cases}$$

gde $M1$ predstavlja emprisku prethodnu situaciju[46], a $M2$ je neempirijska prethodna situacija[47], i gde su:

$$\bar{L} = \frac{1}{N_f}\sum_{n\in D} f_n L(y_n, d_T(x_n)),\ \bar{L}_j = \frac{1}{N_{f,j}}\sum_{n\in D} f_n L(y_n, d_T(x_n))I(y_n = j),$$

$$N_f = \sum_{n\in D} f_n ,\quad N_{f,j} = \sum_{n\in D} f_n I(y_n = j).$$

Ako se pretpostavi da su $L(y_n, d_T(x_n))$ međusobno zavisni, onda se varijansa $R(T)$ ocenuje preko:

$$Var(R(T)) = \begin{cases} \sum_j \pi(j)^2 \frac{s_j^2}{N_{f,j}} & Y \text{ je kategoricka promenljiva, M2} \\ \frac{s^2}{N_f} & Y \text{ je kvantitativna ili kategoricka promenljiva i M1} \end{cases}$$

gde su:

$$s_j^2 = \frac{1}{N_{f,j}}\sum_{n\in D} f_n (L(y_n, d_T(x_n)) - \bar{L}_j)^2 I(y_n = j) = \frac{1}{N_{f,j}}\sum_{n\in D} f_n L^2(y_n, d_T(x_n))I(y_n = j) - \bar{L}_j^2$$

$$s^2 = \frac{1}{N_f}\sum_{n\in D} f_n (L(y_n, d_T(x_n)) - \bar{L}_j)^2 = \frac{1}{N_f}\sum_{n\in D} f_n L^2(y_n, d_T(x_n)) - \bar{L}_j^2 .$$

Iz svega navedenog sledi da je:

$$R(T \mid D) = \begin{cases} \frac{1}{N_f}\sum_{t\in\tilde{T}}\sum_j C(j^*(t) \mid j) N_{f,j}(t) & Y \text{ je kategorijska promenljiva, M1} \\ \sum_j \frac{\pi(j)}{N_{f,j}}\sum_{t\in\tilde{T}} C(j^*(t) \mid j) N_{f,j}(t) & Y \text{ je kategorijska promenljiva, M2} \\ \frac{1}{N_f}\sum_{t\in\tilde{T}}\sum_{n\in D(t)} f_n (y_n - \bar{y}(t))^2 & Y \text{ je kvantitativna promenljiva} \end{cases}$$

[46] Od engleske reči "emprirical prior situation".
[47] Od engleske reči "non-empirical prior situation".

$$Var(R(T \mid D)) =$$

$$= \begin{cases} \dfrac{1}{(N_f)^2}\left\{\sum_j \sum_{t\in\tilde{T}} N_{f,j}(t)C(j^*(t)\mid j)^2 - N_f R(T\mid D)^2\right\} & Y \text{ kategorijska promenljiva M1} \\ \sum_j \left(\dfrac{\pi(j)}{N_{f,j}}\right)^2 \left[\sum_{t\in\tilde{T}} N_{f,j}(t)C(j^*(t)\mid j)^2 - \dfrac{\left\{\sum_{t\in\tilde{T}} N_{f,j}(t)C(j^*(t)\mid j)\right\}^2}{N_{f,j}}\right] & Y \text{ kategorijska promenljiva M2} \\ \dfrac{1}{N_f^2}\left\{\sum_{t\in\tilde{T}}\sum_{n\in D(t)} f_n(y_n - \bar{y}(t))^4 - N_f R(T\mid D)^2\right\} & Y \text{ kvantitativna promenljiva} \end{cases}$$

gde je:

$$N_{f,j}(t) = \sum_{n\in D(t)} f_n I(y_n = j).$$

Ocenjena standardna greška za $R(T \mid D)$ se dobija preko:

$$(R(T \mid D)) = \sqrt{\text{var}(R(T \mid D))}.$$

Ocena rizika stabla se obeležava kao $R(T \mid D) = \sum_{t\in\tilde{T}} R(t \mid D)$, gde $R(t \mid D)$ predstavlja doprinos grane t riziku stabla kao:

$$R(t \mid D) = \begin{cases} \dfrac{1}{N_f}\sum_j N_{f,j}(t)C(j^*(t)\mid j) & Y \text{ je kategoricka promenljiva, M1} \\ \sum_j \dfrac{\pi(j)N_{f,j}(t)}{N_{f,j}} C(j^*(t)\mid j) & Y \text{ je kategoricka promenljiva, M2} \\ \dfrac{1}{N_f}\sum_{n\in D(t)} f_n(y_n - \bar{y}(t))^2 & Y \text{ je kvantitativna promenljiva} \end{cases}$$

Ocena rizika stabla T sa metodom ponovne zamene. Ovaj metod koristi isti skup podataka kako i uzorak za analizu $\hbar$ koji se koristi i za razvijanje stabla T da bi se izračunao rizik stabla, odnosno:

$$R(t) = R(t \mid \hbar)$$

$$R(T) = R(T \mid \hbar) = \sum_{t\in\tilde{T}} R(t)$$

$$Var(R(T)) = Var(R(T \mid \hbar)).$$

Ukrštena validacija ocene rizika za stablo T. Ukrštena validacija ocene je moguća samo kada je stablo razvijeno automatskim procesom. Neka je T stablo koje je razvijeno za sve podatke iz skupa podatka $\hbar^0$.

Neka $V \geq 2$ pozitivan ceo broj.

1) Deli se $\hbar^0$ u dva uzorka koji se međusobno isključuju $\hbar_v^{'}$, $v=1,\ldots,V$. Neka $\hbar_v$ je $\hbar^0-\hbar_v^{'}$, $v=1,\ldots,V$.
2) Za svako v, uzima se $\hbar_v$ kao uzorak za analizu i razvija se stablo T_v od skupa $\hbar_v$ koristeći ista pravila zaustavljanja koja je koristio istraživač da se razvije stablo T.
3) Kada se razvije T_v i dodeljivanje $j_v^*(t)$ ili $\bar{y}_v(t)$ za granu t iz T_v je završeno, uzima se $\hbar_v^{'}$ kao uzorak testa i na bazi njega se ocena rizika $R^{ts}(T_v)$.
4) Ponavljaju se prethodni koraci za svako $v=1,\ldots,V$. Ponderirani prosek ocena rizika uzoraka se koristi kako "V - preklop[48]" ocena rizika T u ukrštenoj validaciji.

V – preklop ocena rizika za stablo T u ukrštenoj validaciji $R^{cv}(T)$ i njena varijansa ocenjuju se preko:

$$R^{CV}(T)=\begin{cases}\sum_j \pi(j)\dfrac{1}{N_{f,j}^{)}}\sum_v N_{v,f,j}^{'}R^{ts}(T_v\mid j) & Y \text{ je kategorijska promenljiva, M2}\\ \dfrac{1}{N_f^0}\sum_v N_{v,f}^{'}R^{ts}(T_v) & Y \text{ je kvantitativna ili kategorijska promenljiva i M1}\end{cases}$$

$$Var(R^{CV}(T))=$$
$$=\begin{cases}\dfrac{1}{\left(N_f^0\right)^2}\sum_v\sum_j\sum_{t\in\tilde{T}_v} N_{v,f,j}^{'}(t)C(j_v^*(t)\mid j)^2-N_f^0R^{cv}(T)^2 & Y \text{ je kategorijska promenljiva, M1}\\ \sum_j\dfrac{\pi(j)^2}{N_{f,j}^0}\left[\sum_v\sum_{t\in\tilde{T}} N_{v,f,j}^{'}(t)C(j_v^*(t)\mid j)-\dfrac{\left\{\sum_v N_{v,f,j}^{'}R^{ts}(T_v\mid Y=j)\right\}^2}{N_{f,j}^0}\right] & Y \text{ je kategorijska promenljiva, M2}\\ \dfrac{1}{\left(N_f^0\right)^2}\left\{\sum_v\sum_{t\in\tilde{T}}\sum_{n\in\hbar_v^{'}(t)} f_n(y_n-\bar{y}_v(t))^4-N_f^0R^{cv}(T)^2\right\} & Y \text{ je kvantitativna promenljiva}\end{cases}$$

gde je:

$$N_f^0=\sum_{n\in\hbar^0}f_n\,,\quad N_{f,j}^0=\sum_{n\in\hbar^0}f_nI(y_n=j)\,,$$

$$N_{v,f}^{'}=\sum_{n\in\hbar_v^{'}}f_n\,,\ N_{v,f,j}^{'}=\sum_{n\in\hbar_v^{'}}f_nI(y_n=j)\,,\ \ N_{v,f,j}^{'}(t)=\sum_{n\in\hbar_v^{'}(t)}f_nI(y_n=j)\,.$$

[48] Od engleske reči "V – fold".

4.6.Pregled dobitka[49]

Pregled dobitka daje kratak pregled stabla pomoću deskriptivne statistike za svaku konačnu granu. Ovo omogućava da istraživač prepozna relativni doprinos svake konačne grane i da identifikuje podskupove konačnih grana koji su najkorisniji. Pregled se može koristiti za sve prikazane algoritme.

Vidovi pregleda dobitka

U zavisnosti od vida zavisne promenljive, različite statistike su prikazane za pregled dobitka.

Ako je zavisna promenljiva Y kvantitativna, onda se koristi **prosečno orijentisan pregled dobitka** gde se upotrebljavaju statistike povezane sa sredinom grane za promenljive Y. Na ovaj način, istraživač može da identifikuje konačnu granu koja daje najveći (ili najmanji) prosek za zavisnu promenljivu.

Ako je zavisna promenljiva Y kategorijska, onda se prikazuju dve statistike. Prva statistika daje **pregled dobitka ciljne kategorije**. Na bazi nje istraživač može da identifikuje konačne grane koje imaju najveći relativni doprinos ciljnoj kategoriji. Druga statistika daje pregled dobitka **prosečne vrednosti profita**. Istraživač može je koristi kada je zainteresovan da identifikuje konačne grane koje imaju relativno visoke vrednosti prosečnog profita.

Moguća su tri različita pregleda dobitka: **prikaz grane po grane, kumulativni prikaz i prikaz u procentima**. Nijhov cilj je i da pomognu istraživaču da indentifikuje bitne konačne grane kako bi razumeo rezultate stabla.

Koriste se sledeće oznake:

- Y označava zavisnu promenljivu za opservaciju n.
- f_n označava frekvencioni ponder za opservaciju n.
- N_f označava ukupni broj opservacija u D, $N_f = \sum_{n \in D} f_n$.
- $N_f(t)$ označava ukupni broj opservacija u $D(t)$, $N_f(t) = \sum_{n \in D(t)} f_n$.
- $N_{f,j}$ označava ukupni broj opservacija iz kategorije j koje spadaju u D, $N_{f,j} = \sum_{n \in D} f_n I(y_n = j)$.

[49] Od engleske reči "gain summary".

- $N_{f,j}(t)$ označava ukupni broj opservacija iz kategorije j koje spadaju u $D(t)$, $N_{f,j}(t) = \sum_{n \in D(t)} f_n I(y_n = j)$.
- $\bar{y}(t)$ označava sredinu zavisne promenjive u $D(t)$, $\bar{y}(t) = \frac{1}{N_f(t)} \sum_{n \in D(t)} f_n y_n$.
- $j^{"}$ označava ciljnu kategoriju koja je predmet interesa i uzima vrednost $\{1,\ldots,J\}$. Ovu kategoriju definiše istraživač. Ako nije naznačeno, onda je $j^{"} = 1$.
- $r(j)$, $e(j)$ označavaju respektivno prihod i troškove za kategoriju j.
- $pv(j)$ označava vrednost profita za klasu j, $pv(j) = r(j) - e(j)$.
- $j^*(\tilde{t})$ označava pripadnost kategoriji određeno preko konačne grane $\tilde{t}$.
- $\pi(j), j = 1,\ldots,J$ označava a priori verovatnoću za $Y = j$, $j = 1,\ldots,J$.
- $M1$ označava emprisku apriori situaciju za kategorijsku promenljivu Y. CHAID i iscrpni CHAID su algoritmi koji imaju prethodnu empirijsku situaciju.
- $M2$ označava neempirijsku apriori situaciju za kategorijsku promenljivu Y.

4.6.1.Pregled dobitka: Prikaz grana po grana

Pregled dobitka grana po grana koristi statistike za svaku granu. Pri tom se koriste sledeći termini i oznake:

Konačna grana označava istovetnost grane i njena oznaka je $\tilde{t}$.

Veličina: n označava ukupni broj opservacija. Broj opservacija u konačnoj grani označava se sa $N_j(\tilde{t})$.

Veličina: % označava procenat opservacija u grani. Označava se sa $p_j(\tilde{t})100\%$ gde se verovatnoča $p_j(\tilde{t})$ dobija kao:

$$p_f(\tilde{t}) = \begin{cases} \dfrac{N_f(\tilde{t})}{N_f} & M1 \text{ ili } Y \text{ je kvantitativna promenljiva} \\ \displaystyle\sum_j \frac{\pi(j) N_{f,j}(\tilde{t})}{N_{f,j}} & M2 \end{cases}$$

Dobitak: n (pregled dobitka samo za ciljne kategorije) označava ukupni broj opservacija u ciljnoj kategoriji $j^{"}$ u grani, $N_{f,j^{"}}(\tilde{t})$.

Dobitak: % (pregled dobitka samo za ciljne kategorije) označava procenat opservacija u ciljnoj kategoriji $j^{"}$ u uzorku koji pripada grani. Označava se $p_j(\tilde{t} \mid j^{"})100\%$, gde je $p_j(\tilde{t} \mid j^{"})$:

$$p_j(\tilde{t} \mid j^{"}) = \frac{N_{f,j^{"}}(\tilde{t})}{N_{f,j^{"}}}.$$

Skor. U zavisnosti od tipa pregleda dobitka, skor se različno definiše i imenuje. Oznaka ostaje ista $s(\tilde{t})$.

Odziv: % (pregled dobitka samo za ciljne kategorije). Stopa koja prikazuje odnos između broja opservacija ciljne kategorije $j^{"}$ u grani i ukupni broj opservacija u grani:

$$s(\tilde{t}) = \begin{cases} \dfrac{N_{f,j^{"}}(\tilde{t})}{N_f(\tilde{t})} & M1 \\ \dfrac{1}{p_f(\tilde{t})} \cdot \dfrac{\pi(j^{"}) N_{f,j^{"}}(\tilde{t})}{N_{f,j^{"}}} & M2 \end{cases}$$

Prosečan profit (pregled dobitka samo za prosečne vrednosti profita). Prosečna vrednost profita za granu se dobije kao:

$$s(\tilde{t}) = \begin{cases} \dfrac{\displaystyle\sum_j N_{f,j^{"}}(\tilde{t}) \cdot pv(j)}{N_f(\tilde{t})} & M1 \\ \dfrac{1}{p_f(\tilde{t})} \cdot \displaystyle\sum_j \frac{\pi(j^{"}) N_{f,j^{"}}(\tilde{t}) \cdot pv(j)}{N_{f,j^{"}}} & M2 \end{cases}$$

Sredina (pregled dobitka samo za prosečno orjentirane vrednosti). Sredina $\bar{y}(\tilde{t})$ kvantitativne zavisne promenljive Y u grani:

$$s(\tilde{t}) = \bar{y}(\tilde{t}).$$

Povratak investicija (pregled dobitka samo za prosečne vrednosti profita). Povratak investicije za granu se izračunava kao prosečni profit podeljen sa prosečnim troškovima:

$$ROI(\tilde{t}) = \frac{s(\tilde{t})}{s_0(\tilde{t})}$$

gde $s_0(\tilde{t})$ predstavlja prosečni trošak za granu $\tilde{t}$ i izračunava se preko jednačine za $s(\tilde{t})$, pri čemu se $pv(j)$ zamenjuje sa $e(j)$.

Index (%). Za pregled dobitka ciljne kategorije, indeks predstavlja stopu izražen u % koja je inače odnos između skora za granu i proporcije opservacija iz klase j'' u uzorku. Označava se sa $is(\tilde{t})100\%$ gde $(\tilde{t})$, a određuje kao:

$$is(\tilde{t}) = \begin{cases} \dfrac{s(\tilde{t})}{N_{f,j''} / N_f} & M1 \\ \dfrac{s(\tilde{t})}{\pi(j'')} & M2 \end{cases}$$

Za pregled dobitka samo za prosečne vrednosti profita, koristi se indeks u % koji se dobija kao pokazuje odnos između skora za granu i prosečne vrednosti profita za uzorak pomnožen sa 100:

$$is(\tilde{t}) = \begin{cases} \dfrac{s(\tilde{t})}{\sum_j N_{f,j''} pv(j) / N_f} & M1 \\ \dfrac{s(\tilde{t})}{\sum_j \pi(j'') pv(j)} & M2 \end{cases}$$

Za pregled dobitka samo za prosečno orientirane vrednosti, predstavlja stopa (u %) koja pokazuje odnos između skora za granu i skora $s(t=1)$ za početnu granu $t=1$, gde je :

$$is(\tilde{t}) = \frac{s(\tilde{t})}{s(t=1)}$$

Ako je imenitelj 0, ovaj indeks se ne izračunava.

4.6.2.Pregled dobitka: kumulativni prikaz

Za kumulativni prikaz pregleda dobitka, sve grane se prvo sortiraju po vrednosti skorova $s(\tilde{t})$. Da bi formule bile jednostavnije, pretpostavlja se da su izabrane grane

$\left\{\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_{|\tilde{T}|}\right\}$ su već sortirane ili u rastući ili u opadajući niz, u zavisnosti od izbora istraživača.

Konačna grana se označava sa $\tilde{t}$ i označava istovetnost ove grane.

Kumulativna veličina n**, kumulativna veličina %, kumulativni dobitak** n**, kumulativni dobitak: %**. Definisane su kao kumulativna suma stavki koje idu od grane na granu sve do konačne grane koja je predmet interesa. Ako je $a(\tilde{t}_i)$ statistika za granu po granu, onda je $\oplus a(\tilde{t}_s) = \sum_{i=1}^{s} a(\tilde{t}_i)$ njen kumulativni deo do grane $\tilde{t}_s$. Ove četiri kumulativne statistike su označene, respektivno $\oplus N_f(\tilde{t}_s)$, $\oplus p_f(\tilde{t}_s)$, $\oplus N_{f,j''}(\tilde{t}_s)$ i $\oplus p_f(\tilde{t}_s \mid j'')$.

Kumulativni skor, za kumulativni odziv, je stopa koja prikazuje odnos između broja opservacija u ciljnoj kategoriji j'' do grane i ukupnog broja opservacija u grani. Za kumulativni prosečni profit, kumulativni skor je prosečna vrednost profita do grane. Za kumulativnu sredinu, on je sredina svih y_n do grane $\tilde{t}_s$. U svim slučajevima koristi se ista formula. Ipak u računima treba koristiti odgovarajuće formule za $s(\tilde{t})$ i $p_j(\tilde{t})$. Kumulativni skor se označava sa $\oplus s(\tilde{t}_s)$.

$$\oplus s(\tilde{t}_s) = \begin{cases} \dfrac{\sum_{i=1}^{s} s(\tilde{t}_i) \cdot N_f(\tilde{t}_i)}{\sum_{i=1}^{s} N_f(\tilde{t}_i)} & M1 \text{ ili } Y \text{ kvantitativna promenljiva} \\ \dfrac{\sum_{i=1}^{s} s(\tilde{t}_i) \cdot p_f(\tilde{t}_i)}{\sum_{i=1}^{s} p_f(\tilde{t}_i)} & M2 \end{cases}$$

Kumulativna vrednost za povratak investicija (pregled dobitka samo za prosečne vrednosti profita) do grane je:

$$\oplus \text{ Povratak investicije} = \frac{\oplus s(\tilde{t}_s)}{\oplus s_0(\tilde{t}_s)}$$

gde je $\oplus s_0(\tilde{t}_s)$ kumulativni trošak koji se izračunava preko jednačine $\oplus s(\tilde{t}_s)$ gde se $pv(\tilde{t})$ zamenjuje sa $e(\tilde{t})$.

Kumulativni indeks (%) za ciljnu kategoriju kumulativni pregled dobitka, je odnos između kumulativnog skora dobitka za granu i proporcije opservacija kategorije $j^{"}$ u uzorku, izražen u %. Označava se sa $\oplus is(\widetilde{t}_s)100\%$, gde je:

$$\oplus is(\widetilde{t}_s) = \begin{cases} \dfrac{\oplus s(\widetilde{t}_s)}{N_{f,j^{"}} / N_f} & M1 \\ \dfrac{\oplus s(\widetilde{t}_s)}{\pi(j^{"})} & M2 \end{cases}$$

Za kumulativni pregled dobitka prosečne vrednosti profita, ova statistika predstavlja procentualni odnos između kumulativnog skora dobitka za granu i prosečne vrednosti profita za uzorak:

$$\oplus is(\widetilde{t}_s) = \begin{cases} \dfrac{\oplus s(\widetilde{t}_s)}{\sum_j N_{f,j^{"}} \cdot pv(j) / N_f} & M1 \\ \dfrac{\oplus s(\widetilde{t}_s)}{\sum_j \pi(j^{"}) \cdot pv(j)} & M2 \end{cases}$$

Za prosečno orijentisan kumulativni prikaz dobitka, statistika je procentualni odnos između kumulativnog skora dobitka za granu i skora $s(t=1)$ za početnu granu $t=1$, gde je:

$$\oplus is(\widetilde{t}_s) = \frac{\oplus s(\widetilde{t}_s)}{s(t=1)} = \sum_{i=1}^{s} is(\widetilde{t}_i)$$

Ako je imenitelj 0, ova statistika se ne izračunava.

4.6.3.Procentni prikaz

Kao i kod kumulativnog pregleda dobitka, sve grane se prvo sortiraju po vrednosti skorova $s(\widetilde{t})$. Da bi formule bile jednostavnije, pretpostavlja se da su grane formirane u skupu $\left\{\widetilde{t}_1, \widetilde{t}_2, \ldots, \widetilde{t}_{|\widetilde{T}|}\right\}$ već uređene ili u rastući ili u opadajući niz, u zavisnosti od odluke istraživača. Neka q bude bilo koji pozitivni celi broj deljiv sa 100. Vrednost q će se koristiti kao procentni priraštaj[50] za percentile, i definiše ga istraživač (osnovna vrednost $q=10$, ali može se promeniti). Za fiksiranu vrednost q, broj percentila koji treba da se proučava je $100/q$. p-ti percentil koji treba ispitati je

[50] Od engleske reči "increment".

$pq\%$-il, i njegova veličina je $N_{f\cdot pq} = N_f \cdot pq\%$, $p = 1,\ldots,100/q$. Za bilo koji $pq\%$-il, neka s_p i $s_p^{'}$ su dva najmanja cela broja u $\{1,\ldots,|\tilde{T}\}$ tako da je:

$$N_{f\cdot pq} \in \left(\oplus N_f(\tilde{t}_{s_p-1}), \oplus N_f(\tilde{t}_{s_p})\right], N_{f\cdot pq} \in \left[\oplus N_f(\tilde{t}_{s_p^{'}-1}), \oplus N_f(\tilde{t}_{s_p^{'}})\right)$$

gde je $\oplus N_f(\tilde{t}_0) \equiv 0$.

Konačne grane. Istovetnost svih konačnih grana koje pripadaju p-tom priraštaju. Grana $\tilde{t}$ pripada p-tom priraštaju ako $\tilde{t} \in [s_{p-1}^{'}, s_p]$.

Percentil (%) je p-ti percentil $pq\%$.

Percentil n je ukupan broj opservacija u percentilu, $N_{f\cdot pq} = \lfloor N_f \cdot pq\% \rfloor$, gde $[x]$ označava najbliži celi broj od x.

Dobitak n **(samo za percentilni prikaz dobitka ciljne kategorije)**. Ukupan broj opservacija u kategoriji $j^{"}$ koji pripadaju u $pq\%$. Označava se kao $\Diamond N_{f,j^{"}}(p)$.

$$\Diamond N_{f,j^{"}}(p) = \oplus N_{fj^{"}}(\tilde{t}_{s_p-1}) + \frac{N_{f\cdot pq} - \oplus N_f(\tilde{t}_{s_p-1})}{N_f(t_{s_p})} N_{fj^{"}}(\tilde{t}_{s_p})$$

gde je $\oplus N_{f,j}(\tilde{t}_0)$ definisano da bude 0.

Dobitak: % (samo za percentilni prikaz dobitka ciljne kategorije). Procenat opservacija u kategoriji $j^{"}$ koje pripadaju $pq\%$. Označava se kao $\Diamond p_{f,j^{"}}(p)100\%$ gde $\Diamond p_{f,j^{"}}(p)$ je:

$$\Diamond p_{f,j^{"}}(p) = \frac{\Diamond N_{f,j^{"}}(p)}{N_{f,j^{"}}}.$$

Percentilni skor za percentilni prikaz dobitka ciljne kategorije je ocena odnosa između broja opservacija u ciljnoj kategoriji $j^{"}$ koje pripadaju $pq\%$ i ukupnog broja opservacija u percentilu. Za percentilni prosečni profit, percentilni skor je ocena prosečne vrednosti profita u $pq\%$. Za prikaz dobitka prosečno orijentisanih vrednosti, percentilni skor je ocena stope svih skorova dobitka za sve grane u percentilu. Uvek se koristi ista formula:

$$\Diamond s(p) = \begin{cases} \dfrac{\oplus N_f(\tilde{t}_{s_p-1}) \cdot \oplus s(\tilde{t}_{s_p-1}) + \left\{N_{f\cdot pq} - \oplus N_f(\tilde{t}_{s_p-1}) \cdot s(\tilde{t}_{s_p})\right\}}{N_{f\cdot pq}} & M1 \\ \dfrac{\oplus p_f(\tilde{t}_{s_p-1}) \cdot \oplus s(\tilde{t}_{s_p-1}) + \left\{p_{f\cdot pq} - \oplus p_f(\tilde{t}_{s_p-1}) \cdot s(\tilde{t}_{s_p})\right\}}{p_{f\cdot pq}} & M2 \end{cases}$$

gde je:

$$p_{f\cdot pq} = \oplus p_f(\tilde{t}_{s_p-1}) + \frac{N_{f\cdot pq} - \oplus N_f(\tilde{t}_{s_p-1})}{N_f(\tilde{t}_{s_p})} p_f(\tilde{t}_{s_p})$$

Percentilna vrednost povraćaja investicija (pregled dobitka samo za prosečne vrednosti profita). Definicija percentilne vrednosti povraćaja investicija je:

$$\Diamond \text{ Povratak investicije (p)} = \frac{\Diamond \text{s(p)}}{\Diamond \text{s}_0\text{(p)}}$$

gde je $\Diamond s_0(p)$ percentilni trošak koji se izračunava u obrascu za $\Diamond s(p)$ gde se $pv(\tilde{t})$ zamenjuje sa $e(\tilde{t})$.

Percentilni indeks (%) za ciljnu kategoriju percentilnog pregleda dobitka je procentualni odnos između percentilnog skora dobitka za $pq\%$ i proporcije opservacija kategorije j'' u uzorku. Označava se sa $\Diamond is(\tilde{t}_s)100\%$, gde je $\Diamond is(\tilde{t}_s)$:

$$\Diamond is(p) = \begin{cases} \dfrac{\Diamond s(p)}{N_{f,j''} / N_f} & M1 \\ \dfrac{\Diamond s(p)}{\pi(j'')} & M2 \end{cases}$$

Za percentilni pregled dobitka prosečne vrednosti profita, ova statistika predstavlja odnos (u %) između percentilnog skora dobitka za $pq\%$ i prosečne vrednosti profita za uzorak:

$$\Diamond is(p) = \begin{cases} \dfrac{\Diamond s(p)}{\sum_j N_{f,j''} \cdot pv(j) / N_f} & M1 \\ \dfrac{\Diamond s(p)}{\sum_j \pi(j'') \cdot pv(j)} & M2 \end{cases}$$

Za prosečno orjentisani percentilni prikaz dobitka, koristi se procentualni odnos između percentilnog skora dobitka u $pq\%$ i skora $s(t=1)$ za početnu granu $t=1$:

$$\oplus is(p) = \frac{\Diamond s(p)}{s(t=1)}$$

Ako je imenitelj 0 (imenitelj predstavlja prosek svih medijalnih vrednosti y_n iz uzoraka), moguće je da ova statistika nije dostupna za uzorak za analizu ili za uzorak testa.

4.7.Primena analize klasifikacionog stabla u klasifikaciji opština Makedonije u odnosu na njihove karakteristike

Cilj ove analize je da ispitivanjem klasifikacionog stabla za opštine Makedonije utvrdi pripadnost opština jednoj od tri kategorije zavisne promenljive: najrazvijenije opštine, srednje razvijene opštine i slabo razvijene opštine. Osim zavisne promenljive – *Razvijenost opština*, razmatra se i devet, demografskih i ekonomskih promenljivih: *Prirodni priraštaj, Broj prodavnica u maloprodaji, Broj individualnih poljoprivrednih domaćinstava, Ukupan broj stanovnika, Ukupan broj domaćinstava, Broj obrazovanih stanovnika, Ukupno ekonomski aktivno stanovništvo, Ukupan broj zaposlenih lica, Ukupan broj nezaposlenih lica*.

Podaci su dobijeni iz Državnog zavoda za statistiku i odnose se na pokazatelje o stanovništvu i domaćinstvima, iz popisa 2002 godina. Privredno-ekonomski pokazatelji datih opština odnose se na 2008 godinu. Baza podataka je ista kao i kod hijerahiske i k –sredina klaster analize. Posmatraju se 82 opštine.

Korišćena baza podatka ispunjava pretpostavke analize klasifikacijonog stabla jer su korišćene adekvatne merne za analizirane promenljive i za atributivne zavisne promenljive. S obzirom da analiza klasifikacionog stabla ne zahteva da se **nestandardne opservacije** odstranjuju na bazi Mahalanobisovog D^2 rastojanja, sve opštine su obuhvaćene analizom što znači da nema **nedostajućih podataka**

Radi bolje preglednosti, za nezavisne promenljive su uvedene skraćenice: *PP - Prirodni priraštaj*, *PRO - Broj prodavnica u maloprodaji*, *POLJ - Broj individualnih poljoprivrednih domaćinstava*, *STA - Ukupan broj stanovnika, DOM - Ukupan broj domaćinstava, OBR - Broj obrazovanih stanovnika, EKO - Ukupno ekonomski aktivno stanovništvo, ZAP - Ukupan broj zaposlenih lica, NEZ - Ukupan broj nezaposlenih lica.*

Tabela 4.1. sumarnog pregleda modela, obezbeđuje neke informacije o specifičnostima koje su korišćene pri izgradnji ovog modela, kao i o specifičnostima konačnog modela. Prvi deo, specifikacija daje informacije o podešavanjima koja su korišćena kako bi se generisao model klasifikacionog stabla, uključujući i promenljive koje su korištene u analizi. Drugi deo, rezultati prikazuje informacije o broju ukupnih i završnih grana, visini drveta, kao i koje nezavisne promenljive su uključene u završni model.

Tabela 4.1. Sumarni pregled modela

Specifikacija	Algoritam	CHAID
	Zavisna promenljiva	Razvijenost opština
	Nezavisne promenljive	Prirodni priraštaj, Broj prodavnica u maloprodaji, Broj individualnih poljoprivrednih domaćinstava, Ukupan broj stanovnika, Ukupan broj domaćinstava, Broj obrazovanih stanovnika, Ukupno ekonomski aktivno stanovništvo, Ukupan broj zaposlenih lica, Ukupan broj nezaposlenih lica
	Validacija	
	Maksimalna dubina stabla	3
	Minimum opservacija u roditelj-grani	10
	Minimum opservacija u dete-grani	5
Rezultati	Uključene nezavisne promenljive	Broj obrazovanih stanovnika, Broj individualnih poljoprivrednih domaćinstava, Ukupan broj zaposlenih lica, Ukupan broj stanovnika
	Broj grana	12
	Broj krajnih grana	7
	Dubina	3

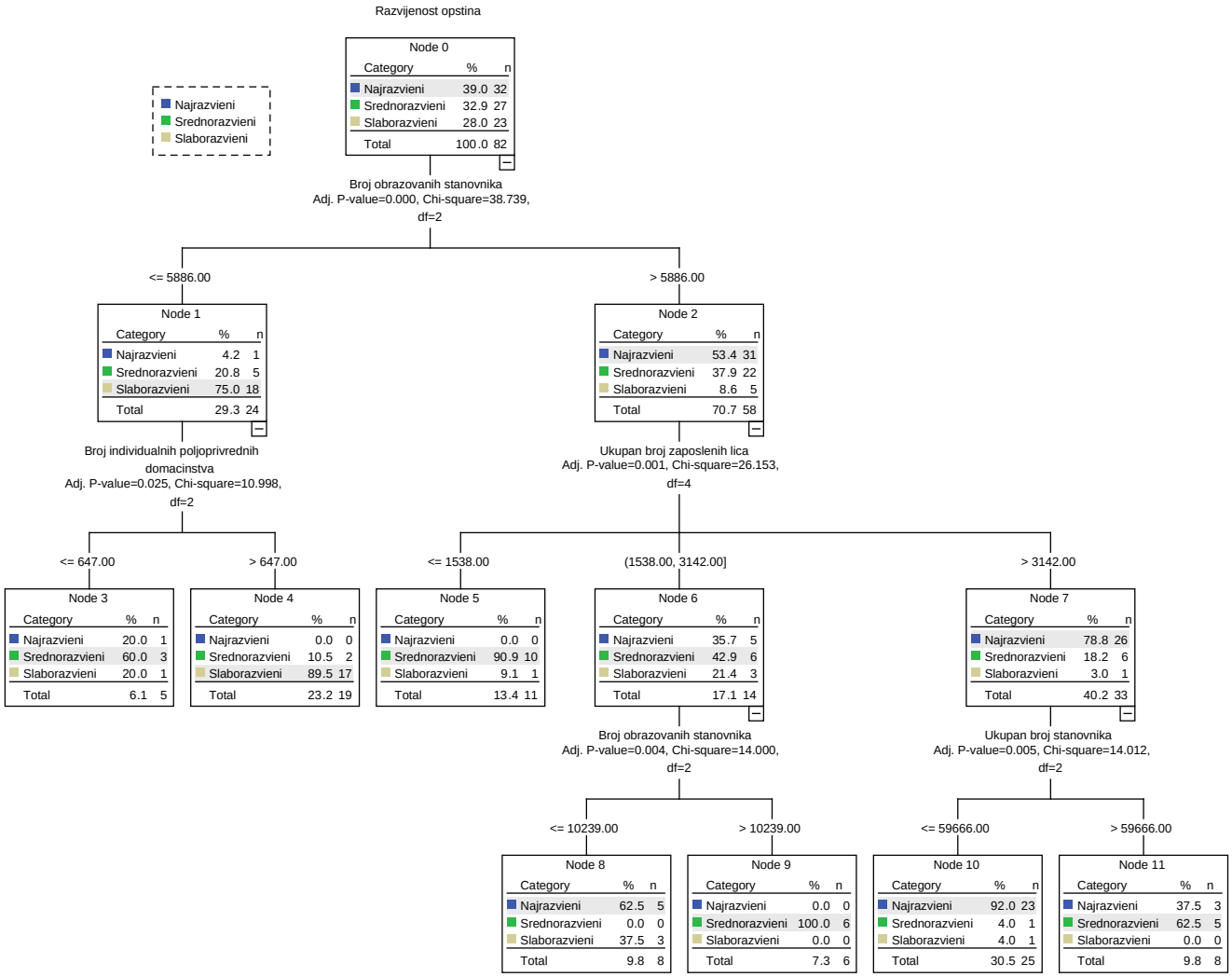
Izvor: Rezultati dobijeni primenom SPSS – a

Od početnih devet nezavisnih promenljivih, samo četiri su uključene u završni model. Ostale promenljive: *Prirodni priraštaj, Bbroj prodavnica u maloprodaji, Ukupan broj domaćinstava, Ukupno ekonomski aktivno stanovništvo, Ukupan broj nezaposlenih* lica ne doprinose značajano izgradnji modela, tako da su one automatski isključene iz konačnog modela.

Šema klasifikacionog stabla pokazuje da koristeći CHAID metod, možemo odrediti da je nezavisna promenljiva *Broj obrazovanih stanovnika* najbolji procenjivač zavisne promenljive *Razvijenost opština*. Ova nezavisna promenljiva deli opštine na dve grane. U prvoj granu su slaborazvijene opštine koje imaju manje od 5886 obrazovanih stanovnika – grana 1, dok u drugoj grani – grana 2 su najrazvijene opštine koje imaju više od 5886 obrazovanih stanovnika.

Za granu u koju spadaju slaborazvijene opštine, osim *Broja obrazovanih stanovnika*, i *Broj individualnih poljoprivrednih domaćinstava* je značajan procenjivač razvijenosti opština. Ova nezavisna promenljiva deli kategoriju na dve grane, i to granu – 3 i granu 4. U grani četiri od ukupno 19 opština, 89,5% ili 17 opština su slaborazvijene. Pošto nema daljeg grananja ove grane, ona se smatra završnom granom.

Slika 4.1. Klasifikaciono stablo



Izvor: Rezultati dobijeni primenom SPSS – a

Za kategorije srednjerazvijene i najrazvijenije opštine, sledeći najbolji procenjivač je nezavisna promenljiva *ukupan broj zaposlenih lica*.

Za srednjerazvijene opštine, model uključuje još jedan procenjivač, a to je nezavisna promenljiva *broj obrazovanih stanovnika*.

Za Najrazvijenije razvijene opštine, nezavisna promenljiva *ukupan broj stanovnika* je još jedan procenjivač. Po ovoj nezavisnoj promenljivoj, 92% ili 23 opštine od ukupno 25 opština iz ove kategorije su najrazvijenije opštine.

U Tabeli 4.2. sadržane su većinu bitne informacije za šemu klasifikacionog stabla. Za svaku granu u tabeli prikazan je ukupan broj i procenat opština u svakoj kategoriji

zavisne promenljive. Takođe, prikazanja je i predviđena kategorija za zavisnu promenljivu. Kao i grana za svaku podgranu na drvetu. Nezavisne promenljive su upotrebljene za dalje grananje grana na podgrane.

Tabela 4.2. Informacije o klasifikacionom stablu

Grana		Najrazvijenije opštine		Srednjerazvijene opštine		Slaborazvijene opštine		Ukupno		Predviđena kategorija	Grana - roditelj
		N	%	N	%	N	%	N	%		
	32	32	39,0%	27	32,9%	23	28,0%	82	100,0%	Najrazvijenije	
0	1	1	4,2%	5	20,8%	18	75,0%	24	29,3%	Slaborazvieni	0
1	31	31	53,4%	22	37,9%	5	8,6%	58	70,7%	Najrazvijenije	0
2	1	1	20,0%	3	60,0%	1	20,0%	5	6,1%	Srednjerazvijene	1
3	0	0	0,0%	2	10,5%	17	89,5%	19	23,2%	Slaborazvijene	1
4	0	0	0,0%	10	90,9%	1	9,1%	11	13,4%	Srednjerazvijene	2
5	5	5	35,7%	6	42,9%	3	21,4%	14	17,1%	Srednjerazvijene	2
6	26	26	78,8%	6	18,2%	1	3,0%	33	40,2%	Najrazvijenije	2
7	5	5	62,5%	0	0,0%	3	37,5%	8	9,8%	Najrazvijenije	6
8	0	0	0,0%	6	100,0%	0	,0%	6	7,3%	Srednjerazvijene	6
9	23	23	92,0%	1	4,0%	1	4,0%	25	30,5%	Najrazvijenije	7
10	3	3	37,5%	5	62,5%	0	,0%	8	9,8%	Srednjerazvijene	7

Algoritam: CHAID
Zavisna promenljiva Razvijenost opština
Izvor: Rezultati dobijeni primenom SPSS – a

S obzirom da je klasifikaciono stablo kreirano CHAID metodom, stepeni slobode i nivoi značajnosti (p –vrednost), χ^2 statistike bitni su za dalje grananje (Tabela 4.3.).

Tabela 4.3. χ^2 statistike klasifikacionog stabla

Grana	**Primarna nezavisna promenljiva**				
	Promenljiva	**p –vrednost (a)**	**χ^2**	**Stepeni slobode**	**Vrednost delenja**
0	OBR	0,000	38,739	2	<= 5886
1	OBR	0,000	38,739	2	> 5886
2	POLJ	0,025	10,998	2	<= 647
3	POLJ	0,025	10,998	2	> 647
4	ZAP	0,001	26,153	4	<= 1538
5	ZAP	0,001	26,153	4	(1538, 3142]
6	ZAP	0,001	26,153	4	> 3142
7	OBR	0,004	14,000	2	<= 10239
8	OBR	0,004	14,000	2	> 10239
9	STA	0,005	14,012	2	<= 59666
10	STA	0,005	14,012	2	> 59666

Algoritam: CHAID
Zavisna promenljiva Razvijensot opština
(a) Bonferroni prilagođenje
Izvor: Rezultati dobijeni primenom SPSS – a

U ovom modelu za grananja bitan je nivo niži od 0,0005. Kao rezultat primenjenog modela dobijene su vrednosti nezavisne promenljive za datu granu.

(Napomena: Za ordinalne i kvantitativne nezavisne promenljive može se videti (i kod šeme, kao i kod tabele klasifikacionog stabla), intervala izražene u obliku vrednost 1, vrednost 2, šta u stvari znači "veći od vrednosti 1 i manji od ili jednak vrednosti 2."). U ovom primeru, takva je promenljiva *ukupan broj zaposlenih lica.*

Tabela 4.4. Statistika dobitka za grane za najrazvijenije opštine

Grana	Grana		Dobitak		Odziv	Indeks
	N	%	N	%	N	%
10	25	30,5%	23	71,9%	92,0%	235,8%
8	8	9,8%	5	15,6%	62,5%	160,2%
11	8	9,8%	3	9,4%	37,5%	96,1%
3	5	6,1%	1	3,1%	20,0%	51,3%
4	19	23,2%	0	0,0%	0,0%	0,0%
5	11	13,4%	0	0,0%	0,0%	0,0%
9	6	7,3%	0	0,0%	0,0%	0,0%

Algoritam: CHAID
Zavisna promenljiva: Razvijenost opština
Izvor: Rezultati dobijeni primenom SPSS – a

Tabela 4.4., Tabela 4.5. i Tabela 4.6. sadrže informacije samo o završnim granama u modelu za posmatrane grupe opština, jer one predstavljaju najbolju ocenu klasifikacije.

Tabela 4.5. Statistika dobitka za grane za srednjerazvijene opštine

Grana	Grana		Dobitak		Odziv	Indeks
	N	%	N	%	N	%
9	6	7,3%	6	22,2%	100,0%	303,7%
5	11	13,4%	10	37,0%	90,9%	276,1%
11	8	9,8%	5	18,5%	62,5%	189,8%
3	5	6,1%	3	11,1%	60,0%	182,2%
4	19	23,2%	2	7,4%	10,5%	32,0%
10	25	30,5%	1	3,7%	4,0%	12,1%
8	8	9,8%	0	0,0%	0,0%	0,0%

Algoritam: CHAID
Zavisna promenljiva: Razvijenost opština
Izvor: Rezultati dobijeni primenom SPSS – a

Budući da vrednost statistike dobitka daje informaciju o ciljnim kategorijama, ove tabele se koriste samo ukoliko je određena jedna ili više ciljnih kategorija. U ovom

primeru postoje tri ciljne kategorije, tako da ovde postoje tri statistike dobitaka za grane.

Veličina "N za granu" je broj opština u svakoj završnoj grani, a veličina "%" predstavlja procentualni iznos ukupnog broja opština u svakoj grani.

Veličina "N kod statistike dobitka" predstavlja broj opština u svakoj završnoj grani u okviru ciljne kategorije; a veličina "%" predstavlja procenat opština u okviru ciljne kategorije koja uzima u obzir ukupni broj opština u ciljnoj kategoriji (u ovom primeru, imamo tri ciljne kategorije).

Tabela 4.6. Statistika dobitka za grane za slaborazvijene opštine

Grana	Grana		Dobitak		Odziv	Indeks
	N	%	N	%	N	%
4	19	23,2%	17	73,9%	89,5%	319,0%
8	8	9,8%	3	13,0%	37,5%	133,7%
3	5	6,1%	1	4,3%	20,0%	71,3%
5	11	13,4%	1	4,3%	9,1%	32,4%
10	25	30,5%	1	4,3%	4,0%	14,3%
11	8	9,8%	0	0,0%	0,0%	0,0%
9	6	7,3%	0	0,0%	0,0%	0,0%

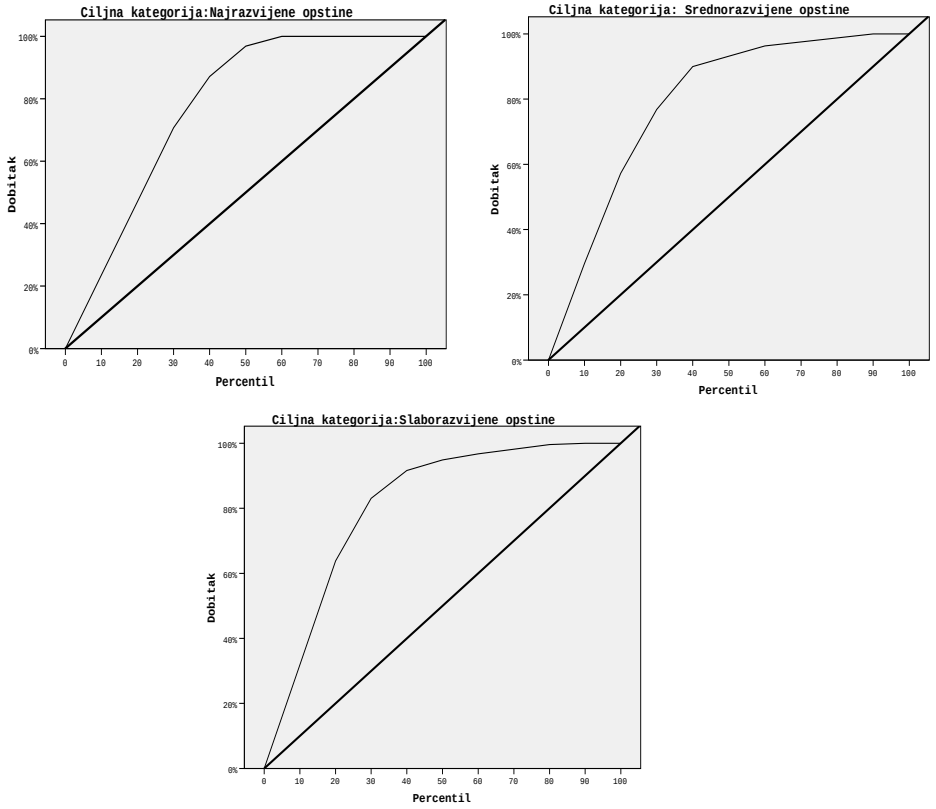
Algoritam: CHAID
Zavisna promenljiva: Razvijenost opština
Izvor: Rezultati dobijeni primenom SPSS – a

Kod kategorijskih zavisnih promenljivih: veličina Odziv predstavlja procenat opština u grani u okviru ciljne kategorije, veličina Indeks predstavlja odnos veličine Odziv i veličine % statistike dobitka, odnosno predstavlja količnik između procenta opština u grani u okviru ciljne kategorije i procenta opština u okviru ciljne kategorije koji uzima u obzir ukupni broj opština u ciljnoj kategoriji za ceo uzorak.

Vrednost statistike Indeksa je u stvari pokazatelj koliko se procenat posmatrane ciljne kategorije za tu granu razlikuje od očekivanog procenta za ciljnu kategoriju u celom uzorku. Procenat ciljne kategorije u nultoj grani (a to je u stvari stablo) predstavlja očekivani procenat pre nego što su uzeti u obzir efekti bilo koje od nezavisnih promenljivih.

Vrednost Indeksa veća od 100% znači da je veće učešće opština u ciljnoj kategoriji od njihovog učešća u ciljnoj kategoriji za ceo uzorak. Nasuprot tome, vrednost Indeksa manja od 100% znači da je manje učešće opština u ciljnoj kategoriji nego u ciljnoj kategoriji za ceo uzorak.

Slika 4.2. Statistike dobitka za ciljne grupe opština



Izvor: Rezultati dobijeni primenom SPSS – a

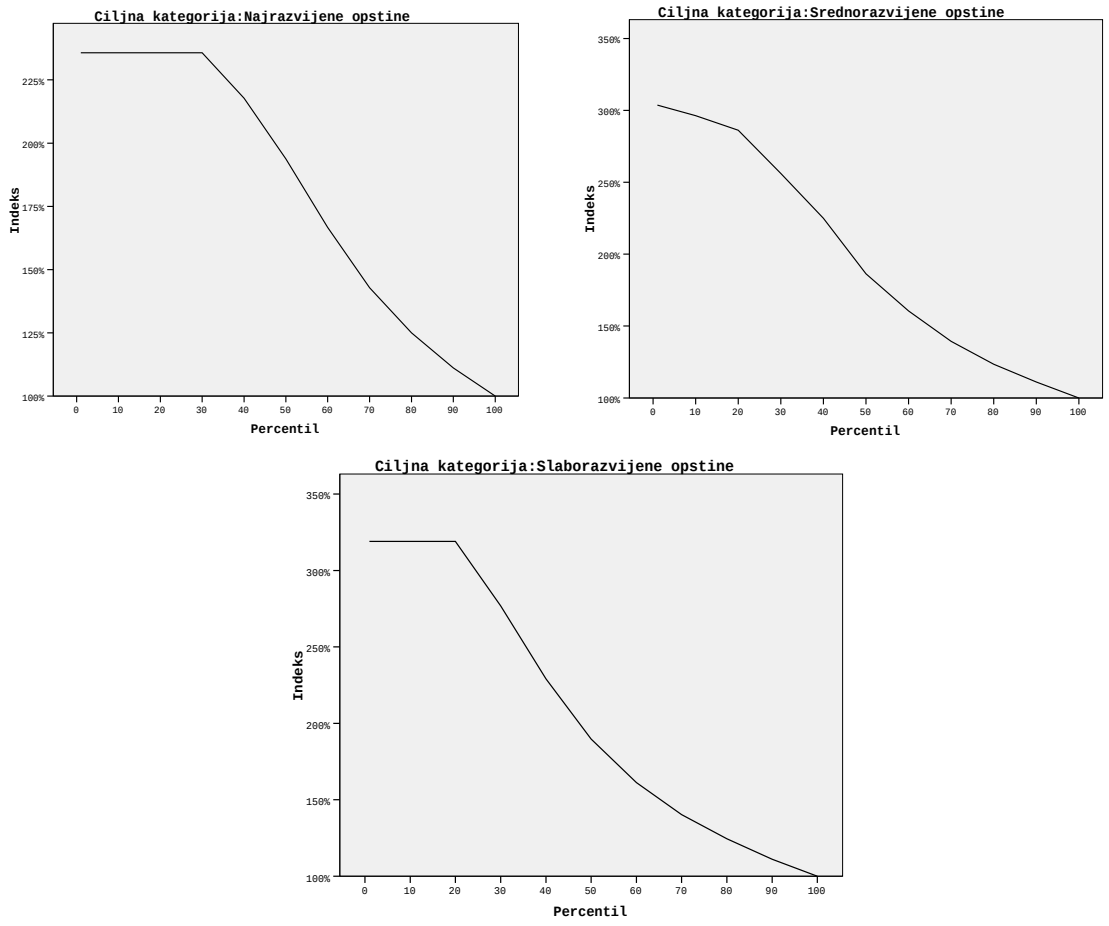
Dijagrami statistike dobitka (Slika 4.2.) ukazuju da je model prilično dobar za sve tri kategorije zavisne promenljive. Dijagram akumuliranih dobitaka uvek započinje od 0% i završava se sa 100% kako se krećete od jednog do drugog kraja horizontalne ose. Za dobar model, dijagram statistike dobitka počinje naglo da raste prema 100%, zatim se rast stabilizuje, i na kraju rast počinje da usporava što ima za posledicu da nagib funkcije počinje da opada. Model koji ne daje informacije slediće pravu, koja je dijagonala prvog kvadranta.

Dijagrami statistika indeksa (Slika 4.3.) takođe ukazuje da je model dobar i to za sve tri kategorije opština. Dijagram akumuliranih statistika indeksa nastoji da otpočne od nivoa znatno višim od 100% i da posle određenog konstantnog nivoa počinje da opada dok ne dostigne nivo 100%.

Kod dobrog modela, vrednost statistike indeksa treba da počne mnogo iznad nivoa 100%, zatim da ostane na tom nivou kako se kreće duž horizontalne ose, a zatim se naglo spušta prema nivou od 100%.

Ukoliko model ne odgovara, funkcija će oscilirati oko nivoa 100%.

Slika 4.3. Statistike indeksa po ciljnim grupama



Izvor: Rezultati dobijeni primenom SPSS – a

Tabela 4.7. Ocena rizika

Ocena	Standardna greška
0,159	0,040

Algoritam: CHAID
Zavisna promenljiva: Razvijenost opština
Izvor: Rezultati dobijeni primenom SPSS – a

Ocena rizika (Tabela 4.7.) omogućava brzu ocenu kvaliteta rada modela. Ocena rizika od 0,159 pokazuje da su kategorije koje je procenio model (najrazvijene, srednjerazvijene i slaborazvijene opštine) pogrešne u 15,9% slučajeva. Tako je "rizik" od pogrešnog klasifikovanja opštine približno 16%.

Rezultati klasifikacije modela (Tabela 4.8) su u skladu sa procenom rizika. Model je ispravno klasifikovao približno 84,1% opština. Najloše su klasifikovanje

slaborazvijene opštine. Ispravno je razvrstano 73,9% slaborazvijenih opština. Ostale dve kategorije imaju skoro isti procent dobro kalsifikovanih opština.

Tabela 4.8. Klasifikacija modela

Originalne kategorije	Predviđene kategorije			
	Najrazvijene opštine	Srednjerazvijene opštine	Slaborazvijene opštine	Procent tačnih vrednosti
Najrazvijene opštine	28	4	0	87,5%
Srednjerazvijene opštine	1	24	2	88,9%
Slaborazvijene opštine	4	2	17	73,9%
Ukupan procent	40,2%	36,6%	23,2%	84,1%

Algoritam: CHAID
Zavisna promenljiva: Razvijenost opština
Izvor: Rezultati dobijeni primenom SPSS – a

Predviđene vrednosti. Na osnovu aktivne baze podataka stvorene su četiri nove promenljive: prva prikazuje završnu granu za svaku opštinu, druga prikazuje predviđenu vrednost zavisne promenljive za svaku opštinu, treća prikazuje verovatnoću da opština pripada datoj kategoriji zavisne promenljive. S obzirom da postoje tri moguće kategorije za zavisnu promenljivu, formiraju se tri promenljive i to:

1) verovatnoća da opština pripada kategoriji najrazvijene opštine,

2) verovatnoća da opština pripada kategoriji srednjerazvijene opštine,i

3) verovatnoća da opština pripada kategoriji najmanjerazvijene opštine.

Četvrta promenljiva je predviđena verovatnoća koja je jednostavno proporcija opština svake kategorije zavisne promenljive u završnim granama.

U Tabeli 4.9. je prikazano kojoj kategoriji razvijenosti pripada opština na osnovu analize klasifikacionog stabla. 13 opština od ukupno 82 opštine je pogrešno klasifikovano. Opštine koje se vode kao najrazvijenije: Gazi Baba, Bitola, Dojran i Prilep, po analizi, pripadaju u grupi srednjerazvijenih opština. Srednjerazvijena opština Kičevo po analizi bi trebalo da bude u grupi najrazvijenijih opština, dok srednjerazvijene opštine Sopište i Centar Župa bi trebale da budu u grupi slaborazvijenih opština.

Iz grupe slaborazvijenih opština: Berovo, Vasilevo, Kratovo i Češinovo, treba svrstati u grupu najrazvijenih opština, a opštine Vranešnica i Dolneni u grupu srednjerazvijenih opština.

Tabela 4.9. Predviđena kategorija zavisne promenljive za svaku opštinu

Redni broj	Opština	Orig. kategorije	Predviđene kategorije	Redni broj	Opština	Orig. kategorije	Predviđene kategorije
1	Butel	1	1	42	Kičevo	2	1
2	Gazi Baba	1	2	43	Konče	3	3
3	Gorče Petrov	1	1	44	Kočani	1	1
4	Karpoš	1	1	45	Kratovo	3	1
5	Kisela Voda	1	1	46	Kriva Palanka	1	1
6	Saraj	2	2	47	Krivogaštani	3	3
7	Čair	2	2	48	Kruševo	1	1
8	Šuto Orizari	2	2	49	Kumanovo	2	2
9	Aračinovo	2	2	50	Lipkovo	2	2
10	Berovo	3	1	51	Lozovo	3	3
11	Bitola	1	2	52	Mavr. i Rostuša	2	2
12	Bogdanci	1	1	53	Mak. Kamenica	1	1
13	Bogovinje	2	2	54	Mak. Brod	3	3
14	Bosilovo	1	1	55	Mogila	3	3
15	Brvenica	2	2	56	Negotino	1	1
16	Valandovo	1	1	57	Novaci	3	3
17	Vasilevo	3	1	58	Novo Selo	1	1
18	Vevcani	2	2	59	Oslomej	2	2
19	Veles	1	1	60	Ohrid	1	1
20	Vinica	1	1	61	Petrovec	2	2
21	Vranešnica	3	2	62	Pehčevo	3	3
22	Vrapčište	2	2	63	Plašnica	2	2
23	Gevgelija	1	1	64	Prilep	1	2
24	Gostivar	2	2	65	Probištip	1	1
25	Gradsko	3	3	66	Radoviš	1	1
26	Debar	2	2	67	Rankovce	3	3
27	Debarca	3	3	68	Resen	1	1
28	Delčevo	1	1	69	Rosoman	3	3
29	Demir Kapija	3	3	70	Sveti Nikole	1	1
30	Demir Hisar	1	1	71	Sopište	2	3
31	Dojran	1	2	72	Staro Nagorican.	3	3
32	Dolneni	3	2	73	Struga	2	2
33	Drugovo	3	3	74	Strumica	1	1
34	Želino	2	2	75	Studeničani	2	2
35	Zajas	2	2	76	Tearce	2	2
36	Zelenikovo	2	2	77	Tetovo	2	2
37	Zrnovci	3	3	78	Centar Župa	2	3
38	Ilinden	1	1	79	Časka	3	3
39	Jegunovce	2	2	80	Česinovo	3	1
40	Kavadarci	1	1	81	Čučer - Sandovo	1	1
41	Karbinci	3	3	82	Štip	1	1

Izvor: Rezultati dobijeni primenom SPSS – a

5.Zaključak

Osnovni cilj ovog rada je da detaljno prikaže i obrazloži primenu multivarijacionih tehnika klasifikacije: klaster analizu, diskriminacionu analizu i analize klasifikacionog stabla na konkretnim bazama podataka pri rešavanju različitih emprijskih ekonomskih problema.

Prvo poglavlje rada se odnosi na multivarijacionu tehniku **klaster analize**. Osnovna ideja klaster analize je da se izvrši grupisanje opservacija ili objekata u klastere, tako da su objekti u istom klasteru sličniji međusobno nego objektima koji se nalaze u ostalim klasterima. Glava je koncipirana u dva dela, gde se prvi deo odnosi na teoretsko izlaganje modela, dok se drugi deo odnosi na primenjenu klaster analizu i to za dve različite baze podataka.

U **teoretskom delu** dat je prikaz istraživačkog dizajna klaster analize, gde je posebna pažnja posvećena pripremi podataka za klaster analizu, naročito pitanju nestandardnih opservacija i standardizaciji promenljivih. U ovom delu rada su navedene pretpostavke klaster analize, kriterijumi za određivanje broja klastera, interpretacija, validacija i profiliranje klastera. Pravilna primena koraka kompletne klaster analize, doprinosi tačnosti i širokoj aplikativnosti dobijenih rezultata. Zato je posebna pažnja posvećena algoritmima klaster analize u delu merama bliskosti i metodama klaster analize, pri čemu gde su detaljno opisani mehanizmi na kojima je zasnovana ova klasifikaciona procedura. U delu koji se odnosi na **primenu klaster analize** prikazane su tri osnovne tehnike: hijerahijska klaster analiza, k - sredina klaster analiza i dvostepena klaster analiza.

Hijerahijska i k - sredina klaster analiza su primenjene za klasifikaciju opština Makedonije na osnovu demografskih i ekomiskih karakteristika. Dvostepena klaster analiza, koja daje bolje rezultate za veće baze podataka, primenjena je za definisanje klastera od 200 makedonskih kompanija na osnovu izabranih karakteristika.

Sledeće poglavlje posvećeno je **diskriminacionoj analizi,** čiji je osnovni cilj da oceni vezu između jedne kategorijski zavisne promenljive i skupa metričkih objašnjavajućih promenljivih. Višestruka diskriminaciona analiza ima široku primenu u situacijama gde je primarni cilj da se identifikuje grupa kojoj pripada objekat ili opservacija. Objekti se klasifikuju grupe, a cilj je da se preko nezavisnih promenljivih predvide i objasne razlozi zašto se objekat nalazi u određenoj grupi, koju je istraživač odabrao.

Koncept ove glave je sličan kako i kod klaster analize, odnosno prvo je dat **teorijski deo** u kome su prikazane osnove istraživačkog dizajna, pretpostavke diskriminacione analize, interpretacija i validacija rezultata kompletnog procesa. U posebnom delu detaljno su objašnjeni algoritmi ove analize, koja se može koristiti za diskriminaciju i klasifikaciju elementa dve grupe ili više grupa (populacija). Nakon što su objašnjeni algoritmi, poseban deo je posvećen i Fišerovom pristupu klasifikaciji. U empriskom delu ovaj statistički metod je primenjen diskriminaciju i klasifikaciju podataka iz dve baze. **Diskriminaciona analiza** je primenjena diskriminaciju (razdvajanje) zemalja članica Europske Unije i zemalja koje nisu članice Evropske Unije na osnovu višedimenzionalnog kriterijuma, a zatim za klasifikaciju opština Makedonije u odnosu na izabrane karakteristike.

Treća klasifikaciona procedura multivarijacione analize **analiza klasifikacionog stabla**, kreira stabla klasifikovanja i odlučivanja, na osnovu kojih se vrši bolja identifikacija grupa, otkrivaju veze između grupa, i predviđaju buduća dešavanja. Kao i kod prethodnih klasifikacionih procedura, poglavlje posvećeno ovoj analize je podeljeno na **teoretski deo** i deo **primene analize** u rešavanju realnih ekonomskih problema. U teoretskom delu objašnjen je kompletan proces analize, koji obuhvata istraživački dizajn i pretpostavke analize klasifikacionog stabla. Algoritmi koji su od posebne važnosti su detaljno prikazani preko metoda rasta klasifikacionog stabla, odnosno CHAID i iscrpnog CHAID algoritma, CART algoritma i QUEST algoritma. Takođe, prikazan je i proces dodeljivanja i ocena rizika kako i pregled dobitka. Empriski deo je posvećen klasifikaciji opštine Makedonije preko klasifikacionog stabla, a za čije kreiranje je korišćen CHAID metod.

Na početku rada izloženo je šest hipoteza koje se odnose na multivarijacione metode za klasifikacije i na njihovu aplikaciju za rešavanju realne ekonomskih problema.

Nakon izlaganja teorijskih osnova i primene klasifikacionih metoda u radu, može se zaključiti da su **multivarijacione tehnike za klasifikaciju korisan alat za rešavanje empirijskih problema. One generišu značajne zaključke pa su bitno sredstvo kvantitativne analize**. Zaključak potvrđuje i činjenica da je u ovom radu primenom klasifikacijone procedure uspešno rešena šest problema (3 preko klaster analize, 2 preko diskriminacione analize i 1 preko analize klasifikicionog stabla) i za svaki problem je dobijeno rešenje koje ima suštinsku i praktičnu vrednost.

Multivarijacione tehnike za klasifikaciju su korisne tehnike za rad sa velikim bazama. Ovo potvrđuje i to što su u radu korišćene tri baze podataka. Prva baza obuhvatila je 84 opštine u Makedoniji za koje je analizirano deset ekonomskih i demografskih promenljivih. Druga baza se odnosila na 200 najuspešnijih makedonskih kompanija i za analizu koristila šest promenljivih. Treću bazu činila su zemlje članice Evropske Unije i zemlje koje nisu članice Evropske Unije. Za 48 zemalja korišćeno je sedam promenljivih u analizi. Korišćenjem sve tri baze podatka uspešno su sprovedeni postupci multivarijacione tehnike klasifikacije (nad nekim bazama gde je to bilo potrebno, sprovedena je transformacija promenljivih). Dobijeni su rezultati koji imaju praktičan značaj jer se mogu koristiti za donošenje zaključaka potrebnih da bi se donele i unapredile strategije i da se pomogne u rešavanju konkretnih problema. Ovo potvrđuje da su sve tri tehnike korisne za rad sa bazom podataka.

Klaster analiza je pokazala da je moguće izvršiti klasifikaciju 84 opštine Makedonije po njihovim karakteristikama, i to korišćenjem hijerahijske klaster analize i k-sredina klaster analize. Obe tehnike su se pokazale uspešnim u analizi, dok su dobijeni centroidi klastera opisali klastere preko najvažnih promenljivih, tako da su otkrivene osnovne karakteristike najrazvijenih, srednjerazvijenih i slaborazvijenih opština. Dobijeni zaključci mogu naći primenu u kreiranju razvojne ekonomske politike zemlje.

Za situacije kada je baza podataka velika, preporučuje se dvostepena klaster analiza. U primeru koji je obuhvatio 200 makedonskih kompanija, klaster analiza se pokazala jako korisnom, jer prikazaje četiri grupe koje formiraju osnovnu strukturu kompanija, gde se kao posebni kriterijum izdvajaju veličina kompanije i oblasti. Ovo potvrđuje uspešnu aplikaciju dvostepene klaster analize i ispunjenje prethodno formulisane hipoteze.

Diskriminacijona analiza uspešno je ocenila vezu između kategorijske zavisne promenljive (*privredna razvijenost opština*) i skupa kvantitativnih objašnjavajućih promenljivih makedonskih opština. Takođe, ona je uspešno identifikovala grupu kojoj bi pripadala svaka opština. Diskriminaciona analiza je uspešno klasifikovala zemlje članice Evropske unije i zemlje koje nisu članice Evropske unije preko njihovih osnovnih ekonomskih i demografskih indikatora.

Dobijeni zaključci su aplikativni, odnosno, u prvom slučaju pomažu u kreiranju ekonomske politike razvoja i grupisanja opština (različita politika za različitu grupu), a u drugom slučaju ukazuju na potencijalne članice Evropske Unije, kao i na osnovne promenljive koje su od posebne važnosti za članstvo. Hipoteza za primenu diskriminacione analize je ispunjena.

Poslednja hipoteza odnosi se na poslednju klasifikacionu proceduru - analizu klasifikacionog stabla. Analizirano je 84 makedonskih opština, da bi se kreiralo stablo klasifikovanja i da bi se bolje identifikovale grupe opština po stepenu njihove razvijenosti. Dobijeni rezultati mogu se uspešno primeniti u kreiranju strategije regionalnog razvoja, kao i u identifikaciji ključnih promenljivih ekonomskog razvija opština. Na ovaj način je potvrđeno ispunjenje poslednje hipoteze.

Osim dobijenih konkretnih rezultata, ovaj rad je stvo**rio i dobru analitičku osnovu za slična buduća istraživanja u oblasti klasifikacione procedure i njihove praktične primene**. Rad predstavlja alternativu za analizu kompleksne baze podataka i dobijanja korisnih i kvaltetnih rezultata koji bi unapredili proces donošenja odluka. Za sveobuhvatnu, brzu i uspešnu primenu klasifikacionih procedura od posebne koristi za analizu je bio i **statistički softer SPSS**.

Doprinos ovog rada je bolje razumevanje različitih multivarijacionih tehnika klasifikacije, njihovo prikazavanje na jednostavan način uz prakičnu ilustraciju. Multivarijaciona analiza ima svoju široku primenu kako u mikroekonomskim istraživanjima (marketing, menadžment, organizacijsko ponašanje) tako i u makroekonomskim istraživanjima, istraživanjima javnog mnenja, psihologiji, medicini, sociologiji, poljoprivredi.

Pored teorijskih postavki, rad nastoji da prikaže funkcionisanje metoda, odnosno da objasni proces analize i dobijene rezultate. Multivarijaciona analiza je jedna od mlađih grana statistike, tako da ovaj rad preko prikazane tehnike i praktične aplikacije treba da doprinese popularizaciji ove analize.

Konkretni empiriski rezultati u ovom radu daju značajne informacije u rešavanju različitih praktičnih problema, konkretno, probleme klasifikacija opština u Makedoniji u odnosu na različite promenljive, klasifikaciju makedonskih kompanija, kao i diskriminaciju i klasifikaciju evropskih zemalja. Takođe, korišćenje ovih metoda prikazuje samo jednu od mnogih mogućih primena ovih tehnika, pa se nadamo da ovaj rad može da posluži i kao podsticaj za neka druga buduća statistička istraživanja.

Korišćena literatura

1. Abonyi, János& Feil, Balázs. (2007) *Cluster analysis for data mining and system identification*. Boston and Basel, Switzerland: Birkhäuser Basel.
2. Aldenderfer, M. S., and R. K. Blashfield. (1984) *Cluster analysis*. Thousand Oaks, CA: Sage.
3. Anderberg, M. (1973) *Cluster Analysis for Applications*. New York: Academic Press.
4. Anderson, T. E. (2003) *An introduction to Multivariate Statistical Analysis* (3rd edition). New York: John Wiley.
5. Banićević, D., Vasić, V. (2006) "Business bank's discriminatory analysis: conservative or aggressive credit policy". Zlatibor: *SymOgr 2006*.
6. Berry, M. J. A., and G. Linoff. (2004) *Data Mining Techniques: For Marketing, Sales and Customer Relationship Management* (2nd edition). New York: John Wiley.
7. Berthold, M., and D. J. Hand (2003) *Intelligent Data Analysis* (2nd edition). Berlin, Germany: Springer – Verlag.
8. Bigss, D., Ville, B., and Suen, E. (1991) A Method of Choosing Multiway Partitions for Classification and Decision Trees. *i*, 18, 1, 49 – 62.
9. Breiman, L., Friedman, J. H., Olshen, R., Stone, C. J. (1984) *Classification and Regression Tree*. Wadsworth & Broks/Cole Advanced Books & Software, Pacific California.
10. Everitt, B. S., S. Landau and M. Leese. (2001) *Cluster Analysis* (4th edition) London: Hodder Arnold.
11. Everitt, B. S., & Rabe-Hesketh, S. (1997) *The analysis of proximity data*. London: Arnold.
12. Fisher, R. A. (1936) "The Use of Multiple Measurments in Taxonomic Problems". *Annals of Eugenics*, 7, 179 – 188.
13. Fisher, R. A. (1938) "The statistical Utilization of Multiple Measurments". *Annals of Eugenics,* 8, 376 – 386.

14. Garson, D., (2009) *Cluster analysis from Statnotes: Topics in Multivariate analysis*, retrieved from http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm.
15. George H. D. (1984) *Introduction to multivariate analysis*. Thousand Oaks, CA: Sage Publications.
16. Gnanadesikan, R. (1997) *Methods for Statistical Data Analysis of Multivariate Observations* (2nd edition). New York: Wiley – Interscience.
17. Goodman, L. A. Simple Models for the Analysis of Association in Cross – Classifications Having Ordered Categories (1979) *Journal of the American Statistical Association*, 74, 537 – 552.
18. Green, P. E. (1978) *Analyzing Multivariate Data* Hinsdale, IL: Holt, Reinehart and Winston.
19. Hair, J.F., W.C. Black, B.J. Babin, R.E. Anderson and R.L. Tatham. (2006) *Multivariate Data Analysis* (6th edition). Upper Saddle River, NJ: Prentice Hall.
20. Harris, R. J. (2001) *A Primer of Multivariate Statistics* (3rd edition). Hillsdale, NJ: Lawrence Erlbaum Associates.
21. Hartigan, J. A. (1975) *Clustering Algorithms*. New York: John Wiley.
22. Hartigan, J. A. (1985) "Statistical Theory in Clustering". *Journal of Classification* 2: 63 – 76.
23. Hilld, M. (1996) "Allocation Rules and Their Error Rates". *Journal of the Royal Statistical Society* (B), 28, 1 – 31.
24. Huberty, Carl J. (1994) *Applied discriminant analysis*. NY: Wiley-Interscience. (Wiley Series in Probability and Statistics).
25. Huberty, C. J., S. Olejnik. (2006) *Applied Manova and Discriminant analysis* (2nd edition). A John Wiley & Sons, Inc. Publications.
26. Jain, A. K., and R. C. Dubes. (1988) *Algorithms for Clustering Data*. Upper Saddle River, NJ: Prentice Hall.
27. Jajuga, Krzystof; Sokolowski; Andrzej; & Bock, Hans-Hermann. (2002) *Classification, clustering and data analysis*. Y: Springer.
28. Jardine, N., and R. Sibson. (1975) *Mathematical Taxonomy*. New York: Wiley.

29. Johnsong, N., and D. Wichern. (2002) *Applied Multivariate Statistical Analysis* (5th edition). Upper Saddle River, NJ: Prentice Hall.
30. Kass, G.V. (1980) An Exploratory Technique for Investigating Large Quantites of Categorical Data. *Applied Statistics*, 20, 2, 119 – 127.
31. Kaufman, Leonard & Rousseeuw Peter J. (2005) *Finding groups in data: An introduction to cluster analysis.* NY: Wiley-Interscience.
32. Kendall, M. G. (1975) *Multivariate Analysis*. New York: Hafner Press.
33. Klecka, William R. (1980) *Discriminant analysis. Quantitative Applications in the Social Sciences Series*, No. 19. Thousand Oaks, CA: Sage Publications.
34. Kovačić, Z. (1994) *Multivarijaciona analiza*. Ekonomski fakultet – Beograd.
35. Krstić B., Lakić N., Janković R., Radosavljević Jovanka (1992) "Evaluation of economic achievement of large dairy farms with the application of Discriminant analysis", *Review of Reseach Work at the Faculty of Agriculture, Vol. 37, No. 2*, 31 – 37, Belgrade.
36. Lachenbruch, P. A. (1975) *Discriminant analysis*. NY: Hafner.
37. Lakić N., Maletić R. (1998) "Degree of Separability of Singled out Clusters Based on the Indicator Production Conditions", *Review of Reseach Work at the Faculty of Agriculture, Vol. 43, No. 1*, 123 – 131, Belgrade.
38. Lakić N., Maletić R. (1999) "Separability Degree of Clusters Based on the Production Results Indicator", *Review of Reseach Work at the Faculty of Agriculture, Vol. 44, No. 1*, 89 – 97, Belgrade.
39. Lakić N., Stevanović S. (2003) "Ranking of Vojvodina Muncipalities according to multidimensional denominator of livestock production commodities", *Journal of Agricultural Sciences, Vol. 48, No. 2*, 217 – 226, Faculty of Agriculture - Belgrade.
40. Lim, T. S., Loh, W. Y. And Shih, Y. S. (2000) A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty – three Old and New Classification Algorithms. *Machine Learning*, 40.
41. Loh, W. Y. And Shih, Y. S. (1997) Split selection methods for classification trees. *Statistica Sinica*, Vol. 7, 815 – 840.
42. Mardia, K. V., J. T. Kent, and J. M. Bibby. (2003) *Multivariate Analysis*. London: Academic Press.

43. McLachlan, Geoffrey J. (2004) *Discriminant analysis and statistical pattern recognition*. NY: Wiley-Interscience. (Wiley Series in Probability and Statistics).
44. Meyers S.A, Gamst G., Guarno A. J. (2005) *Applied multivariate research*. Sage Publications.
45. Milligan, Glenn W., and Martha C. Cooper. (1985) "An Examination of Procedures for Determening the Number of Clusters in a Data Set". *Psychometrica* 50 (2): 159 – 179.
46. Morrison, D. F. (1976) *Multivariate Statistical Methods.* New York: McGraw-Hill.
47. Murray, G. D. (1997) "A Cautionary Note on Selection of Variables in Discriminant Analysis", *Applied Statistics*, 26, no. 3, 246 – 250.
48. Shepard, R. *Metric Structures in Ordinal Data*. (1996) Journal of Mathematical Psychology 3: 287 – 315.
49. SPSS. (2004) *SPSS 13.0 Command Syntax Reference*. Chicago: SPSS Inc.
50. Stevens, J. (2002) *Applied multivariate statistics for social sciences*. Lawrence Earlbaum Associates.
51. Tabachnick, Barbara G. and Linda S. Fidell. (2001) *Using multivariate statistics*, (4th edition). Boston: Allyn and Bacon.
52. Tabachnick, B. G., L. S. Fidell. (2007) *Using Multivariate Statistics* (5th edition). Pearson Education, Inc.
53. Tinsley EAH, Brown SD. (2000) *Handbook of applied multivariate statistics and mathematical modelling*. Academic Press.
54. Timm, N. H. (2002) *Applied Multivariate Research*. Springer – Verlag New York, Inc.
55. Vasić, V., Banićević, D. (2006) Classification tree analysis to evaluate credit risk. Kopaonik: *YU INFO 2006*, 6, CD: ISBN 86-85525-01-2.
56. Vasić, V., Banićević, D. (2006) "Discriminant analysis algorithm in clients credit risks estimation". Kopaonik: *YU INFO 2006*.
57. Vasić V., Banićević D., Vojvodičan M. (2008) "Two–step cluster analysis algorithm in identification of groups of bank's clients". Kopaonik: XIV scientific conference *YU INFO 2008*, 6, CD: ISBN 987-86-85525-03-2.

58. Vasić, V., Ćojbašić, V. (2005). Metode stepenastog postupka u diskriminacionoj analizi. Vrnjačka Banja: *SYMOPIS 2005*.
59. Vasić V,. Trpkova M. (2009) "Hierarchical cluster analysis: potential of the tourism up growths in the poorly developed Macedonian municipalities". *Proceedings from the 26th Symposium for operational research SYM-OP-IS 2009*, Ivanjica, 631 – 634.
60. Welch, B. L. (1939) "Note on Discriminant Functions". Biometrika, 31, 218 – 220.
61. Young, F. W. and R. M. Hamer. (1987) *Multidimensional Scaling: History, Theory, and Applications*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.