Klaster analiza u kreditnom skoriranju

Žagar, Ivana

Sveučilište J. J. Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni diplomski studij matematike
Financijska matematika i statistika

Ivana Žagar

Klaster analiza u kreditnom skoriranju

Diplomski rad

Osijek, 2023.

Sveučilište J. J. Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni diplomski studij matematike
Financijska matematika i statistika

Ivana Žagar

Klaster analiza u kreditnom skoriranju

Diplomski rad

Mentor: prof. dr. sc. Nataša Šarlija
Komentor: prof. dr. sc. Kristian Sabo

Osijek, 2023.

Sadržaj

Uvod

Često istraživači raznih znanosti imaju potrebu grupirati podatke u određene grupe ili skupine koji su slični po određenim obilježjima pri čemu se grupe međusobno razlikuju. Upravo se u takvim situacijama može koristiti klaster analiza. Klaster analiza vrsta je statističkog grupiranja čiji je cilj grupirati skup slučajeva ili pojedinaca u grupe koje se nazivaju klasteri. Podaci koji se nalaze u istom klasteru trebaju biti međusobno što sličniji, a dobiveni klasteri što različitiji. Na primjer, u biologiji se klaster analiza koristi za razvrstavanje vrsta, u geologiji za grupiranje minerala, u ekonomiji za grupiranje podataka na temelju određenih makroekonomskih varijabli te u mnogim drugim znanostima za grupiranje po raznim kriterijima. U ovom ćemo radu provesti klaster analizu na podacima jedne banke za mala i srednja poduzeća kako bismo ih klasterirali u različite klastere. Cilj klasteriranja je grupirati poduzeća tako da se u svakoj grupi odnosno klasteru nalaze poduzeća s istim ili sličnim karakteristikama što otvara mogućnost izgradnje modela skoriranja za svaki klaster posebno čime bi se mogla napraviti bolja procjena kreditnog rizika. Rad je podijeljen na dva djela. U prvom je djelu dana teorijska podloga klaster analize. Objašnjene su najčešće korištene metode klasteriranja i opisan je njen matematički smisao. Također, objašnjeno je kreditno skoriranje i postupci izrade modela kreditnog skoriranja te je dan osvrt na nekoliko istraživanja koja opisuju primjenu klaster analize u kreditnom skoriranju u kombinaciji s drugim metodama. U drugom je djelu rada provedena klaster analiza na stvarnim podacima jedne banke pomoću hijerarhijske metode klasteriranja. Objašnjeni su financijski pokazatelji i njihove vrijednosti, a obzirom na njih opisani su i dobiveni klasteri.

1 Klaster analiza

Grupiranje podataka koje se još naziva i klaster analiza statistička je procedura stvaranja grupa objekata ili klastera na način da su objekti u jednom klasteru vrlo slični, dok se objekti u različitim klasterima poprilično razlikuju [9]. Grupiranje objekata, koji mogu biti osobe, ustanove, poduzeća, jedinice lokalne uprave i slično, provodi se na osnovi njihovih zajedničkih obilježja. Iako je klaster analiza namijenjena grupiranju objekata, može se koristiti i za grupiranje varijabli. S obzirom da se osobine objekata definiraju pomoću varijabli, one tako ulaze u analizu [9]. Pretpostavka na kojoj počiva klaster analiza jest mogućnost nalaženja prirodnog načina razvrstavanja podataka smislenog za istraživača. Primjerice, u kontekstu istraživanja tržišta, klaster analiza se koristi za identifikaciju kategorija kao što su dobne skupine, urbana, ruralna ili prigradska lokacija.
Cilj klaster analize je grupirati objekte sa istim ili sličnim obilježjima zajedno. Kako bi istraživač to mogao napraviti potrebno je utvrditi koliko su objekti slični, odnosno različiti. Dakle, potrebno je definirati klastere i objekte koji im pripadaju. U tom smislu, možemo reći da je ideja korištenja klastera kombinirati kriterije i zahtjeve kako bi svi objekti u njemu [9]:

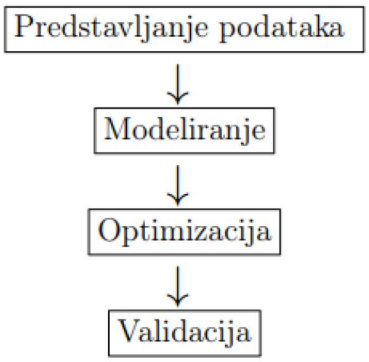
a) imali ista ili blisko povezana svojstva;

b) pokazivali male međusobne udaljenosti ili nesličnosti;

c) imali neki odnos ili vezu s barem jednim drugim objektom u skupini;

d) bili jasno razlikovani od objekata u drugim klasterima.

Postupak provedbe klaster analize naziva se klasteriranje. "Klasteriranje je postupak podjele nekog skupa podataka na unaprijed zadani broj klastera tako da je sličnost među elementima klastera najveća a razlika (udaljenost) među klasterima najveća"[17]. Kako bi algoritam za klasteriranje bio što bolji, on prema [9] uključuje četiri faze dizajna: predstavljanje podataka, modeliranje, optimizacija i validacija. Faza predstavljanja podataka unaprijed određuje kakve se strukture klastera mogu otkriti u podacima. Faza modeliranja definira pojam klastera i kriterije koji odvajaju poželjne grupne strukture od nepovoljnih.



Slika 1: Faze dizajna klaster analize (vidjeti [9])

Cilj istraživača je korištenjem klaster analize odgovoriti na tri ključna pitanja [6].

- Kako mjeriti sličnosti objekata?
- Kako formirati klastere?
- Kako utvrditi konačan broj klastera?

Klaster analiza je proces grupiranja objekata bez nadzora, odnosno nije poznato koliko klastera postoji u podacima prije pokretanja modela. S obzirom da je većina algoritama za klasteriranje vrlo osjetljiva na početne pretpostavke, potrebno joj je pristupiti s velikim oprezom, jer će ona rezultirati rješenjem i u slučaju krivih pretpostavki. Zbog toga je pažljiv odabir varijabli ključan dio ove analize. Kod kreiranja klastera treba uzeti u obzir samo one klastere i ona rješenja koja se mogu logički objasniti. Kako rješenje ovisi o analiziranim varijablama, njegova generalizacija nije moguća. Odnosno, ne postoji statistička osnova za zaključivanje sa uzorka na populaciju i ne postoji garancija jedinstvenog rješenja [13]. Stoga, dodavanje novih varijabli može znatno utjecati na konačno rješenje. Podjela objekata u klastere, u skladu s definiranim ciljevima, konačni je rezultat klaster analize. Odluke oko izbora mjere sličnosti, kriterija za svrstavanje u klaster i konačno, odabir broja klastera glavne su odrednice praktične provedbe klaster analize. U prvom se koraku izbor mjere sličnosti između svakog objekta temelji na osnovi sličnosti varijabli u procesu klasteriranja. Tri su metode određivanja sličnosti u klaster analizi: mjere udaljenosti, mjere korelacije i mjere asocijacije.

1.1 Mjere sličnosti

Prije klasteriranja potrebno je definirati mjeru sličnosti, odnosno udaljenosti [12]. Mjere sličnosti i udaljenosti igraju važnu ulogu u klaster analizi, a koriste se za kvantitativno opisivanje sličnosti ili različitosti dvaju objekata [9]. Mjera sličnosti definirana je kao udaljenost između

različitih objekata. Što je udaljenost manja, objekti su sličniji, odnosno što je koeficijent sličnosti veći, to su objekti sličniji [13]. Koeficijent sličnosti prikazuje snagu odnosa između dva objekta te se on računa za sve parove objekata, a najsličniji se objekti klasteriraju u klastere. Odabir mjere sličnosti ovisi o tipu podataka, a oni mogu biti kvantitativni(numerički) ili kvalitativni(kategorijalni). Numerički podaci, odnosno varijable poprimaju vrijednosti iz skupa realnih brojeva, a njihov se odnos može prikazati nekom funkcijom udaljenosti [9]. Stoga se za takve podatke koriste mjere udaljenosti i mjere korelacije. S druge strane, kategorijalne su varijable raspoređene u neke kategorije, a njihov se odnos ne može prikazati nekom funkcijom udaljenosti, stoga se za takve podatke koristi mjera asocijacije.

1.1.1 Mjere udaljenosti

Najčešće korištena mjera sličnosti u klaster analizi je mjera udaljenosti. Mjera udaljenosti zapravo je mjera različitosti između varijabli. Što je mjera različitosti veća, to je veća razlika između dva objekta. Najčešće korištene mjere udaljenosti između dviju točaka $x = (x_1, x_2, \ldots, x_n)$ i $y = (y_1, y_2, \ldots, y_n)$ su [9]:

- Manhattan udaljenost:
$$d_1(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$
- Euklidska udaljenost:
$$d_2(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$
- Minkowski udaljenost:
$$d_p(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}, \quad p > 1$$
- Prosječna udaljenost:
$$d(x, y) = \left(\frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2 \right)^{\frac{1}{2}}$$

Vrlo se često umjesto udaljenosti koristi tzv. *kvazimetrička funkcija* o kojoj ćemo nešto više reći u 1.3.1.

1.1.2 Mjere korelacije

Mjere korelacije temelje se na određivanju koeficijenta korelacije između objekata. Visoka korelacija označava veliku sličnost među objektima, dok niska korelacija upućuje na malu sličnost [9].

1.1.3 Mjere asocijacije

Mjere asocijacije koriste se za uspoređivanje kategorijalnih varijabli. Pomoću njih se određuje stupanj slaganja između svakog para objekata po svim atributima željenih karakteristika [13].

Neka su x i y dvije kategorijalne vrijednosti. Tada se za njihovu mjeru podudaranja koristi diskretna metrika definirana s [9]:

$$\delta(x, y) = \begin{cases} 1, & \text{za } x = y, \\ 0, & \text{za } x \neq y. \end{cases}$$

Kada su x i y kategorijalne varijable opisane s n atributa, različitost između njih, mjerena diskretnom metrikom, definirana je izrazom

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

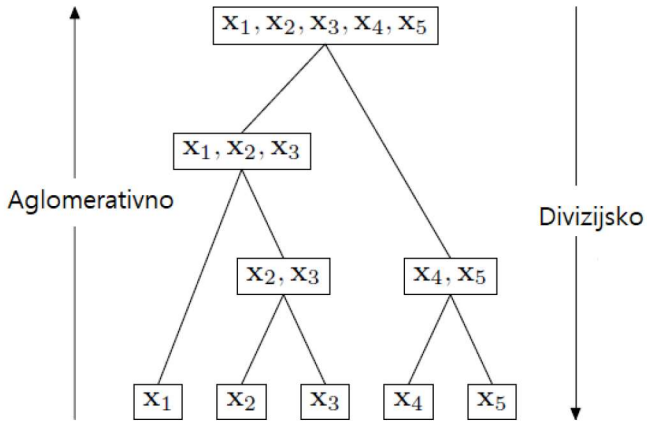
1.2 Hijerarhijske metode za klasteriranje

Općenito algoritme klasteriranja možemo podijeliti u dvije kategorije: hijerarhijski algoritmi i particijski algoritmi [9]. Hijerarhijski algoritmi podrazumijevaju izgradnju hijerarhijske strukture objekata pomoću dendograma. Dendogram je grafički prikaz rezultata u obliku dvodimenzionalnog hijerarhijskog dijagrama nalik stablu, koji ilustrira spajanja ili razdvajanja podataka. Više o dendogramu može se pogledati u [9].

Hijerarhijski algoritmi započinju tako da se izračuna udaljenost između svih objekata zajedno. Nakon što se njihova međusobna udaljenost izračuna, kreće se u stvaranje klastera. Postoje dvije vrste hijerarhijskih algoritama, to su hijerarhijski algoritmi koji razdvajaju (divizijski) i aglomerativni hijerarhijski algoritmi. Kod divizijskih algoritama skup od n podataka dijeli se u manje, finije grupe. Kažemo da se podaci razdvajaju od vrha prema dnu, odnosno algoritam počinje s jednim velikim klasterom koji sadrži sve objekte te ih zatim dijeli u dva nova klastera tako da su objekti u jednom klasteru što različitiji od objekata u drugom klasteru. Postupak dijeljenja svakog klastera nastavlja se dok svaki pojedini objekt ne postane zasebni klaster, odnosno dok broj klastera ne bude jednak broju objekata. U aglomerativnim hijerarhijskim algoritmima postupak je obrnut, odnosno ide se od dna prema vrhu. Dakle, algoritmi počinju s klasterima od kojih svaki sadrži jedan objekt, stoga na početku imamo n klastera. Nakon toga se na temelju matrice sličnosti dva najsličnija objekta klasteriraju u novi klaster. Postupak spajanja klastera ponavlja se $n - 1$ puta, odnosno dok se svi pojedini klasteri ne sjedine u jedan klaster [9].

Iako su oba algoritma zadovoljavajuća, aglomerativni se algoritmi koriste mnogo češće od divizijskih pa ćemo u daljnjem radu detaljnije obraditi aglomerativne algoritme.

Aglomerativni algoritmi se s obzirom na način na koji se određuje sličnost među klasterima mogu podijeliti na tri metode: metode povezivanja, Wardova metoda i centroidna metoda. Metode povezivanja i centroidna metoda u svakom koraku spajaju pojedinačne objekte ili klastere koji su najbliži, a razlikuju se samo po definiciji udaljenosti, odnosno sličnosti između podataka. S druge strane, Wardova metoda se razlikuje od prethodnih jer prilikom spajanja klastera analizira varijancu između objekata.



Slika 2: Aglomerativno i divizijsko hijerarhijsko klasteriranje

1.2.1 Metode povezivanja

Metode povezivanja dijele se na tri metode ovisno o tome kako se određuje reprezentant klastera [13].

1. Jednostruko povezivanje (engl. *single-linkage method*)
2. Potpuno povezivanje (engl. *complete-linkage method*)
3. Prosječno povezivanje (engl. *average linkage*)

Jednostruko povezivanje

Jednostruko povezivanje jedna je od najjednostavnijih hijerarhijskih metoda klasteriranja. Još se naziva metoda najbližeg susjeda ili metoda minimalne udaljenosti. U toj se metodi definira sličnost između dva klastera kao najmanja udaljenost između bilo kojeg objekta iz jednog klastera i bilo kojeg objekta iz drugog klastera [9, 13]. Klasteri se stvaraju na način da se objekti čija je udaljenost najmanja klasteriraju u novi klaster. Nakon kreiranja tog klastera, sljedeći se objekti povezuju s njim tako što se gleda najmanja udaljenost između samostalnih objekata i objekata već kreiranog klastera. Postupak se dalje ponavlja na način da se udaljenost između dva klastera određuje na temelju udaljenosti između bilo koja dva objekta tih klastera.

Neka su primjerice Z_i, Z_j, Z_k neka tri klastera. Tada se udaljenost između Z_k i $Z_i \cup Z_j$ može izračunati kao

$$\begin{aligned} D(Z_k, Z_i \cup Z_j) &= \frac{1}{2}D(Z_k, Z_i) + \frac{1}{2}D(Z_k, Z_j) - \frac{1}{2}|D(Z_k, Z_i) - D(Z_k, Z_j)| \\ &= \min\{D(Z_k, Z_i), D(Z_k, Z_j)\}, \end{aligned}$$

gdje je D udaljenost između dva klastera. Odnosno,

$$D(X, Y) = \min_{x \in X, y \in Y} d(x, y),$$

pri čemu su X i Y dva neprazna disjunktna klastera [9].

Potpuno povezivanje

Metoda potpunog povezivanja suprotna je metodi jednostrukog povezivanja. U svakom se koraku sličnost između dva klastera određuje tako da je udaljenost između njihovih elemenata maksimalna [9]. Točnije, udaljenost između klastera Z_k i $Z_i \cup Z_j$ računa se kao

$$\begin{aligned} D(Z_k, Z_i \cup Z_j) &= \frac{1}{2}D(Z_k, Z_i) + \frac{1}{2}D(Z_k, Z_j) + \frac{1}{2}|D(Z_k, Z_i) - D(Z_k, Z_j)| \\ &= \max\{D(Z_k, Z_i), D(Z_k, Z_j)\}, \end{aligned}$$

gdje je D udaljenost između dva klastera. Odnosno,

$$D(X, Y) = \max_{x \in X, y \in Y} d(x, y),$$

pri čemu su X i Y dva neprazna disjunktna klastera [9].

Prosječno povezivanje

Prosječno povezivanje je metoda povezivanja koja povezuje klastere ovisno o tome kolika je prosječna udaljenost između svih parova objekata pri čemu jedan član pripada jednom klasteru, a drugi član pripada drugom klasteru. Tada se, prema [9], udaljenost između klastera Z_k i $Z_i \cup Z_j$ računa se kao

$$D(Z_k, Z_i \cup Z_j) = \frac{|Z_i|}{|Z_i| + |Z_j|}D(Z_k, Z_i) + \frac{|Z_j|}{|Z_i| + |Z_j|}D(Z_k, Z_j),$$

gdje je D udaljenost između dva klastera. Dakle, za dva neprazna disjunktna klastera X i Y vrijedi

$$D(X, Y) = \frac{1}{|X||Y|} \sum_{x \in X, y \in Y} d(x, y).$$

1.2.2 Centroidna metoda

U ovoj se metodi sličnost između klastera definira kao udaljenost između centroida klastera [9], [13]. Centroid klastera je srednja vrijednost objekata u klasteru po svim varijablama uključenima u analizu klastera. Dodavanjem novih objekata u klastere vrijednost centroida se mijenja. Prema [9], udaljenost između klastera Z_k i $Z_i \cup Z_j$ računa se kao

$$D(Z_k, Z_i \cup Z_j) = \frac{|Z_i|}{|Z_i| + |Z_j|}D(Z_k, Z_i) + \frac{|Z_j|}{|Z_i| + |Z_j|}D(Z_k, Z_j) - \frac{|Z_i||Z_j|}{(|Z_i| + |Z_j|)^2}D(Z_i, Z_j).$$

Pri čemu je D udaljenost između dva neprazna disjunktna klastera X i Y definirana kao

$$D(X, Y) = \frac{1}{|X||Y|} \sum_{x \in X, y \in Y} d(x, y) - \frac{1}{2|X|^2} \sum_{x, y \in X} d(x, y) - \frac{1}{2|Y|^2} \sum_{x, y \in Y} d(x, y).$$

1.2.3 Wardova metoda

Wardova metoda hijerarhijska je procedura klasteriranja u kojoj se dva klastera spajaju po kriteriju greške sume kvadrata (SSE). Zbog toga se Wardova metoda često naziva i metoda minimalne varijance. Za dani skup podataka Z, suma kvadrata grešaka je

$$SSE(Z) = \sum_{x \in Z} (x - \mu(Z))(x - \mu(Z))^T,$$

gdje je $\mu(Z)$ srednja vrijednost od Z, tj.

$$\mu(Z) = \frac{1}{|Z|} \sum_{x \in C} x.$$

Pretpostavimo da postoji k klastera $Z_1, Z_2, \ldots, Z_k$ u jednom klasteriranju. Tada je gubitak informacija prema [9] predstavljen s

$$SSE = \sum_{i=1}^{k} SSE(Z_i).$$

U ovoj se metodi za izračunavanje udaljenosti najčešće primjenjuje euklidska udaljenost. Wardova se metoda smatra vrlo učinkovitom, a njen cilj je kreiranje klastera s malim brojem objekata i s približno jednakim brojem objekata u svakom klasteru [13].

Kao glavna prednost hijerarhijskog klasteriranja navodi se njezina brzina i jednostavnost [13]. Naime, hijerarhijske metode jednim provođenjem nude cijeli skup mogućih rješenja koje istraživači mogu analizirati. Također, zbog široke upotrebe hijerarhijskih metoda došlo je do razvoja mjera sličnosti za gotovo svaki tip varijabli i vrstu istraživanja. S druge strane, hijerarhijske metode osjetljive su na outliere, a ako su početni objekti pogrešno svrstani, mogu dovesti do pogrešnih zaključaka.

1.3 Klaster analiza u matematičkom smislu

Promotrimo sada problem klasteriranja elemenata u disjunktne neprazne podskupove, odnosno klastere.

Neka je $\mathcal{A}$ skup koji sadrži $m \geq 2$ elemenata koje želimo klasterirati u disjunktne podskupove $\pi_1, \ldots, \pi_k, 1 \leq k \leq m$, takve da vrijedi

$$\bigcup_{i=1}^{k} \pi_i = \mathcal{A}, \quad \pi_i \bigcap \pi_j = \emptyset, \; i \neq j, \quad |\pi_j| \geq 1, \quad j = 1, \ldots, k, \tag{1}$$

na osnovu jednog ili više obilježja uz korištenje raznih kriterijskih funkcija cilja [18]. Ovako definirane rastave skupa $\mathcal{A}$ na podskupove $\pi_1, \ldots, \pi_k$, koji zadovoljavaju (1) označavamo s $\Pi = \{\pi_1, \ldots, \pi_k\}$ i zovemo *particija* skupa $\mathcal{A}$, a elemente particije zovemo *klasteri*. Skup svih particija skupa $\mathcal{A}$ sastavljenih od k klastera koje zadovoljavaju (1) označavamo s $\mathcal{P}(\mathcal{A}; k)$ [20]. Broj svih particija skupa $\mathcal{A}$ sastavljenih od k klastera jednak je *Stirlingovom broju*

druge vrste[1]

$$|\mathcal{P}(\mathcal{A};k)| = \frac{1}{k!}\sum_{j=1}^{k}(-1)^{k-j}\binom{k}{j}j^m. \tag{2}$$

Primjer 1.1. Neka je dan skup $\{a,b,c,d\}$. Koliko je dvočlanih particija ovakvog skupa?

Rješenje: Primijetimo da imamo skup od 4 člana i tražimo sve dvočlane particije. To su redom particije:

$$\begin{array}{ll} \{a,b,c\}\cup\{d\} & \{a,b\}\cup\{c,d\} \\ \{a,b,d\}\cup\{c\} & \{a,c\}\cup\{b,d\} \\ \{a,c,d\}\cup\{b\} & \{a,d\}\cup\{b,c\} \\ \{b,c,d\}\cup\{a\} & \end{array}$$

Vidimo kako ima 7 particija. Pogledajmo sada koliko iznosi Stirlingov broj druge vrste koristeći (2).

$$\begin{aligned} |\mathcal{P}(\mathcal{A};2)| &= \frac{1}{2!}\sum_{j=1}^{2}(-1)^{2-j}\binom{2}{j}j^4 \\ &= \frac{1}{2}\left[(-1)^{2-1}\binom{2}{1}1^4 + (-1)^{2-2}\binom{2}{2}2^4\right] \\ &= \frac{1}{2}(-2+16) \\ &= 7 \end{aligned}$$

Kao što možemo vidjeti, broj particija jednak je 7, a to je upravo Stirlingov broj druge vrste za ovaj primjer.

Broj svih mogućih načina da se m različitih elemenata klasterira u k nepraznih skupova može biti vrlo velik broj. Ideja korištenja klaster analize u matematičkom smislu je pronaći optimalnu particiju. U nekim je slučajevima broj klastera u particiji skupa $\mathcal{A}$ određen iz prirode promatranog problema [18]. Ukoliko broj klastera nije unaprijed zadan, prirodno je tražiti optimalnu particiju s klasterima koji su što kompaktnije klasterirani i imaju što bolju međusobnu razdvojenost. Kako bismo mogli primijeniti određene kriterije za pronalazak optimalne particije, podaci moraju biti prikazani skupom realnih brojeva ukoliko je riječ o objektima s jednim obilježjem, odnosno skupom točaka ukoliko je riječ o objektima s više obilježja. Metode za klasteriranje mogu se podijliti na klasteriranje na bazi centra te na hijerarhijske metode koje smo opisali u prethodnom poglavlju.

1.3.1 Klasteriranje na bazi centra

Pretpostavimo da želimo klasterirati objekte koji su reprezentirani s n obilježja. Neka je $\mathcal{A} = \{a_1,\ldots,a_m\}$ skup kojega na osnovi n obilježja želimo klasterirati u k klastera koji zadovoljavaju (1). S obzirom na dana obilježja, svaki element $a_i \in \mathcal{A}$ reprezentirat ćemo jednom točkom $(x_{i1},\ldots,x_{in}) \in \mathbb{R}^n$ koju ćemo prema [18] označiti s $\mathbf{a}_i$. Podaci iz skupa $\mathcal{A}$

[1]Stirlingovi brojevi dobili su naziv po Jamesu Stirlingu. Postoje Stirlingovi brojevi prve i druge vrste. Ilustrativno, Stirlingov broj druge vrste može se opisati kao broj načina da se n različitih kuglica stavi u k jednakih kutija, pri čemu niti jedna kutija ne smije ostati prazna. Još se koristi oznaka $S(n,k)$. Više o Stirlingovim brojevima može se pročitati u [7].

ne moraju biti nužno različiti. Kao što smo ranije rekli, podatke ćemo klasterirati u klastere ovisno o vrijednostima mjere sličnosti. Kao mjeru sličnosti koristit ćemo neku od funkcija udaljenosti.

Definicija 1.1. Funkciju $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ koja ima svojstvo pozitivne definitnosti

$$d(x,y) \geq 0,\ \forall x, y \in \mathbb{R}^n \quad \text{i} \quad d(x,y) = 0 \Leftrightarrow x = y,$$

zovemo *kvazimetrička funkcija* na $\mathbb{R}^n$.

Prema [19] najčešće korištene kvazimetričke funkcije su *funkcija najmanjih kvadrata* i l_1-metrička funkcija, poznatija kao *Manhattan metrička funkcija*. O njima ćemo nešto više reći u nastavku.

U svrhu klasteriranja objekata s n obilježja, svakom klasteru $\pi_j \in \Pi$ pridružit ćemo njegov *centar* c_j. Uvedimo najprije oznaku globalnog minimuma. Skup svih točaka u kojima funkcija $h\colon \mathcal{D} \to \mathbb{R}$ postiže globalni minimum označavamo s $\underset{x\in\mathcal{D}}{\operatorname{argmin}}\, h(x)$. Neka je zadana neka kvazimetrička funkcija $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, tada svakom klasteru $\pi_j \in \Pi$ pridružujemo njegov centar c_j na sljedeći način

$$c_j := \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \sum_{\mathbf{a}_i \in \pi_j} d(x, \mathbf{a}_i). \tag{3}$$

Nadalje, na skupu svih particija $\mathcal{P}(\mathcal{A}; k)$ skupa $\mathcal{A}$ sastavljenih od k klastera definiramo *funkciju cilja* [19] $\mathcal{F}\colon \mathcal{P}(\mathcal{A}; k) \to \mathbb{R}_+$ s

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{\mathbf{a}_i \in \pi_j} d(c_j, \mathbf{a}_i), \tag{4}$$

tada $d-$optimalnu k-particiju Π^* tražimo rješavanjem sljedećeg optimizacijskog problema

$$\mathcal{F}(\Pi^*) = \min_{\Pi \in \mathcal{P}(\mathcal{A};k)} \mathcal{F}(\Pi). \tag{5}$$

Problem traženja optimalne particije možemo preformulirati na problem tražnja optimalnih centara.

Definicija 1.2. Neka je $\mathcal{A} = \{\mathbf{a}_i = (x_{i1}, \dots x_{in}) \in \mathbb{R}^n : i = 1, \dots, m\}$ skup točaka iz $\mathbb{R}^n$. Kažemo da je particija Π^* optimalna u *smislu najmanjih kvadrata* (**LS-optimalna**[2]) ako je Π^* rješenje optimizacijskog problema (4)-(5), a kvazimetrička funkcija $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ definirana s

$$d(a_1, a_2) = \|a_1 - a_2\|_2^2. \tag{6}$$

Centri $c_1, \dots, c_k$ klastera $\pi_1, \dots, \pi_k$ nazivaju se *centroidi*[3] i određeni su s

$$c_j = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \sum_{\mathbf{a}_i \in \pi_j} \|\mathbf{a}_i - x\|_2^2 = \frac{1}{|\pi_j|} \sum_{\mathbf{a}_i \in \pi_j} \mathbf{a}_i, \quad j = 1, \dots, k, \tag{7}$$

a funkcija cilja s

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{\mathbf{a}_i \in \pi_j} \|c_j - \mathbf{a}_i\|_2^2. \tag{8}$$

Definicija 1.3. Neka je $\mathcal{A} = \{\mathbf{a}_i = (x_{i1}, \dots, a_{in}) \in \mathbb{R}^n : i = 1, \dots, m\}$ skup točaka iz $\mathbb{R}^n$. Kažemo da je particija Π^* optimalna u *smislu najmanjih apsolutnih odstupanja* (**LAD-optimalna**[4]) ako je Π^* rješenje optimizacijskog problema (4)-(5), a metrička funkcija $d\colon \mathbb{R}^n \times$

[2]engl. Least Squares

[3]Centroid skupa vektora $c(A) = (\bar{x}, \bar{y}) \in \mathbb{R}^2, \bar{x} = \frac{1}{m}\sum_{i=1}^m x_i, \bar{y} = \frac{1}{m}\sum_{i=1}^m y_i$.

[4]engl. Least Absolute Deviations

$\mathbb{R}^n \to \mathbb{R}_+$ definirana s

$$d(x, y) = \|x - y\|_1. \tag{9}$$

Centri $c_1, \ldots, c_k$ klastera $\pi_1, \ldots, \pi_k$ određeni su s

$$c_j = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \sum_{\mathbf{a}_i \in \pi_j} \|\mathbf{a}_i - x\|_1 = \operatorname{med}(\pi_j) \quad j = 1, \ldots, k, \tag{10}$$

a funkcija cilja s

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{\mathbf{a}_i \in \pi_j} \|c_j - \mathbf{a}_i\|_1. \tag{11}$$

Pogledajmo primjere klasteriranja objekata s jednim obilježjem.

Primjer 1.2. Odredimo sve dvočlane particije skupa $\mathcal{A} = \{1, 3, 4, 8\}$ koje zadovoljavaju (1). Zatim odredimo pripadne centre i vrijednosti kriterijske funkcije cilja $\mathcal{F}$ u smislu najmanjih kvadrata.

π_1	π_2	c_1	c_2	$\mathcal{F}(\Pi)$	
$\{1\}$	$\{3,4,8\}$	1	5	$0 + 14 =$	14
$\{3\}$	$\{1,4,8\}$	3	13/3	$0 + 74/3 =$	24.67
$\{4\}$	$\{1,3,8\}$	4	4	$0 + 26 =$	26
$\{8\}$	$\{1,3,4\}$	8	8/3	$0 + 14/3 =$	4.67
$\{1,3\}$	$\{4,8\}$	2	6	$2 + 8 =$	10
$\{1,4\}$	$\{3,8\}$	5/2	11/2	$9/2 + 25/2 =$	17
$\{1,8\}$	$\{3,4\}$	9/2	7/2	$49/2 + 1/2 =$	25

Tablica 1: Biranje optimalne particije skupa $\mathcal{A}$ primjenom *LS-kvazimetričke funkcije*

Primjer 1.3. Odredimo sve dvočlane particije skupa $\mathcal{A} = \{1, 3, 4, 8\}$ koje zadovoljavaju (1). Zatim odredimo pripadne centre i vrijednosti kriterijske funkcije cilja $\mathcal{F}$ u smislu najmanjih apsolutnih odstupanja.

π_1	π_2	c_1	c_2	$\mathcal{F}(\Pi)$	
$\{1\}$	$\{3,4,8\}$	1	4	$0 + 5 =$	5
$\{3\}$	$\{1,4,8\}$	3	4	$0 + 7 =$	7
$\{4\}$	$\{1,3,8\}$	4	3	$0 + 7 =$	7
$\{8\}$	$\{1,3,4\}$	8	3	$0 + 3 =$	3
$\{1,3\}$	$\{4,8\}$	1	6	$2 + 4 =$	6
$\{1,4\}$	$\{3,8\}$	1	3	$3 + 5 =$	8
$\{1,8\}$	$\{3,4\}$	8	4	$7 + 1 =$	8

Tablica 2: Biranje optimalne particije skupa $\mathcal{A}$ primjenom *LAD-metričke funkcije*

Primijetimo da se minimalna vrijednost funkcije iz Primjera 1.3 postiže na particiji $\Pi^* = \{\{8\}, \{1, 3, 4\}\}$ koja je također bila i LS-optimalna dvočlana particija (vidi Primjer 1.2). U ovim se primjerima to dogodilo slučajno. Općenito se LS-optimalna i LAD-optimalna particija ne moraju podudarati.

1.3.2 Traženje optimalne particije

Traženje optimalne particije putem pretraživanja cijelog skupa $\mathcal{P}(\mathcal{A}; k)$ nije moguće izvršiti u prihvatljivom vremenu. Pretraživanje skupa svih particija čiji se klasteri međusobno nastavljaju također može biti vremenski vrlo zahtjevno. Stoga ćemo predstaviti najčešće korišten algoritam za pronalaženje particije dosta bliske optimalnoj. *k-means* algoritam iterativni je proces koji u konačno mnogo koraka daje rješenje i u svakom koraku snižava vrijednost funkcije cilja. Što je vrijednost funkcije cilja manja, time je "rasipanje" manje, odnosno klasteri su kompaktniji i međusobno bolje razdvojeni.

k-means algoritam u slučaju podataka s više obilježja

k-means algoritam u slučaju podataka s više obilježja može se opisati u sljedeća dva koraka [19].

ALGORITAM 1.1. (*k-means* algoritam)

Korak A **Pridruživanje**. Poznavanjem međusobno različitih točaka $z_1, \ldots, z_k$ skup $\mathcal{A}$ treba klasterirati u k disjunktnih klastera $\pi_1, \ldots, \pi_k$ korištenjem **principa minimalnih udaljenosti**

$$\pi_j = \{a \in \mathcal{A} \colon d(z_j, a) \leq d(z_s, a),\ \forall s = 1, \ldots, k\}, \quad j = 1, \ldots, k.$$

Korak B **Korekcija**. Za poznatu particiju $\Pi = \{\pi_1, \ldots, \pi_k\}$ skupa $\mathcal{A}$, treba definirati centre klastera

$$c_j = \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{a \in \pi_j} d(x, a), \quad j = 1, \ldots, k.$$

U **Koraku A** cijeli skup $\mathcal{A}$ se principom minimalnih udaljenosti razdjeljuje u k skupina prema njihovoj d-bliskosti pojedinim točkama z_j.

U **Koraku B** svakom klasteru particije Π pridružujemo njegove centre. Ako je d LS-kvazimetrička funkcija, centri su centroidi klastera, a ako je d LAD-metrička funkcija, centri su medijani elemenata klastera.

k-means algoritam ponavlja navedene korake dok se particije ne počnu ponavljati, odnosno dok vrijednost funkcije cilja prestane opadati, a centri klastera trenutne i prethodne particije međusobno se podudaraju. U oba se slučaja algoritam može pokrenuti zadavanjem početne particije ili zadavanjem početnih centara.

1.4 Određivanje primjerenog broja klastera u particiji

Kako bismo što bolje razumjeli i kasnije interpretirali rezultate klasteriranja, kažimo nešto o metodama za pronalazak primjerenog broja klastera.

1.4.1 Metoda lakta

Metoda lakta (enlg. *elbow method*) empirijska je metoda za pronalaženje primjerenog broja klastera za skup podataka koja promatra ukupnu sumu kvadrata unutar klastera. U ovoj se metodi odabire mogući raspon vrijednosti za k, a zatim se koristeći svaku od k vrijednosti provede neka od metoda klasteriranja. Primjerena vrijednost za k je ona gdje ukupna suma kvadrata iznenada počne padati. Vizualno, dijagram izgleda kao lakat te se za optimalan broj klastera uzima onaj k u kojem se lakat nalazi [22, 14].

1.4.2 Metoda prosječne siluete

Metoda prosječne siluete (engl. *average silhouette method*) mjeri koliko je dobro opažanje klasterirano i procjenjuje prosječnu udaljenost između klastera. Izračunava koeficijent siluete svake točke koji mjeri koliko je točka slična vlastitom klasteru u usporedbi s drugim klasterima. Koeficijent siluete za svako opažanje j izračunava se na sljedeći način [5].

i) Za svako opažanje j potrebno je izračunati prosječnu udaljenost a_j između j i svih ostalih točaka klastera kojemu j pripada.

ii) Za sve druge klastere C kojima j ne pripada, potrebno je izračunati prosječnu udaljenost $d(j, C)$ točke j do svih klastera C. Najmanju takvu udaljenost označimo s $b_j = \min d(j, C)$.

iii) Koeficijent siluete $j-$te točke dan je formulom: $s_j = \frac{b_j - a_j}{\max\{b_j, a_j\}}$.

Nakon što se izračuna koeficijent siluete svake točke u skupu podataka, crta se dijagram kako bi se dobio vizualni prikaz o tome koliko je dobro skup podataka klasteriran u k klastera. Dijagram siluete prikazuje mjeru koliko je svaka točka u jednom klasteru blizu točkama u susjednim klasterima i tako pruža način za vizualnu procjenu parametra poput broja klastera, a ima raspon od -1 do 1. Ukoliko je koeficijent siluete 1 ili blizu 1 tada je uzorak daleko od susjednih klastera. Ako je koeficijent veći od 0 znači da je uzorak vrlo blizu granice dvaju klastera ili na samoj granici. Koeficijent koji je manji od 0 ukazuje na to da su uzorci dodijeljeni krivom klasteru ili su outlieri [22].

Osim navedenih metoda, za određivanje primjerenog broja klastera mogu se koristiti i Calinski-Harabasz index [16], Davies-Bouldin index [16, 4], Dunn index [4], Partition coefficient [4] i dr.

2 Kreditno skoriranje

U zemljama s razvijenim sustavom financijske usluge krediti se izdaju samo onim klijentima koji su prošli poseban postupak procjene kreditne sposobnosti koji se naziva kreditno skoriranje [29]. Kreditno skoriranje je postupak procjene kreditne sposobnosti uzimajući u obzir različite kriterije, primjerice prihode klijenta, imovinu, prethodnu kreditnu povijest i slično. Odluku o pitanju davanja kredita i razmatranje uvjeta kreditiranja provodi kreditni odbor banke. S obzirom da se te odluke temelje na subjektivnom mišljenju pojedinih članova odbora o riziku kreditiranja, uz poštivanje pravila ocjene kreditne sposobnosti propisane od banke, donesene odluke ne odražavaju uvijek stvarnu sliku. Problem subjektivnosti pri donošenju odluke moguće je riješiti pomoću statističkih alata za obradu podataka koji implementiraju bodove za ocjenu kreditne sposobnosti kao što je kreditno skoriranje. Kreditno skoriranje je alat koji se odnosi na modele i sustave koji daju numeričku kreditnu ocjenu za svakog klijenta, pri čemu je kreditna ocjena pokazatelj kreditne sposobnosti dužnika i vjerojatnosti[5] da će isti kasniti u otplati kredita [8]. Tu ocjenu, odnosno kvantitativnu mjeru nazivamo *skor* [25]. Skor se za svakog klijenta izračunava koristeći skor-karticu koja obuhvaća skupinu karakteristika koje su raspodijeljene prema atributima, od kojih svaki ima statistički izveden skor. Skorovi pojedinačnih obilježja se zbrajaju te se dobije ukupan skor koji se uspoređuje s graničnom vrijednošću, a ona predstavlja najvišu razinu rizika s kojom je

[5]engl. Probability of default

financijska institucija spremna raditi. Primjerice, ukoliko financijska institucija odluči postaviti graničnu vrijednost na 0,6, to bi značilo da ne želi odobriti kredit onim poduzećima za koje je vjerojatnost da će kasniti u otplati veća od 60% [1]. Na temelju te usporedbe donosi se odluka o odobravanju odnosno neodobravanju kredita [25]. Predviđanje kreditne sposobnosti klijenata iznimno je važno za svakog kreditora. Naime, odobravanje kredita klijentu koji će kasniti u plaćanju ili neće vratiti kredit u cijelosti donosi gubitke kreditoru. S druge strane, odbijanje klijenta koji je potencijalno dobar kreditoru donosi manju zaradu. Prema [29] glavne prednosti korištenja kreditnog skoriranja su:

- smanjenje rizika od neplaćanja kredita, odnosno smanjenje broja loših klijenata;
- povećanje kreditnog portfelja smanjenjem subjektivnih odbijanja zahtjeva za kredit;
- ubrzanje procesa donošenja odluke o izdavanju kredita, odnosno smanjenje vremena i troškova potrebnih za procjenu.

Prije nego li opišemo proces razvoja modela, predstavimo osnovne specifikacije sustava bodovanja kako je navedeno u pravilima Baselskog odbora [8].

1. Potrebno je smisleno razlikovati rizik. Točnije, sustav kreditnog rejtinga mora biti osmišljen tako da razlikuje rizik. Ocjene sustava moraju biti pravilno definirane kako bi predstavljale različite razine kreditnog rizika.

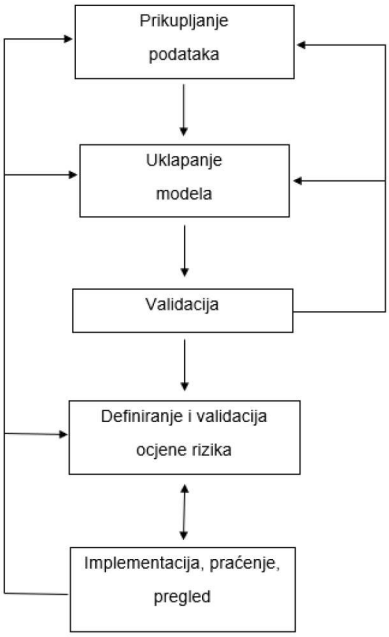
2. Svi klijenti i krediti u kreditnom portfelju trebali bi se barem jednom godišnje dopuniti, uzimajući u obzir nove informacije o statusu i napretku klijenta.

3. Rad sustava kreditnog bodovanja mora se stalno nadzirati zbog pravilnog funkcioniranja i ispravnosti. Osim toga, sustav treba biti podvrgnut dostatnim i učinkovitim kontrolama.

4. Ispravan odabir atributa procjene rizika. Programeri bi trebali biti u mogućnosti pokazati da se kreditna sposobnost klijenta može ispravno analizirati na temelju čimbenika rizika i atributa koje razmatraju sustavi kreditnog bodovanja. Analiza bi se trebala temeljiti na budućoj uspješnosti klijenta na temelju trenutno dostupnih informacija o klijentu i vanjskom okruženju.

5. Baza podataka koja se koristi za razvoj i ocjenu sustava kreditnog bodovanja mora biti reprezentativna za stvarnost. Mora sadržavati povijesne podatke o karakteristikama klijenta, prošlim kreditnim rezultatima, ocjenama, povijesti plaćanja i slično.

S obzirom na navedeno, možemo zaključiti kako je razvoj kreditnog skoriranja složen proces. Na Slici 3 ilustriran je pregled procesa kreditnog skoriranja.



Slika 3: Proces kreditnog skoriranja (vidjeti [8])

2.1 Postupak izrade modela kreditnog skoriranja

Kao što smo ranije naveli, kako bismo predvidjeli kreditnu sposobnost klijenta, potrebno je odrediti skor koji se izračunava koristeći skor-karticu. Izrada skor-kartice ne počinje prikupljanjem podataka, već planiranjem unaprijed, prije početka bilo kakvog rada. Što uključuje utvrđivanje razloga ili cilja projekta, identifikaciju ključnih sudionika u razvoju i implementaciji skor-kartice, dok je cilj izgradnje osigurati održivost projekta. Opišimo stoga osnovne korake u izgradnji skor-kartice [21].

2.1.1 Studija provedivosti

Prvi korak u izradi skor-kartice je identifikacija i određivanje prioriteta organizacijskih ciljeva za taj projekt. Točnije, potrebno je definirati troškove i koristi izrade, faze izrade, zahtjeve i odgovornosti za svaku fazu izrade te implementaciju [25]. Također, treba odlučiti hoće li se ići na unutarnju ili vanjsku izradu skor-kartice, odnosno hoće li se ona izrađivati unutar tvrtke ili će se kupiti od vanjskih dobavljača [21]. Općenito se banke, ukoliko imaju dovoljno podataka, odlučuju za internu izgradnju skor-kartice zbog brže i jeftinije izgradnje te veće fleksibilnosti. Međutim, ta odluka može ovisiti o čimbenicima kao što su dostupnost izvora, stručnost u razvoju kartice, vremenski okvir i slično. Kako bi se ta odluka olakšala, navedimo neke prednosti i nedostatke[6] vanjske izrade skor-kartice [21].

- Jeftinija je za vrlo male portfelje, a skupa za velike portfelje.
- Manje je rizična za kreditora jer nema potrebe za iskorištavanjem resursa. Međutim, manji rizik donosi i manju nagradu u smislu povrata na uloženo.
- Manja fleksibilnost. Primjerice, zbog ograničenih troškova istraživanje detaljnih strategija segmentacije može biti ograničeno.

[6]Više o prednostima i nedostacima vanjske i unutarnje izrade skor-kartice pogledati u [21].

Nakon toga, potrebno je napraviti plan projekta, identificirati moguće rizike te odrediti projektni tim. Ukoliko je odlučeno ići na izgradnju modela, prate se sljedeći koraci.

2.1.2 Prikupljanje podataka

Kako bi se zadržale karakteristike koje se koriste u izradi skor-kartice prvo je potrebno prikupiti podatke. Općenito se podaci prikupljaju iz zahtjeva za kreditom, izvještaja kreditnog biroa, geo-demografske baze i drugo [25]. Primjerice, za stanovništvo se analiziraju podaci kao što su dob, stambeni status, postojeći odnos s bankom, geografski položaj, prosječne vrijednosti različitih pokazatelja temeljenih na izvještajima kreditnog biroa i svi drugi kriteriji koji će pomoći u izradi profila klijenta. Uzorak koji se koristi za modeliranje skor-kartice treba biti iz recentnog vremenskog razdoblja i treba reprezentirati buduće podnositelje kreditnih zahtjeva [21]. Na kraju je potrebno zadržati sve one karakteristike za koje će se statističkom analizom utvrditi značajnost u predikciji budućeg ponašanja te je potrebno poznavati kvalitetu podataka.

2.1.3 Analiza karakteristika

U ovom se koraku, tijekom određenog vremenskog razdoblja, prikupljaju zahtjevi za kreditima. Zatim se za svakog klijenta, odnosno zahtjev određuje radi li se o dobrom, lošem, neodređenom, neaktivnom ili odbijenom klijentu. Prikupljene karakteristike, zajedno s dobrom/lošom klasifikacijom čine uzorak iz kojeg se razvija skor-kartica. Svaki kreditor određuje svoju definiciju dobrih i loših klijenata u ovisnosti o tome čemu je namijenjen skoring model i koji je klijent stvarno dobar odnosno loš za kreditora. Definicija onoga što zahtjev čini lošim oslanja se na nekoliko razmatranja [21].

- Definicija mora biti u skladu s projektnim ciljevima. Ako je cilj povećati profitabilnost, onda se definicija mora postaviti na točku kašnjenja u kojoj račun postaje neprofitabilan.
- Definicija mora biti u skladu s proizvodom ili svrhom za koju se izrađuje skor-kartica. Na primjer bankrot, prijevara, potraživanja i naplate.
- Definicija se mora lako tumačiti i pratiti. Primjerice, klijent koji kasni u plaćanju barem jedne rate kredita više od 60 dana smatra se lošim [25]. Odabir jednostavnije definicije olakšava upravljanje i donošenje odluke. Na primjer, ako kažemo da ima 4% loših zahtijeva, to znači da je kod 4% zahtijeva zabilježeno kašnjenje u plaćanju. Na temelju sporazuma o adekvatnosti kapitala (Basel II) definicija neispunjavanja obveza općenito je 90 dana kašnjenja.

Kao što smo ranije rekli, svaki kreditor određuje svoju definiciju lošeg klijenta. Na primjer, ako banka utvrdi da je broj loših klijenata promatranih u 90 dana nizak, može se odlučiti za izračun skor-kartice za predviđanje od 60 dana. Također, ako je i tada broj loših klijenata nizak, može ići na promatranje razdoblja od 30 dana. S druge strane, klijenti koji nikada nisu kasnili u podmirivanju svojih obveza ili su kasnili manje od tjedan dana te kod kojih nije bilo potraživanja smatraju se dobrim klijentima [21]. Važno je napomenuti da dobri klijenti moraju zadržati svoj status tijekom cijelog vremenskog razdoblja, dok se loš klijent može definirati u bilo kojem trenutku, dostizanjem određene faze neplaćanja. Ostaje pitanje klijenata koji nisu ni dobri ni loši. Takvi se klijenti nazivaju neodređeni. Općenito su to klijenti koji često kasne s plaćanjem, ali nikada ne više od postavljene granice dana

kašnjenja te na kraju ispune svoje obveze. Ukoliko je broj neodređenih klijenata vrlo mali, tada se oni mogu isključiti ili mogu biti dodijeljeni dobroj klasifikaciji.

2.1.4 Segmentacija portfelja

U većini slučajeva korištenje nekoliko skor-kartica omogućuje bolju procjenu rizika od korištenja jedne kartice. To je obično slučaj u velikim portfeljima gdje postoji mnogo različitih podskupova i gdje jedna skor-kartica neće učinkovito funkcionirati za sve. Proces identifikacije tih podskupova naziva se segmentacija [21]. Dva su glavna načina na koja se segmentacija izvršava [21].

1. Generiranje ideja o segmentaciji na temelju operativnog, iskustvenog i industrijskog znanja, a zatim potvrđivanje tih ideja pomoću analitike.

2. Generiranje jedinstvenih segmenata korištenjem statističkih tehnika kao što su klasteriranje ili stabla odlučivanja.

Uobičajeno se segmentacija provodi kako bi se identificirao optimalan skup segmenata koji maksimizira statističku izvedbu. Segmentacija temeljena na iskustvu uključuje ideje generirane iz poslovnog znanja i iskustva. Općenito se područja segmentacije koja se koriste u industriji temelje na demografiji, vrsti proizvoda, izvoru poslovanja, dostupnosti podataka i vrsti podnositelja zahtjeva. Nakon što se ideje o segmentaciji generiraju, potrebno je provesti daljnju analizu kako bi se one empirijskim putem potvrdile. Kod statistički utemeljene segmentacije najčešće se koristi klasteriranje, o kojem je nešto više rečeno u poglavlju 1.

S obzirom da korištenje nekoliko skor-kartica omogućuje bolju procjenu rizika od korištenja jedne kartice, pitamo se zašto netko ne bi implementirao sve izgrađene skor-kartice? Ovdje se javlja pitanje troška. Naime, trošak razvoja, trošak implementacije, trošak obrade podataka, razvoj strategije i praćenje te veličina segmentacije čine veliki trošak kod manjih portfelja. Kod velikih portfelja ti troškovi mogu biti opravdani u usporedbi s koristima.

2.1.5 Metodologija izrade modela kreditnog skoriranja

Kod modeliranja skor-kartice potrebno je prvo postaviti varijable i veze između varijabli za koje se smatra da utječu na rizik neplaćanja. Zatim je potrebno upotrijebiti neke statističke procedure kojima se varijable dodaju, odnosno oduzimaju iz modela pri čemu se kod svakog koraka mjeri poboljšanje prediktivnosti modela [25]. Postoje razne matematičke tehnike za izgradnju skor-kartice, a izbor najprikladnije tehnike ovisi o kvaliteti dostupnih podataka, vrsti željenih ishoda, veličini uzorka, implementaciji, interpretabilnosti rezultata i sl. Dobrim odabirom tehnike razultati skor-kartice će se moći ispravno protumačiti. Neke od tehnika mogu se pogledati u [15]. Većina metoda procjene kreditnog rizika temelji se na tradicionalnim statističkim metodama poput logističke regresije [23], k-*means* algoritma [11], klasifikacijskih stabla odlučivanja [10] ili modela neuronskih mreža [27], kao i analize klastera koju ćemo i sami obraditi.

2.1.6 Pregled plana provedbe

Kako bi se osigurala realna očekivanja, u ovom se koraku planovi testiranja i implementacija trebaju pregledati. Kod implementacije skoring modela potrebno je postaviti granične vrijednosti. Također, potrebno je provoditi testiranja kako bi se osiguralo da su skor-kartica i ostali parametri postavljeni korektno. Osim toga, potrebno je arhivirati sve podatke te je

potrebna efikasna komunikacija između odjela gdje se zahtjevi za kredit procesuiraju i odjela marketinga.

2.2 Prednosti i nedostaci kredit skoring modela

Izgradnjom i korištenjem kredit skoring modela ostvaruju se mnoge prednosti, kako za kreditore tako i za klijente. Prema [1] prednosti za kreditore su:

- preciznije procjenjivanje rizika;
- kvalitetnije donošenje odluka;
- povećanje brzine odlučivanja, a time i broja kredita koje kreditori mogu pregledati;
- mogućnost kreiranja kreditnih programa za ciljanu populaciju.

Prednosti za mala poduzeća su [1]:

- povećana dostupnost kredita malim poduzetnicima;
- cijena kredita određena je prema očekivanom riziku;
- skraćeno vrijeme postupka odobravanja i realizacije kredita.

Dodatno, prema [3] glavne prednosti kredit skoring modela su objektivnost, konzistentnost, jednostavnost, laka interpretacija te uobičajena i shvatljiva metodologija. S druge strane, kao nedostaci kredit skoring modela navode se [3]:

- slabo eliminiranje pristranosti procesa nastalih u prošlosti;
- automatizacija samo postojeće kreditne prakse banke;
- loša prediktivnost ukoliko dođe do promjene populacije u odnosi na originalnu populaciju prema kojoj je model dizajniran.

2.3 Klaster analiza u kreditnom skoriranju

Kao što je ranije navedeno, uporaba različitih metoda za izradu modela kreditnog skoriranja, uključujući statističke, matematičke, modele strojnog učenja i druge, poboljšava prediktivnost modela. Korištenje više tehnika za obradu podataka donosi veliku prednost jer takvi modeli mogu bolje predvidjeti rizik vezan uz odobravanje kredita za svakog klijenta, nego korištenje svake metode zasebno. U tu svrhu, pogledajmo nekoliko istraživanja koja opisuju primjenu klaster analize u kreditnom skoriranju u kombinaciji s drugim metodama.

Glavna ideja u istraživanjima procjene kreditnog rizika sastoji se od izgradnje pravila klasifikacije koja ispravno dijele klijente banaka na dobre i loše. U radu [28] predlaže se sustav koji se temelji na kombinaciji nenadziranih i nadziranih klasifikacija. Točnije, istražuje se kombinacija klaster analize i modela stabla odlučivanja. U prvom koraku, korištenjem algoritma klasteriranja, klijenti su segmentirani u klastere sličnih karakteristika. Zatim su, u drugom koraku, za svaku grupu izrađena stabla odlučivanja i definirana su pravila klasifikacije za svaku grupu klijenata uzimajući u obzir različite atribute. Istraživanje je provedeno na skupovima podataka o kreditnom riziku retail klijenata iz njemačkih i japanskih banaka. U prvom slučaju analizirani su klijenti njemačke banke koji su segmentirani u četiri klastera.

Prvi klaster čine prilično mladi ljudi s velikim iznosom kredita i dugim rokom otplate. Drugi klaster čine osobe srednje dobi s prosječnim iznosom kredita i prosječnim rokom otplate. Skup mladih, s niskim iznosom kredita i relativno kratkim rokom otplate čine treći klaster i posljednji, četvrti klaster čini skupina starijih osoba s prosječnim iznosom kredita i prosječnim rokom otplate. U drugom su slučaju analizirani klijenti japanskih banaka. I u ovom su slučaju klijenti klasterirani u četiri klastera. Prvi klaster čine poprilično mladi ljudi s prosječnim stanjem na računu, zaposleni na duže vrijeme s dugoročnom otplatom kredita. Klijente svrstane u treći klaster opisuje kratak rok rada i kratak rok otplate kredita. Drugi i četvrti klaster sadrže podatke poprilično starih klijenata. Klijenti raspoređeni u drugi klaster dobro su situirani, zaposleni na duže vrijeme s dugim rokom otplate kredita. Klijenti raspoređeni u četvrti klaster imaju prilično nisko stanje na računu, nisku stopu plaćanja i kratak rok otplate. Konačno, potvrđena je mogućnost povezivanja nenadziranih i nadziranih tehnika za procjenu kreditnog rizika. Rezultati dobiveni na skupovima podataka o kreditnom riziku pokazali su veću preciznost i jednostavnost pravila dobivenih za svaki klaster, nego za pravila povezana s cijelim skupom podataka.

Kreditno skoriranje vrlo je popularna i često korištena tehnika za analizu kreditnog rizika, međutim jedan model kreditnog skoriranja možda neće moći generirati zajedničko pravilo za klasifikaciju klijenata. Zbog toga se ono često kombinira s drugim tehnikama. U radu [2] predložena je segmentacija klijenata pomoću k-*means* algoritma. Točnije, [2] govori o segmentaciji klijenata koja se provodi pomoću k-*means* algoritma, a za svaki je segment razvijen poseban model kredit skoringa. Osim toga, za svaki je segment odabran zaseban prag kako bi se ostvario minimalni relativni trošak pogrešne klasifikacije. Podaci za ovo istraživanje uključuju informacije o kreditima izdanim retail klijentima između 2005. i 2014. godine. Cilj toga rada je segmentirati klijente u različite segmente na temelju njihovih sličnosti i modelirati svaki segment zasebno. Postavljanjem pragova za klasifikaciju na temelju segmentiranih modela, željela se postići veća preciznost u identifikaciji rizika i smanjiti trošak pogrešne klasifikacije. Korištenje k-*means* algoritma rezultiralo je segmentiranjem podataka u pet klastera. Drugi i četvrti klaster okarakterizirani su vrlo visokom stopom neplaćanja s visokim kamatama i niskim kreditnim rezultatima. Prvi i peti klaster imaju niske stope neispunjenja obveza i visoke kreditne rezultate koji sugeriraju nizak rizik neplaćanja. Treći klaster ima umjereni rizik neplaćanja te visoke kreditne rezultate.

Kako je fokus ove studije trošak pogrešne klasifikacije, za odabir najboljeg praga primjenjuje se pristup relativnog troška klasifikacije. S obzirom da jedan prag možda neće točno klasificirati kredite s različitim razinama rizika, dodjeljivanje zasebnih pragova za kredite sa sličnim rizicima može povećati preciznost u klasifikaciji. Kako bi se odabrao prag koji daje najniži relativni trošak pogrešne klasifikacije, odabrano je 20 nasumičnih pragova. Optimizacija praga provedena je sa šest različitih omjera troškova za potpune podatke i segmente sa svim razvijenim modelima. Kao najbolji model odabran je onaj koji daje najniži trošak pri svakom omjeru troškova. Uz smanjenje relativnog troška pogrešne klasifikacije, segmentirano modeliranje bilo je uspješno u smanjenju relativnog troška. Stoga, poboljšanje relativnog troška pogrešne klasifikacije putem segmentiranog modeliranja u odnosu na jedan model bodovanja opravdava primjenjivost segmentiranog modeliranja i odabira praga. Rezultati su pokazali da svaki segment ima drugačije ponašanje prema riziku, stoga optimalan prag varira ovisno o riziku. Nadalje, pojedinačnim odabirom praga za segmente došlo je do poboljšanja relativnog troška pogrešne klasifikacije u usporedbi s relativnim troškom za jedan model kredit skoringa. Rezultati su pokazali važnost primjene segmentiranog modeliranja i individualnog odabira praga za bolje odluke i uštede ulaganja.

3 Empirijski dio: Primjena klaster analize u kreditnom skoriranju

3.1 Opis podataka i varijabli

U ovom poglavlju provest ćemo ranije opisanu klaster analizu na stvarnim podacima jedne banke. Uzorak s kojim ćemo raditi sastoji se od 6414 opažanja opisanih s 25 varijabli, pri čemu je polovica podataka okarakterizirana kao dobra poduzeća, odnosno poduzeća koja ne kasne u plaćanju, a druga polovica kao loša poduzeća, odnosno poduzeća koja kasne u plaćanju.

Općenito, baze podataka sadrže jedno ili više neuobičajenih opažanja koji se čine kao da ne pripadaju bazi podataka. Takva opažanja nazivamo *stršeće vrijednosti* (*engl. outliers*). Stršećim vrijednostima smatramo one vrijednosti koje su u usporedbi sa ostalim podacim u bazi ili vrlo visoke ili vrlo niske. U našoj bazi postojao je veliki broj takvih vrijednosti. Neke su varijable imale po nekoliko tisuća outliera, stoga takve varijable nisu uključene u analizu. Varijable koje smo uključili u analizu opisane su u Tablici 3. S obzirom da su i druge varijable imale veći broj outliera, njihovo se uklanjanje nije pokazalo dobrim odabirom jer se tada baza podataka smanjila na svega 616 podataka. Iz tog razloga smo se, umjesto brisanja outliera, odlučili zamijeniti ih medijanom. Nakon postupka zamjene outliera medijanom došlo smo do 6008 podataka što će biti konačan broj podataka s kojima ćemo dalje raditi. U tim podacima nalazi se 3106 dobrih poduzeća, odnosno onih koja ne kasne u plaćanju i 2902 loših poduzeća, odnosno onih poduzeća koja kasne u plaćanju. Deskriptivna statistika varijabli s kojima ćemo dalje raditi dana je u Tablici 4.

Priprema podataka za klaster analizu ponekad zahtjeva neku vrstu transformacije podataka. Ukoliko su podaci zadani s više obilježja ta se obilježja mogu međusobno značajno razlikovati. Primjerice, ako promatramo radni staž klijenta koji potražuje kredit, koji dolazi iz segmenta $[2, 15]$ i neto iznos plaće koji dolazi iz segmenta $[3585, 6285]$, za očekivati je da obilježja nisu ravnopravna te je prije primjene neke od metoda za klasteriranje podatke potrebno standardizirati. Kako varijable s većim vrijednostima ne bi dominirale nad onim varijablama s manjim vrijednostima, standardizirali smo podatke.

U praksi se za klasteriranje podataka koriste razne statističke metode i pristupi. U ovom radu, prilikom kreiranja klastera, korištena je kombinacija hijerarhijskog i nehijerarhijskog klasteriranja, odnosno Wardova metoda i *k-means* algoritam o čijim ćemo rezultatima nešto više reći u nastavku. Razlika koju treba istaknuti je da *k-means* algoritam zahtjeva od korisnika da definira broj klastera koje treba stvoriti, dok hijerarhijsko klasteriranje to ne čini. Iz tog razloga, opišimo prvo metodu hijerarhijskog klasteriranja.

Naziv varijable	Naziv pokazatelja/objašnjenje	Izračun	Kategorija
KoefTekuceLikvidnosti	Koeficijent tekuće likvidnosti	kratkotrajna imovina/kratkoročne obveze	Pokaztelji likvidnosti
KoefUbrzaneLikvidnosti	Koeficijent ubrzane likvidnosti	(kratkotrajna imovina - zalihe)/ kratkoročne obveze	
KoefTrenutneLikvidnosti	Koeficijent trenutne likvidnosti	novac/kratkoročne obveze	
KoefObrtaUkupneImovine	Koeficijent obrta ukupne imovine	ukupni prihodi/imovina	Pokazatelji aktivnosti
KoefObrtaKratkotrajneImovine	Koeficijent obrta kratkotrajne imovine	ukupni prihodi/kratkotrajna imovina	
ProdajaObrtniKapital	Prodaja prema obrtnom kapitalu	prihodi od prodaje/(kratkotrajna imovina - kratkoročne obveze)	
KoefZaduzenosti	Koeficijent zaduženosti	(dugoročne obveze + kratkoročne obveze)/imovina	Pokazatelji zaduženosti
OdnosDugaIKapitala	Odnos duga i kapitala	(dugoročne obveze + kratkoročne obveze)/kapital	
KoefVlastitogFinanciranja	Koeficijent vlastitog financiranja	kapital/imovina	
UkKreditiImovina	Ukupni krediti prema imovini	ukupni krediti/imovina	
ObvezeEBITDA	Ukupne obveze prema EBITDA	ukupne obveze/EBITDA	
EkonomicnostUkPoslovanja	Ekonomičnost ukupnog poslovanja	ukupni prihodi/ukupni rashodi	Pokazatelji ekonomičnosti
NetoRentabilnostImovine	Neto rentabilnost imovine	dobit ili gubitak razdoblja/imovina*100	Pokazatelji profitabilnosti
NetoMarzaProfita	Neto marža profita	dobit ili gubitak razdoblja/ukupni prohodi*100	
ZadrzanaDImovina	Zadržana dobit prema prodaju	zadržana dobit ili gubitak/imovina	
DobarLos	Je li poduzeće dobro ili loše	Dobar − ne kasni u plaćanju Loš - kasni u plaćanju	Ostale varijable

Tablica 3: Opis varijabli

Naziv varijable	Minimum	Donji kvantil	Medijan	Aritmetička sredina	Gornji kvantil	Maksimum
Koeficijent tekuće likvidnosti	0	0.2515	1	1.1750	1.4593	6.2325
Koeficijent ubrzane likvidnosti	0	0.1487	0.7	0.9195	1.1707	5.3886
Koeficijent trenutne likvidnosti	0	0.001243	0.023435	0.12535	0.108728	1.191753
Koeficijent obrta ukupne imovine	0	0.03972	0.7	1.01493	1.44311	5.63835
Koeficijent obrta kratkotrajne imovine	0	0.1623	1.2	1.6861	2.3253	9.3589
Prodaja prema obrtnom kapitalu	-3.51962	0	0	0.30517	0.04164	5.25705
Koeficijent zaduženosti	0	0.388	0.8693	0-855	1	3.6184
Odnos duga i kapitala	-5.48756	-1.0334	0.1	-0.05827	0.54029	5.45153
Koeficijent vlastitog financiranja	-2.619073	-0.004546	0.104905	0.127102	0.578174	2.020005
Ukupni krediti prema imovini	0	0	0.0263	0.2869	0.4366	2.0896
Ukupne obveze prema EBITDA	-21.6256	-0.7955	0	0.1943	1.6613	19.2815
Ekonomičnost ukupnog poslovanja	0	0.3868	1	0.7821	1.0582	2.1421
Neto rentabilnost imovine	-57.5183	-1.0662	0	0.7673	4.2366	48.6080
Neto marža profita	-78.4525	0	0	0.4508	5.2039	58.6442
Zadržana dobit prema prodaji	-2.33440	-0.06849	0	-0.03606	0.16976	1.68

Tablica 4: Deskriptivna statistika

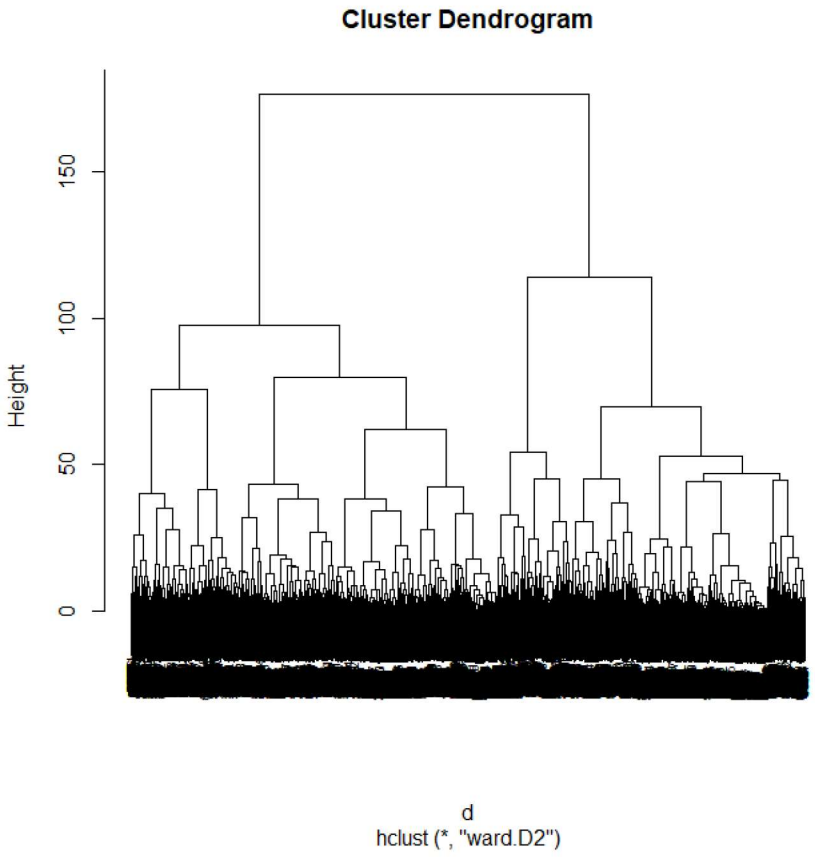
3.2 Hijerarhijsko klasteriranje

Nakon što smo odabrali varijable s kojima ćemo ići u modeliranje, sljedeći korak je generiranje klastera. Prvo ćemo upotrijebiti metodu hijerarhijskog klasteriranja, jer za sada ne znamo u koliko klastera treba klasterirati podatke. S obzirom da ne znamo koja će hijerarhijska metoda dati najjaču strukturu klasteriranja izračunali smo nekoliko keficijenata aglomeracije, čije su vrijednosti prikazane u Tablici 5. Vidimo kako Wardova metoda daje najveći koeficijent aglomeracije, odnosno ona ima najjaču strukturu klasteriranja, stoga ćemo je koristiti kao metodu za konačno hijerarhijsko klasteriranje.

average	single	complete	ward
0.8843333	0.7793288	0.9235692	**0.9936974**

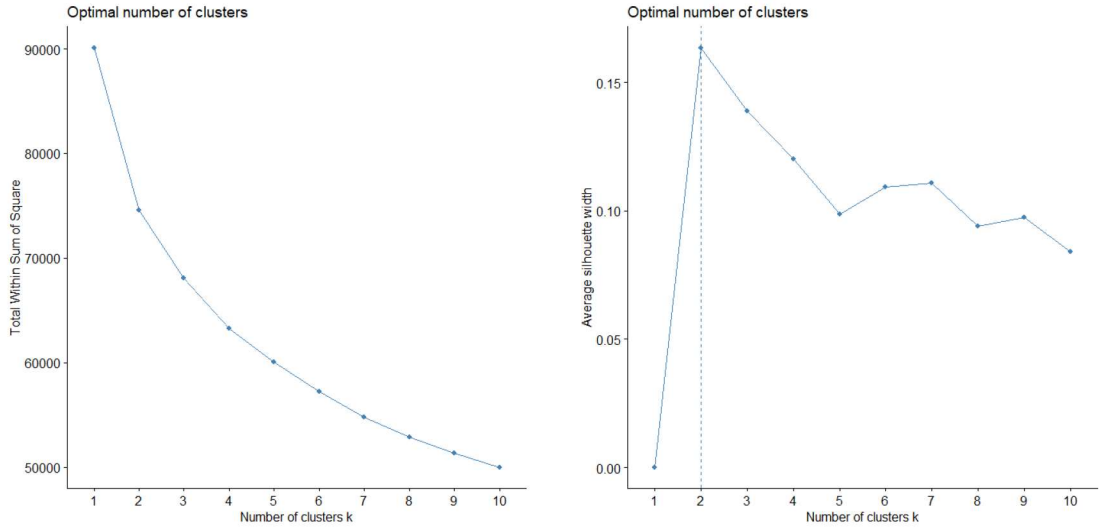
Tablica 5: Koeficijenti aglomeracije

Pomoću Wardove metode, temeljene na euklidskoj udaljenosti načinjen je dendogram koji prikazuje kako funkcionira ova metoda klasteriranja. Možemo uočiti kako se opažanja klasteriraju počevši od parova pojedinačnih opažanja koja su najbliže jedan drugome i spajanjem manjih grupa u veće, ovisno o tome koje su grupe najbliže jedna drugoj. Na kraju se svi naši podaci spajaju u jedan segment.



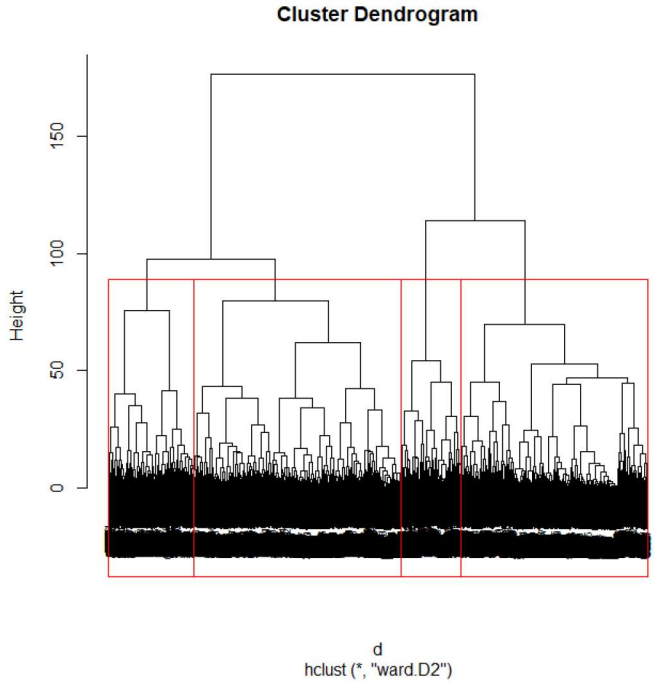
Slika 4: Dendogram

S obzirom da u ovom trenutku ne znamo ništa o primjerenom broju klastera, provest ćemo neke od metoda za pronalaženje primjerene vrijednosti k. Za donošenje najbolje odluke oko izbora broja klastera potrebno je kombinirati više metoda, a mi smo kombinirali metodu lakta i metodu siluete. Vizualni prikaz obiju metoda dan je na Slici 5.



(a) Metoda lakta (b) Metoda prosječne siluete

Slika 5: Grafičko određivanje broja klastera



Slika 6: Dendogram s 4 klastera

U ovom radu odlučili smo podatke klasterirati u 4 klastera, što je na dendogramu (vidi Slika 6) i prikazano. Slika 6 prikazuje podjelu baze podataka u 4 klastera. Klasteri redom sadrže 948, 2314, 2080 i 666 podataka, ali ne prikazuju koji od ta 4 klastera sadrži financijski zdrava, a koji financijski nezdrava mala i srednja poduzeća. Kako bismo bolje shvatili

strukturu svakog klaster, napravili smo njihovu analizu. U Tablici 6 dana je frekvencija dobrih i loših poduzća po klasteru, a Tablica 7 prikazuje prosjek i standardnu devijaciju svih financijskih pokazatelja na temelju naše kasifikacije, koja je 6008 poduzeća klasterirala u 4 klastera.

	Klaster 1	Klaster 2	Klaster 3	Klaster 4
Dobar	665	1507	663	271
Loš	283	807	1417	395

Tablica 6: Frekvencija dobrih i loših poduzeća po klasterima

Varijabla	Klaster 1	Klaster 2	Klaster 3	Klaster 4
Koeficijent tekuće likvidnosti	3.0431 (1.6032)	1.0918 (0.7758)	0.6409 (0.7475)	0.4735 (0.5117)
Koeficijent ubrzane likvidnosti	2.3833 (1.4299)	0.8539 (0.7369)	0.5074 (0.6916)	0.3514 (0.4197)
Koeficijent trenutne likvidnosti	0.3653 (0.3812)	0.1139 (0.2054)	0.0431 (0.1024)	0.0805 (0.1483)
Koeficijent obrta ukupne imovine	1.1182 (0.9434)	1.4095 (1.3086)	0.5752 (1.0257)	0.8702 (1.3015)
Koeficijent obrta kratkotrajne imovine	1.5279 (1.1609)	2.4681 (2.1893)	0.9127 (1.4815)	1.6097 (2.1241)
Prodaja prema obrtnom kapitalu	1.5336 (1.5569)	0.3595 (1.3563)	-0.1189 (0.8418)	-0.3079 (0.9101)
Koeficijent zaduženosti	0.4322 (0.3350)	0.6393 (0.3605)	0.8829 (0.4283)	2.1185 (0.6759)
Odnos duga i kapitala	0.4315 (0.8583)	0.8056 (1.5086)	-0.5180 (1.5252)	-2.3211 (1.0431)
Koeficijent vlastitog financiranja	0.5337 0.3531	0.3439 (0.3584)	0.1046 (0.4271)	-1.1348 (0.6782)
Ukupni krediti prema imovini	0.1511 (0.2936)	0.1902 (0.2721)	0.2845 (0.4246)	0.8232 (0.6709)
Ukupne obveze prema EBITDA	1.2051 (3.8056)	3.1961 (4.7708)	-3.3033 (5.1894)	-0.7506 (6.5387)
Ekonomičnost ukupnog poslovanja	1.0294 (0.4079)	1.0978 (0.2309)	0.3893 (0.3902)	0.5598 (0.4376)
Neto rentabilnost imovine	7.4293 (15.3283)	7.5246 (12.0569)	-7.5908 (13.6205)	-6.0905 (16.4348)
Neto marža profita	8.2817 (17.7995)	7.3809 (13.5429)	-9.3995 (19.3785)	-4.0105 (16.0536)
Zadržana dobit prema prodaji	0.1972 (0.3911)	0.0896 (0.3199)	-0.0421 (0.5015)	-0.7857 (0.6939)

Tablica 7: Srednja vrijednost i standardna devijacija financijskih pokazatelja po klasterima

Prva grupa pokazatelja, odnosno koeficijenti tekuće, ubrzane i trenutne likvidnosti su koeficijenti koji predstavljaju pokazatelje likvidnosti. Pokazatelji likvidnosti mjere sposobnost poduzeća u podmirivanju svojih kratkoročnih obveza [24, 26].

Koeficijent tekuće likvidnosti ocjenjuje sposobnost poduzeća da podmiri svoje tekuće obveze. Općenito, vrijednost koeficijenta trebala bi iznositi najmanje 2 kako poduzeće ne bi imalo problema pri podmirivanju svojih tekućih obveza. Iz Tablice 7 možemo primijetiti kako je prosječna vrijednost koeficijenta tekuće likvidnosti veća od 2 samo u prvom klasteru, dok u ostalim klasterima ta vrijednost opada, ali svugdje je pozitivna. Poduzeća koja se nalaze u prvom klasteru ne bi trebala imati problema u podmirivanju svojih tekućih obveza, jer u prosjeku imaju 3 puta više kratkotrajne imovine nego kratkoročnih obveza. Kako poduzeća u ostalim klasterima imaju prosječnu vrijednost koeficijenta manju od preporučene granice, možemo reći kako kod tih poduzeća postoje potencijalni problemi u podmirivanju tekućih obveza.

Koeficijent ubrzane likvidnosti ocjenjuje sposobnost poduzeća da udovolji svojim kratkoročim obvezama upotrebom svoje najlikvidnije imovine. Preporučena vrijednost koeficijenta ubrzane likvidnosti je 1, a iz Tablice 7 vidimo kako jedino poduzeća koja se nalaze u prvom klasteru imaju prosječnu vrijednost koeficijenta veću od donje granice. Poduzeća u ostalim klasterima imaju prosječnu vrijednost koeficijenta manju od 1, što znači da u prosjeku nemaju dovoljno kratkoročnih sredstava da podmire svoje dospjele obveze, a bez prodaje zaliha.

Koeficijent trenutne likvidnosti u omjer stavlja novac i kratkoročne obveze, odnosno pokazuje je li poduzeće trenutno sposobno podmiriti svoje kratkoročne obveze. Iz prosječne vrijednosti koeficijenta vidimo kako poduzeća koja su dodijeljena prvom klasteru u prosjeku mogu podmiriti 36.53% ukupnih kratkoročnih obveza, dok je taj postotak u ostalim klaster znatno niži. U trećem klasteru prosječni koeficijent je najniži, odnosno poduzeća iz tog klastera u prosjeku mogu podmiriti svega 4.31% ukupnih kratkoročnih obveza.

Druga grupa pokazatelja, odnosno koeficijenti obrta ukupne imovina i obrta kratkotrajne imovine pokazatelji su aktivnosti poduzeća koji pokazuju koliko brzo cirkulira imovina u poslovnom procesu. Općenito, što je koeficijent obrta veći, vrijeme vezivanja je kraće [24, 26].

Koeficijent obrta ukupne imovine u omjer stavlja ukupni prihod i imovinu, odnosno pokazuje koliko je puta godišnje poduzeće obrnulo svoju ukupnu imovinu da ostvari jednu novčanu jedinicu prihoda. Iz Tablice 7 vidimo kako poduzeća koja se nalaze u drugom klasteru u prosjeku najfunkcionalnije koriste imovinu jer je prosječni koeficijent obrta za taj klaster najveći. Poduzeća koja se nalaze u drugom klasteru u prosjeku su ostvarila 1.4095 novčanih jedinica prihoda na jednu novčanu jedinicu ukupne imovine, dok je kod poduzeća koja se nalaze u prvom klasteru dobit u prosjeku 1.1182 novčanih jedinica na jednu novčanu jedinicu ukupne imovine.

Koeficijent obrta kratkotrajne imovine u omjer stavlja ukupne prihode i kratkotrajnu imovinu, odnosno pokazuje koliko je puta godišnje poduzeće obrnulo svoju kratkotrajnu imovinu da ostvari jednu novčanu jedinicu prihoda. Kao i ranije, poduzeća koja pripadaju drugom klasteru imaju najveći prosječni koeficijent obrta kratkotrajne imovine, odnosno ta su poduzeća u prosjeku kratkotrajnu imovinu obrnuli više od 2 puta. Najmanji prosječni prihod ostvarila su poduzeća koja se nalaze u trećem klasteru. Oni su u prosjeku ostvarili 0.9127 prihoda na jednu novčanu jedinicu kratkotrajne imovine.

Iduću grupu pokazatelja čine pokazatelji zaduženosti koji pokazuju strukturu kapitala i iz kojih se izvora poduzeća financiraju te jesu li poduzeća prezadužena ili se mogu još zadužiti [24, 26].

Koeficijent zaduženosti u omjer stavlja ukupne obveze i ukupnu imovinu te pokazuje do koje se mjere poduzeće koristi zaduživanjem kao oblikom financiranja. Što je koeficijent veći to je poduzeće zaduženije. Općenito, koeficijent zaduženosti ne bi trebao prelaziti 0.5. Iz Tablice 7 vidimo kako jedino poduzeća koja se nalaze u prvom klasteru imaju prosječni koeficijent zaduženosti niži od 0.5, odnosno ta se poduzeća više financiraju iz vlastitih nego iz tuđih izvora. S druge strane, poduzeća koja su dodjeljena ostalim klasterima imaju koeficijent zaduženosti u prosjeku znatno veći od preporučene granice. Poduzeća koja se nalaze u četvrtom klasteru imaju koeficijent zaduženosti u prosjeku 2.1185, odnosno možemo reći kako ta poduzeća imaju previsoku zaduženost.

Suprotni koeficijent koeficijentu zaduženosti jest koeficijent vlastitog financiranja. On u omjer stavlja ukupni kapital i ukupnu imovinu te se preporuča da koeficijent vlastitog financiranja bude veći od 0.5. Usporedimo li koeficijente iz Tablice 7 vidimo kako najveći koeficijent vlastitog financiranja imaju poduzeća koja se nalaze u prvom klasteru, a u prosjeku ta vrijednost iznosi 0.5337. Poduzeća koja se nalaze u ostalim klasterima imaju prosječnu vrijednost koeficijenta manju od preporučene granice, dok poduzeća iz četvrtog klastera imaju negativan koeficijent. Razlog tome je to što je za ta poduzeća koeficijent zaduženosti vrlo visok pa je posljedično koeficijent vlastitog financiranja vrlo nizak.

Jedini pokazatelj ekonomičnosti u našoj bazi podataka je ekonomičnost ukupnog poslovanja koji u omjer stavlja ukupne prihode i ukupne rashode. Ovaj pokazatelj pokazuje za koliko su posto veći ukupni prihodi od ukupnih rashoda, a cilj je da bude veći od 1 [24, 26]. Ukoliko pogledamo prosječnu vrijednost pokazatelja ekonomičnosti po klasterima možemo zaključiti kako poduzeća koja pripadaju prvom i drugom klasteru ostvaruju više ukupnih prihoda po jedinici ukupnih rashoda s obzirom da im je prosječna vrijednost koeficijenta veća od 1. S druge strane, poduzeća koja se nalaze u trećem i četvrtom klasteru ostvaruju gubitke jer je prosječna vrijednost njihovog koeficijenta ekonomičnosti manja od 1.

Zadnju grupu pokazatelja čine pokazatelji profitablinosti koji mjere sposobnost poduzeća da ostvari određenu razinu dobiti u odnosu na prihode, imovinu ili kapital. Pokazatelji profitablinosti koje promatramo su neto rentabilnost imovine, neto marža profita i zadržana dobit prema prodaji. Oni su za poduzeća koja se nalaze u trećem i četvrtom klasteru negativna, odnosno svi promatrani pokazatelji profitabilnosti za poduzeća koja se nalaze u tim klasterima su u prosjeku negativni, što znači da su ta poduzeća u gubitku [24, 26].

Neto marža profita je pokazatelj koji iskazuje koliko posto ukupnog prihoda zadržava poduzeće u obliku neto dobiti uvećane za rashode od kamata. Promatrajući prosječne vrijednosti neto marže profita za svaki klaster vidimo kako poduzeća koja pripadaju prvom klasteru u prosjeku imaju neto dobiti uvećane za rashode od kamata 8.28% u odnosu na ukupni prihod, dok je prosječna dobit poduzeća iz drugog klastera 7.38%.

Neto rentabilnost imovine, odnosno povrat na ukupnu imovinu poduzeća u omjer stavlja čistu neto dobit sa ukupnom imovinom. Srednja vrijednost neto marže profita za poduzeća iz prvog klastera iznosi 7.43 što znači da je prosječni povrat ukupne imovine tih poduzeća 7.43%. Za poduzeća iz drugog klastera taj je povrat nešto veći i iznosi 7.52%.

Opišimo klastere s obzirom na prosječne vrijednosti financijskih pokazatelja iz Tablice 7.

Klaster 1 sadrži 948 podataka, pri čemu je 665 poduzeća okarakterizirano kao dobra odnosno koja ne kasne u plaćanju, a 283 poduzeća kao loša odnosno koja kasne u plaćanju. U ovom se klasteru nalaze najzdravija poduzeća, odnosno poduzeća čiji su prosječni financijski pokazatelji u najvećoj mjeri dobri. Poduzeća koja se nalaze u ovom klasteru su najlikvidnija te nemaju poteškoća pri podmirivanju svojih kratkoročnih obveza, odnosno rizik neplaćanja je najmanji. Imaju dovoljno kratkoročnih sredstava da podmire dospjele obveze, nisu zadužena te se u najvećoj mjeri financiraju iz vlastitih izvora. Imaju vrlo dobre pokazatelje aktivnosti i u odnosu na druge klastere posluju s dobiti, a njihova poslovna dobit je u prosjeku najveća.

Klaster 2 sadrži 2314 poduzeća, pri čemu je 1507 poduzeća okarakterizirano kao dobra odnosno koja ne kasne u plaćanju, a 807 poduzeća kao loša odnosno koja kasne u plaćanju. U ovom se klasteru nalaze zdrava poduzeća, ali lošija od poduzeća iz Klastera 1. Poduzeća su likvidna, a s obzirom na niže prosječne vrijednosti koeficijenata postoji potencijalna opasnost od nemogućnosti podmirivanja kratkoročnih obveza. Poduzeća iz ovog klastera imaju najbolje pokazatelje aktivnosti. Koeficijenti obrta su najveći, odnosno prosječno trajanje obrta je najkraće. Poduzeća su zadužena, imaju manji udio vlastitog kapitala, ali posluju s dobiti.

Klaster 3 sadrži 2080 podataka, pri čemu je 663 poduzeća okarakterizirano kao dobra odnosno koja ne kasne u plaćanju, a 1417 poduzeća kao loša odnosno koja kasne u plaćanju. Poduzeća koja se nalaze u ovom klasteru nisu likvidna i ne mogu podmiriti svoje kratkoročne obveze, stoga je rizik neplaćanja velik. Zadužena su i imaju mali udio vlastitog kapitala te se velikim dijelom financiraju iz tuđih izvora. Loše koriste svoju imovinu, koeficijenti obrta su mali, odnosno prosječno trajanje obrta je dugo. Poduzeća posluju s najvećim gubitkom, stvaraju manje ukupnih prihoda po jedinici ukupnih rashoda. Takav je rezultat očekivan, s obzirom da je velika razlika između dobrih i loših poduzeća u ovom klasteru.

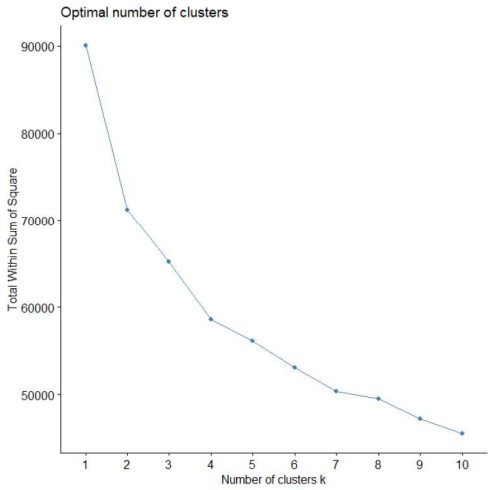
Klaster 4 sadrži 666 podataka, pri čemu je 271 poduzeća okarakterizirano kao dobra odnosno koja ne kasne u plaćanju, a 395 poduzeća kao loša odnosno koja kasne u plaćanju. Poduzeća iz ovog klastera su najmanje likvidna, nisu sposobna podmiriti svoje kratkoročne obveze i rizik neplaćanja za ova poduzeća je najveći. Imaju previsoku zaduženost, u velikoj se mjeri financiraju iz tuđih izvora i obveze su im veće od imovine. Prosječne vrijednosti koeficijenata obrta su niske, odnosno prosječno trajanje obrta je dugo, ali ne duže od poduzeća iz Klastera 3. Poduzeća posluju s gubitkom i prihodi su manji od rashoda.

Iz opisa klastera vidimo kako se svaki klaster po nečemu razlikuje od drugog klastera. Najvidljivija razlika između klastera je veličina svakog klastera, a nakon toga i udio dobrih, odnosno loših poduzeća unutar svakog klastera. Klaster 1 sadrži najlikvidnija i najprofitabilnija poduzeća koja nisu zadužena. Klaster 2 sadrži najaktivnija poduzeća, ali poduzeća imaju problem sa zaduživanjem. Klaster 3 sadrži najmanje financijski zdrava poduzeća, odnosno ona poduzeća koja posluju s gubitkom, koja ne mogu podmiriti svoje obveze i koja imaju loše pokazatelje aktivnosti. Klasteru 4 su također dodijeljena financijski nezdrava poduzeća, ali su pokazatelji aktivnosti bolji od pokazatelja aktivnosti poduzeća iz Klastera 3. S obzirom na to, možemo reći kako su Klasterima 1 i 2 dodijeljena financijski zdrava poduzeća, dok su Klasterima 3 i 4 dodijeljena financijski manje zdrava, odnosno nezdrava poduzeća.

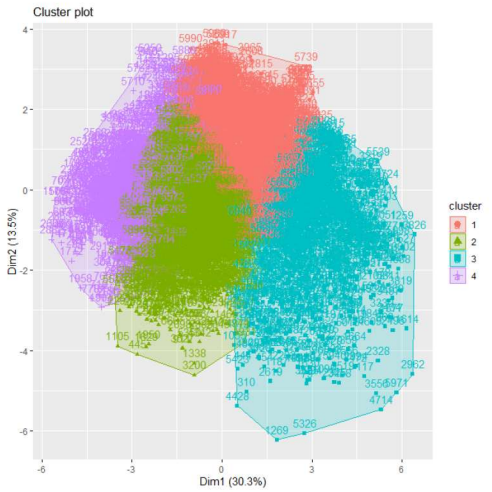
Sada se za svaki klaster može razviti poseban model kreditnog bodovanja. U skladu s time, financijski zdrava mala i srednja poduzeća mogla bi posuditi više novca od banaka po nižim kamatama uz niže zahtjeve za kolateralom, zbog nižeg rizika neplaćanja. S druge strane, mala i srednja poduzeća koja su u lošem financijskom stanju morala bi plaćati više kamate i imali bi nižu gornju granicu zaduživanja uz veće zahtjeve za kolateralom. Na ovaj način banke bi mogle smanjiti iznos nenaplativih zajmova danih malim i srednjim poduzećima, što bi dovelo do poboljšanja kreditne sposobnosti financijskog sustava i pomoglo bi zdravim malim i srednjim poduzećima da dobiju bolje uvjete kreditiranja.

3.3 *k-means* algoritam

S obzirom da se kod *k-means* algoritma broj klastera mora postaviti prije pokretanja algoritma, nakon nekoliko različitih vrijednosti za k odlučili smo se kao i kod hijerarhijske metode za četiri klastera čija je vizualizacija dana na Slici 8. Da je četiri optimalan broj klastera potvrdili smo koristeći tzv. metodu lakta prikazanu na Slici 7.



Slika 7: Metoda lakta



Slika 8: Vizualizacija klastera

Nakon što smo izvršili konačnu analizu klastera potvrdili smo rezultate koje smo dobili pomoću hijerarhijske metode.

4 Zaključak

Mala i srednja poduzeća temelj su zapošljavanja i razvoja privrede. Međutim, često ne uspijevaju dobiti potrebna financijska sredstva na financijskim tržištima. Zbog toga su istraživanja, razvoj i inovacije najvažniji za održivu uspješnost malih i srednjih poduzeća. Potaknuti time, u ovom smo radu odlučili provesti klaster analizu na stvarnim podacima jedne banke. Prije provedbe klaster analize upoznali smo se s njezinom teorijskom podlogom te smo opisali metode za klasteriranje i kreditno skoriranje. Nakon toga smo napravili klaster analizu čiji je cilj grupirati poduzeća tako da se u svakoj grupi odnosno klasteru nalaze poduzeća s istim ili sličnim karakteristikama. Takav način grupiranja podataka otvara mogućnost izgradnje modela skoriranja za svaki klaster posebno čime bi se mogla napraviti bolja procjena kreditnog rizika. S obzirom da financijski izvještaji pružaju informacije o uspješnosti poslovanja poduzeća, bitno ih je analizirati te na temelju dobivenih rezultata donijeti zaključke. Provedbom klaster analize na danoj bazi podataka dobili smo 4 klastera. Kako nismo znali koji od klastera sadrži financijski zdrava, a koji financijski nezdrava mala i srednja poduzeća napravili smo analizu svakog klastera posebno. Nakon analize financijskih pokazatelja za svaki klaster zaključili smo kako su u dva klastera smještena financijski zdrava poduzeća, a u preostala dva klastera smještena su financijski nezdrava, odnosno financijski manje zdrava poduzeća. Klasteriranje malih i srednjih poduzeća na temelju njihovih financijskih pokazatelja jača njihovu financijsku strukturu jer će financijski zdrava poduzeća, zbog nižeg rizika neplaćanja, dobiti bolje uvjete kreditiranja. Jačanje financijskog kapaciteta malih i srednjih poduzeća rezultira nastankom velikih i jakih poduzeća, a samim time dolazi i do ekonomskog razvoja.

Literatura

[1] Z. Bohaček, N. Šarlija, M. Benšić, Upotreba kredit skoring modela za ocjenjivanje kreditne sposobnosti malih poduzetnika, Ekonomski pregled, 54(7-8), str. 565-580, 2003.

[2] A. Byanjankar, Improving Credit Risk Analysis with Cluster Based Modeling and Threshold Selection, Proceedings of the 53rd Hawaii International Conference on System Sciences, 1413-1420, 2020.

[3] J.B. Caouette, E.I. Altman, P. Narayanan, Managing Credit Risk, John Wiley & Sons, New York, 1998.

[4] C.-H. Chou, M.-C. Su, and Eugene Lai. 2004. A new cluster validity measure and its application to image compression. Pattern Analysis and Applications 7, 2 (2004), 205–220.

[5] Determining The Optimal Number Of Clusters: 3 Must Know Methods. https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/

[6] K. Devčić, I. Tonković Pražić, Ž. Župan, Klaster analiza: primjena u marketinškim istraživanjima, stručni rad, Veleučilište Nikola tesla u Gospiću, Gospić, 2012.

[7] B. Dokić, Stirlingovi brojevi, Osječki matematički list, 165-187, 2013.

[8] M. Doumpos, C. Lemonakis, D. NIklis, C. Zopoundis, Analytical Techniques in the Assessment of Credit Risk: An Overview of Methodologies and Applications, Springer, Chania, Greece, June 2018.

[9] G. Gan, C. Ma, J. Wu, Data Clustering: Theory, Algorithms and Applications, SIAM, Philadelphia, 2007.

[10] R.H. Gavis, D.B. Edelman, A.J. Gammerman, Machine learning algorithms for credit-card application, IMA Journal of Management Mathematics, 4, 1992, 43-51.

[11] W.E. Henley, D.E. Hand, Construction of a k-nearst neighbor credit-scoring system, IMA Journal of Management Mathematics, 8, 1997, 305-321.

[12] A. Huang, Similarity Measures for Text Document Clustering, NZCSRSC 2008, Christchurch, New Zeland, April 2008.

[13] Ž. Kiš, Klaster analiza i njega primjena u bankarstvu, diplomski rad, Odjel za matematiku, Osijek, 2012.

[14] K-Mean: Getting The Optimal Number Of Clusters. https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/

[15] Lj. Kvesić, Statističke metode u upravljanju kreditnim rizikom, Sveučilište u Mostaru, Fakultet prirodoslovno-matematičkih i odgojnih znanosti, 2012.

[16] G. Liu, A New Index for Clustering Evaluation Based on Density Estimation, Tsinghua University, 2022.

[17] B. Markić, Neuronska mrežna klasifikacija u menadžerskom računovodstvu, Informatol, 2011., 44 (3), 200-206.

[18] K. Sabo, R. Scitovski, I. Vazler, Grupiranje podataka: klasteri, Osječki matematički list, 149-178, 2010.

[19] R. Scitovski, M. Briš Alić, Grupiranje podataka, Sveučilište J.J. Strossmayera u Osijeku, Odjel za matematiku, Osijek, 2016.

[20] S. Scitovski, N. Šarlija, Cluster analysis in retail segmentation for credit skoring, CRORR 5(2014), 235-245.

[21] N. Siddiqi, Intelligent Credit skoring: Building and Implementing Better Credit Risk skorecards, 2an Edition, John Wiley & Sons, Inc., Hoboken, New Jersey, 2016.

[22] Silhouette Method — Better than Elbow Method to find Optimal Clusters. https://towardsdatascience.com/silhouette-method-better-than-elbow-method-to-find-optimal-clusters-378d62ff6891

[23] A. Steenackers, M.J. Goovaerts, A credit scoring model for personal loans, Insurance Mathematics & Economics, 8, 1989, 31-34.

[24] N. Šarlija, Predavanja za kolegij "Analiza poslovanja poduzeća", Odjel za matematiku, Sveučilište J.J. Strossmayera u Osijeku, 2009.

[25] N. Šarlija, Predavanja za kolegij "Upravljanje kreditnim rizicima", Odjel za matematiku, Sveučilište J.J. Strossmayera u Osijeku, 2008.

[26] D. Vukojević, Analiza poslovne uspješnosti na primjeru društva Kraš d.d., diplomski rad, RRiF Visoka škola za financijski menadžment, Zagreb, 2018.

[27] D. West, Neural network credit scoring models, Computers & Operations Research, 27, 2000, 1131-1152.

[28] D. Zakrezewska, On integrating unsupervised and supervised classification for credit risk evulation, Information technology and control, 36, 98-102, 2007.

[29] Što je bodovanje i kako funkcionira. Bonitetna ocjena i stručna ocjena kreditne sposobnosti dužnika Kreditni rejting i bodovni model. https://spmost.ru/hr/banks/chto-takoe-skoring-i-kak-on-rabotaet-kreditnyi-skoring-i-ekspertnaya-ocenka/

Sažetak

U ovom se radu upoznajemo s klaster analizom i njezinom primjenom u kreditnom skoriranju. Na početku je dana teorijska podloga klaster analize te su obrađene metode za klasteriranje. Opisano je kreditno skoriranje te je dan osvrt na klaster analizu u kreditnom skoriranju. Zatim je provedeno empirijsko istraživanje na stvarnim podacima jedne banke. Financijski pokazatelji malih i srednjih poduzeća su klasterirani u klastere i opisani su dobiveni rezultati.

Ključne riječi: klaster analiza, klasteriranje, klaster, kreditno skoriranje, financijski pokazatelj

Cluster analysis in credit scoring

Summary

In this work, we are introduced to cluster analysis and its application in credit scoring. The theoretical background of the cluster analysis is presented at the beginning as well as the methods used in clustering. There is a description of credit scoring and an overview of cluster analysis used within it. Then there is an empirical research done on real data provided by a bank. Financial indicators of small and medium businesses are clustered and the final results are described.

Keywords: cluster analysis, clustering, cluster, credit scoring, financial indicator

Životopis

Rođena sam 3. studenoga 1996. godine u Virovitici. Pohađala sam Osnovnu školu August Cesarec u Špišić Bukovici. Nakon završetka osnovne škole upisala sam opću gimnaziju u Gimnaziji Petra Preradovića u Virovitici. Srednjoškolsko obrazovanje završila sam 2015. godine te nakon toga upisala preddiplomski studij Matematike u Osijeku na Odjelu za matematiku. Naziv sveučilišne prvostupnice stekla sam 2019. godine s temom završnog rada "Hermitski adjungirani operator" pod mentorstvom doc. dr. sc. Suzane Miodragović. Nakon toga, nastavila sam studij na Odjelu za matematiku upisavši diplomski studij, smjer Financijska matematika i statistika.