Klasifikacija metodom K-srednjih vrijednosti kod predviđanja navika kupaca

Bubnjar, Veronika

SVEUČILIŠTE U ZAGREBU
FAKULTET STROJARSTVA I BRODOGRADNJE

ZAVRŠNI RAD

Veronika Bubnjar

Zagreb, 2022.

SVEUČILIŠTE U ZAGREBU
FAKULTET STROJARSTVA I BRODOGRADNJE

ZAVRŠNI RAD

Mentori:

Doc. dr. sc. Tomislav Stipančić, dipl. ing.

Student:

Veronika Bubnjar

Zagreb, 2022.

Izjavljujem da sam ovaj rad izradila samostalno koristeći znanja stečena tijekom studija i navedenu literaturu.

Zahvaljujem se mentoru prof. dr. sc. Tomislavu Stipančiću na pomoći i savjetima tijekom izrade ovog rada.

Također se zahvaljujem svojoj obitelji, rodbini, prijateljima i svom dečku na neizmjernoj podršci. Svaki teški trenutak oni su učinili ljepšim, boljim i zabavnim.

Veronika Bubnjar

SVEUČILIŠTE U ZAGREBU
FAKULTET STROJARSTVA I BRODOGRADNJE
Središnje povjerenstvo za završne i diplomske ispite
Povjerenstvo za završne i diplomske ispite studija strojarstva za smjerove:
proizvodno inženjerstvo, računalno inženjerstvo, industrijsko inženjerstvo i menadžment, inženjerstvo materijala i mehatronika i robotika

Sveučilište u Zagrebu Fakultet strojarstva i brodogradnje	
Datum	Prilog
Klasa: **602 – 04 / 22 – 6 / 1**	
Ur.broj: **15 - 1703 - 22 -**	

ZAVRŠNI ZADATAK

Student: **Veronika Bubnjar** JMBAG: **0035217917**

Naslov rada na hrvatskom jeziku: **Klasifikacija metodom K-srednjih vrijednosti kod predviđanja navika kupaca**

Naslov rada na engleskom jeziku: **Classification by the K-means method for customer habit predictions**

Opis zadatka:

U svijetu podataka i velike količine informacija algoritmi umjetne inteligencije postaju sve značajniji. U ovisnosti o vrti i strukturi podataka metode strojnog učenja koriste se prilikom predviđanja, odlučivanja, klasifikacije, prepoznavanja, itd.
Koristeći metodu K-srednjih vrijednosti strojnog učenja (eng. *K-means*) u radu je potrebno napraviti računalni model za predviđanje navika kupaca u trgovačkom centru. Rad model je potrebno temeljiti na znanom skupu podataka koji uključuje podatke o spolu, godinama i mjesečnim primanjima svojih mušterija. Za bolje rezultate predviđanja potrebno je odrediti odgovarajuće varijable modela te optimalan broj klastera. Rezultate rada modela potrebno je evaluirati temeljem standardnih evaluacijskih tehnika strojnog učenja (npr. matrice konfuzija). Osim toga, rezultate je potrebno objasniti koristeći tehnike za vizualizaciju informacija. U radu je potrebno navesti korištenu literaturu i eventualno dobivenu pomoć.

Zadatak zadan:	Datum predaje rada:	Predviđeni datumi obrane:
30. 11. 2021.	**1. rok:** 24. 2. 2022. **2. rok (izvanredni):** 6. 7. 2022. **3. rok:** 22. 9. 2022.	**1. rok:** 28. 2. – 4. 3. 2022. **2. rok (izvanredni):** 8. 7. 2022. **3. rok:** 26. 9. – 30. 9. 2022.
Zadatak zadao:		Predsjednik Povjerenstva:
Doc. dr. sc. Tomislav Stipančić		Prof. dr. sc. Branko Bauer

SADRŽAJ

POPIS SLIKA

POPIS OZNAKA

Oznaka	Jedinica	Opis
D	-	Neoznačeni skup primjera
x	-	Značajka, s indeksom *i* označava primjer
y	-	Oznaka
J	-	Funkcija pogreške
b	-	Binarna indikatorska varijabla
K	-	Broj grupa
μ	-	Primjer, s indeksom *k* označava centroid
P	-	vjerojatnost
R	-	Randov indeks
a	-	Broj jednako označenih parova u istim grupama
b	-	Broj različito označenih parova u različitim grupama

SAŽETAK

U završnom radu napravljen je i dokumentiran računalni model koristeći metodu K-srednjih vrijednosti strojnog učenja. Računalni model je napravljen u cilju predviđanja navika kupaca u trgovačkom centru. Temeljen je na zadanoj bazi podataka s određenim karakteristikama kao što su rod, dob, godišnja primanja i navika potrošnje od kojih je potrebno neke uzeti u obzir za određivanje broja grupa, tj. grupiranja. Uz odabir odgovarajućih varijabla dobiven je optimalan broj grupa odnosno klastera. Svi rezultati su određeni odgovarajućom standardnom metodom, metodom koljena. Koristeći tehnike vizualizacije, rezultati su prikazani i objašnjeni.

Ključne riječi:

Strojno učenje, K-srednje vrijednosti, grupiranje, grupe(klasteri), metoda koljena

SUMMARY

In the final paper, a computer model was created and documented using the K-means method of machine learning. A computer model was created to classify the habits of customers in a shopping center by K-means method. It is based on a given database with certain characteristics such as gender, age, annual income, and spending score, some of which need to be taken into algorithm to determine the number of clusters, i.e., clustering. With the selection of appropriate variables, the optimal number of clusters was obtained. All results were determined by the appropriate standard method, the elbow method. The results are presented and explained using visualization techniques.

Key words:

Machine learning, K-means method, clustering, clusters, elbow method

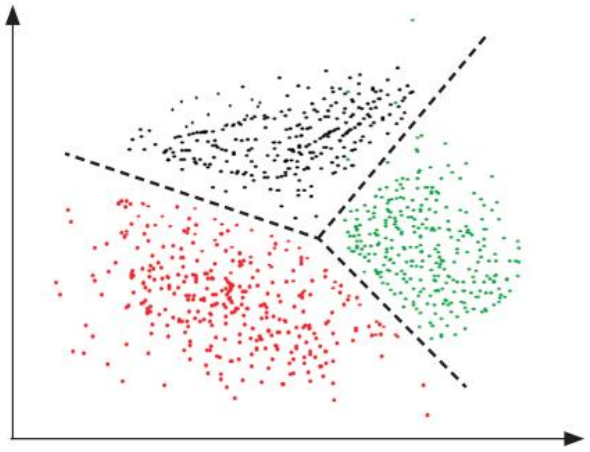
1. UVOD

Kako bi tema ovog rada bila jasnija, važno je definirati neke osnovne pojmove strojnog učenja i objasniti što to jest.

Strojno učenje grana je umjetne inteligencije koja se bavi oblikovanjem algoritama koji svoju učinkovitost poboljšavaju na temelju empirijskih podataka. Strojno učenje jedno je od danas najaktivnijih i najuzbudljivijih područja računarske znanosti, ponajviše zbog brojnih mogućnosti primjene koje se protežu od raspoznavanja uzoraka i dubinske analize podataka do robotike, računalnog vida, bioinformatike i računalne lingvistike. Dva su osnovna pristupa strojnom učenju: nadzirano učenje (klasifikacija i regresija) i nenadzirano učenje (grupiranje). [1]

1.1. Nenadzirano strojno učenje

Nenadzirano učenje pronalazi skrivene uzorke ili intrinzične strukture u podatcima. Primjenjuju se za izvođenje zaključaka iz skupova podataka koje čine samo ulazni podatci bez poznavanja odzivnih podataka (izlaza). Njime se grupiraju uzorci ili otkrivaju strukture, slika 1. Najčešća tehnika nenadziranog učenja je grupiranje (engl. clustering), pri čemu se traže skriveni obrasci ili grupe. Primjenu nalazi pri analizi sekvenca gena, analizi tržišta i prepoznavanju objekta. [2]



Slika 1. K-means algoritam najjednostavniji je algoritam grupiranja.[2]

1.2. Grupiranje

Grupiranje (engl. clustering) jest postupak razdjeljivanja primjera u grupe (engl. clusters), tako da slični primjeri (slični po nekom svojstvu) budu svrstani u istu grupu, a različiti primjeri u različite grupe. Svrha grupiranja jest nalaženje "prirodnih" (intrinzičnih) grupa u skupu neoznačenih podataka. Postoje razni algoritmi grupiranja. Jedan od jednostavnijih je algoritam K-srednjih vrijednosti poznatije K-means clustering koji će se bolje objasniti u slijedećim poglavljima. Naime to je algoritam koji je korišten za izradu računalnog modela koji će poslužiti za temu ovog rada, tj. za predviđanje navika kupaca upravo tom metodom.

Osnovna motivacija za nenadzirano učenje jest što u mnogim slučajevima analize podataka jednostavno nemamo informaciju o tome koji primjer pripada kojoj klasi. To znači da umjesto skupa označenih primjera koji se može prikazati jednadžbom 1:

$$D = \left\{(x^{(i)}, y^{(i)})\right\}_{i=1}^{N} \tag{1.}$$

zapravo raspolažemo skupom neoznačenih primjera koji glasi po jednadžbi 2:

$$D = \left\{x^{(i)}\right\}_{i=1}^{N} \tag{2.}$$

Primjeri mogu biti neoznačeni iz dva razloga: (1) ne znamo ih označiti (ne znamo unaprijed koje klase postoje) ili (2) znamo koje klase postoje, ali označavanje je teško izvedivo (npr. preskupo je). Evo nekih zadataka u kojima bismo tipično koristili nenadzirano strojno učenje:

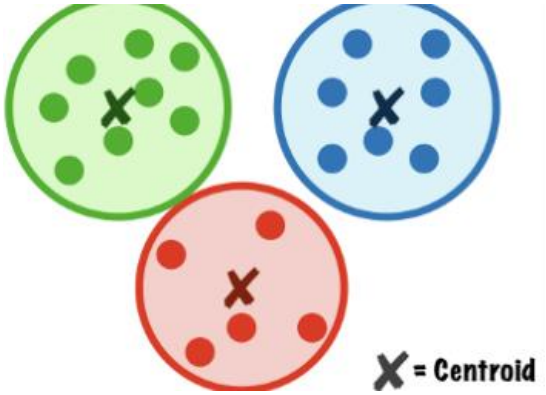
otkrivanje napada korisnika na mreži (engl. intrusion detection) – ovo je primjer otkrivanja vrijednosti koje odskaču (engl. outlier detection), klasifikacija objekata na fotografiji (engl. content-based image retrieval), grupiranje gena sa sličnom izražajnošću (funkcionalnošću), grupiranje tekstova prema autorima, grupiranje novinskih članaka prema temama, te grupiranje klijenata prema ponašanju (segmentacija korisnika).

Postoje različite vrste grupiranja. Prva podjela odnosi se na to generira li algoritam grupe primjera koje imaju nekakvu internu strukturu (podgrupe primjera) ili generira "plošne" grupe koje nemaju nikakvu internu strukturu. Tako razlikujemo:

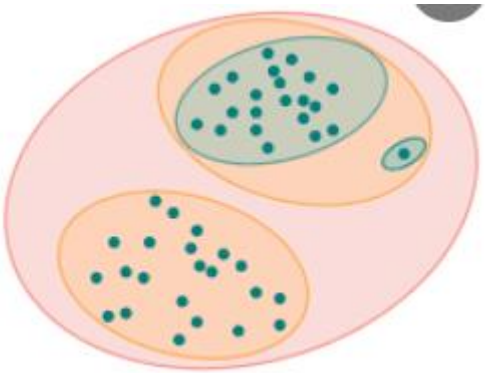
• Particijsko grupiranje – grupe su plošne (engl. flat), tj. ne postoje podgrupe;

• Hijerarhijsko grupiranje – grupe imaju podgrupe, koje imaju svoje podpodgrupe, i tako dalje, rekurzivno.

Drugim riječima, postupak rezultira hijerarhijom grupa. Na donjoj slici 2. prikazan je rezultat particijskog grupiranja ujedno i primjer grupiranja kojim će se bolje objasniti u slijedećim poglavljima, a na slici 3. rezultat hijerarhijskog grupiranja.[3]



Slika 2. Particijsko grupiranje[4]



Slika 3. Hijerarhijsko grupiranje[5]

2. RAČUNALNI MODEL ZA PREDVIĐANJE NAVIKA KUPACA U TRGOVAČKOM CENTRU

Grupiranje klijenata prema ponašanju je tema koja je bolje objašnjena u sljedećem poglavlju, odnosno klasifikacija metodom K-srednjih vrijednosti kod predviđanja navika kupaca.

2.1. Općenito o algoritmu K-srednjih vrijednosti

Najjednostavniji i najpoznatiji algoritam grupiranja jest algoritam k-srednjih vrijednsoti (engl. k-means algorithm). Algoritmom se primjeri iz neoznačenog skupa primjera $D = \{x^{(i)}\}_{i=1}^{N}$ grupiraju u K čvrstih grupa, gdje se parametar K (broj grupa) zadaje unaprijed. Bitno je da se ovdje radi o particijskom grupiranju, jer imamo čvrste grupe (svaki primjer pripada samo jednoj grupi).

Ideja algoritma jest da svaka grupa ima svoju srednju vrijednost (centroid) koja predstavlja grupu. Svaki primjer pripada grupi čiji mu je centroid najbliži (po euklidskoj udaljenosti). Postupak grupiranja je iterativan. Krenuvši od K slučajno odabranih sredina (centroida grupa), svi se primjeri svrstavaju u onu grupu čiji im je centroid najbliži. To može dovesti do pomaka centroida, stoga se u idućem koraku, nakon svrstavanja svih primjera u njima najbližu grupu, ponovno izračunavaju novi centroidi za svaku grupu. No, nakon što su se izračunali novi centroidi, to može dovesti do promjene u pripadanju primjera grupama, pa se primjeri ponovo svrstavaju u grupe tako da svaki primjer bude u grupi čiji im je centroid najbliži. To novo razvrstavanje primjera ponovo može dovesti do promjena centroida, pa se ponovno izračunavaju centroidi za svaku grupu, i tako dalje. Postupak ponavlja ova dva koraka (pridjeljivanje primjera grupama i izračun centroida) sve do konvergencije (dok nema promjene u pripadnosti primjera grupama, odnosno dok nema promjene u centroidima grupa).

Pogledajmo to sada formalno. Mnogi algoritmi grupiranja mogu se formalizirati (i izvesti) tako da se definira funkcija pogreške koju minimiziraju. Ta funkcija se u kontekstu algoritama grupiranja često naziva kriterijska funkcija. Označit ćemo je sa *J*. Za algoritam K-srednjih vrijednosti, kriterijska je funkcija definirana ovako:

$$J = \sum_{k=1}^{K} \sum_{i=1}^{N} b_k^{(i)} \left\| x^{(i)} - \mu_k \right\|^2 \qquad (3.)$$

Intuitivno, ova funkcija zbraja koliko primjeri unutar svake grupe odstupaju od centroida dotične grupe (imamo dvije sume, jedna ide po grupama, a druga po primjerima).

Preciznije, svaka grupa predstavljena je svojim centroidom, $\{\mu_k\}_{k=1}^{K}$. Oznaka $\|\cdot\|$ je *L2*-norma (tj. euklidska norma), definirana kao $\|x-\mu\|^2 = (x-\mu)^T(x-\mu)$, pa je dakle $\left\|x^{(i)}-\mu\right\|^2$ kvadrat euklidske udaljenosti izmedu primjera $x^{(i)}$ i μ. Vrijednost $b_j^{(i)}$ je binarna indikatorska varijabla koja indicira pripada li primjer $x^{(i)}$ grupi *k*: ako $b_k^{(i)}$= 1, onda primjer $x^{(i)}$ pripada grupi *k*, inače joj ne pripada. Prema tome, ukupna pogreška jednaka je zbroju, po svim primjerima kvadrata euklidske udaljenosti primjera $x^{(i)}$ od središta μ_k grupe u koju je taj primjeri svrstan. Cilj je pronaći ono grupiranje koje minimizira pogrešku. Grupiranje je definirano dvama parametrima: indikatorskim varijablama $b_k^{(i)}$ (one definiraju kojoj grupi pripada koji primjer) i centroidima grupa μ_k (oni definiraju gdje u prostoru primjera se grupe nalaze). Dakle, traže se parametri takvi da:

$$\underset{b_1,\ldots,b_K;\,\mu_1,\ldots,\mu_K}{\operatorname{argmin}} \quad J$$

Pogreška je veća što su primjeri dalje od centroida svoje grupe. To znači da, ako je namjera minimiziranje pogreške *J*, svaki primjer $x^{(i)}$ je potrebno svrstati u grupu čije je središte μ_k tom primjeru najbliže, to jest:

$$b_k^{(i)} = \begin{cases} 1 \text{ ako } k = \underset{j}{\operatorname{argmin}}\left\|x^{(i)}-\mu_j\right\|^2 \\ 0 \quad \text{inače} \end{cases} \tag{4.}$$

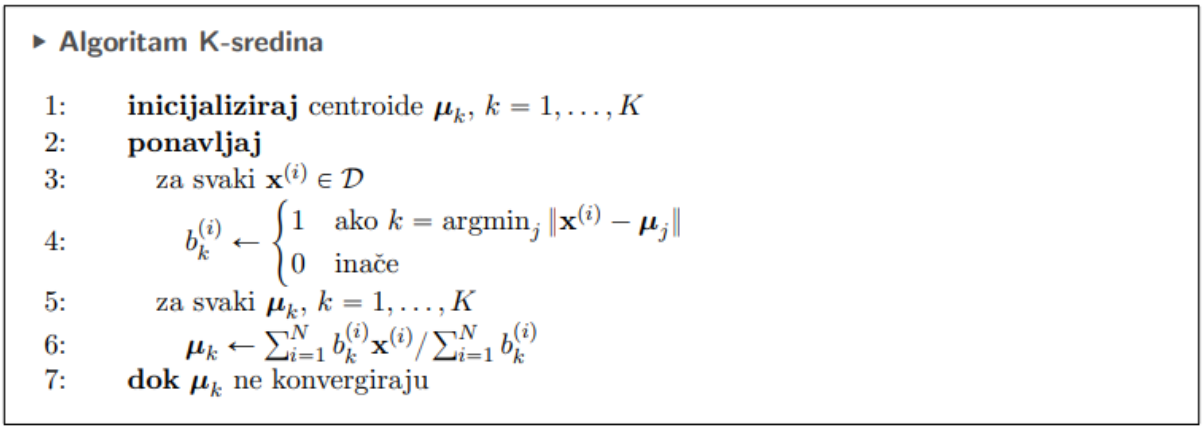
Algoritam K-srednjih vrijednosti radi tako da minimizira upravo kriterijsku funkciju *J*, ali to ne radi analitički jer ovaj optimizacijski problem nema rješenje u zatvorenoj formi. Naime, ako se pokuša derivirati funkciju *J* po oba parametra, nailazimo na problem jer su ti parametri međusobno ovisni: $b_k^{(i)}$ ovise o μ_k i obrnuto, μ_k ovise o $b_k^{(i)}$. Umjesto toga, algoritam K-srednjih vrijednosti optimizaciju provodi iterativno. Algoritam započinje sa slučajno odabranim srednjim vrijednostima μ_k. Zatim se u svakoj iteraciji temeljem izraza za $b_k^{(i)}$ za svaki primjer $x^{(i)}$ izraˇcunava vrijednost $b_k^{(i)}$, odnosno svaki se primjer pridjeljuje grupi ˇcijem je centroidu najbliˇzi. Nakon toga – budu´ci da sad imamo fiksirane vrijednosti $b_k^{(i)}$– moˇzemo izravno minimizirati izraz za kriterijsku funkciju J. Konkretno, postavljanjem $\nabla\mu_k$ *J* = 0 i rješavanjem po μ_k dobiva se:

$$2\sum_{i=1}^{N} b_k^{(i)}\left(x^{(i)}-\mu_k\right) = 0 \tag{5.}$$

iz čega slijedi

$$\mu_k = \frac{\sum_i b_k^{(i)} x^{(i)}}{\sum_i b_k^{(i)}} \tag{6.}$$

Vektor μ_k jednak je srednjoj vrijednosti vektora svih primjera koji su svrstani u grupu k. Budući da je ovime ostvarena promjena vektora μ_k u odnosu na njegovu prethodnu vrijednost, sada treba opet primijeniti izraz za $b_k^{(i)}$i ponovno izračunati koji primjeri pripadaju grupi k. Ova dva koraka ponavljaju se sve dok se ne dosegne stacionarno stanje, odnosno stanje u kojemu nema daljnjih promjena vrijednosti μ_k. kad bi se sve to iskombiniralo u algoritam, onda se može prikazati kao na slici 5.[3]



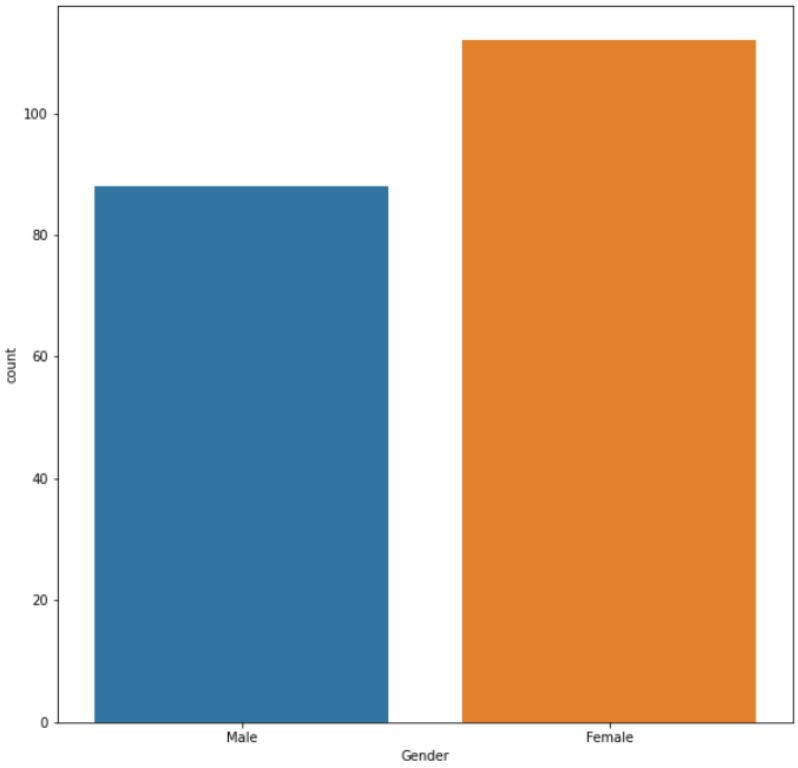
Slika 4. Algoritam K-srednjih vrijednosti[3]

2.2. Eksplorativna analiza baze podataka

Baza podataka koja se koristi u radu sastoji se od dvjesto kupaca, a sadrži neke osnovne podatke kao što je rod, dob, te konkretno bitne karakteristike za ovu problematiku kao što su godišnji prihodi i potrošnja kupaca.

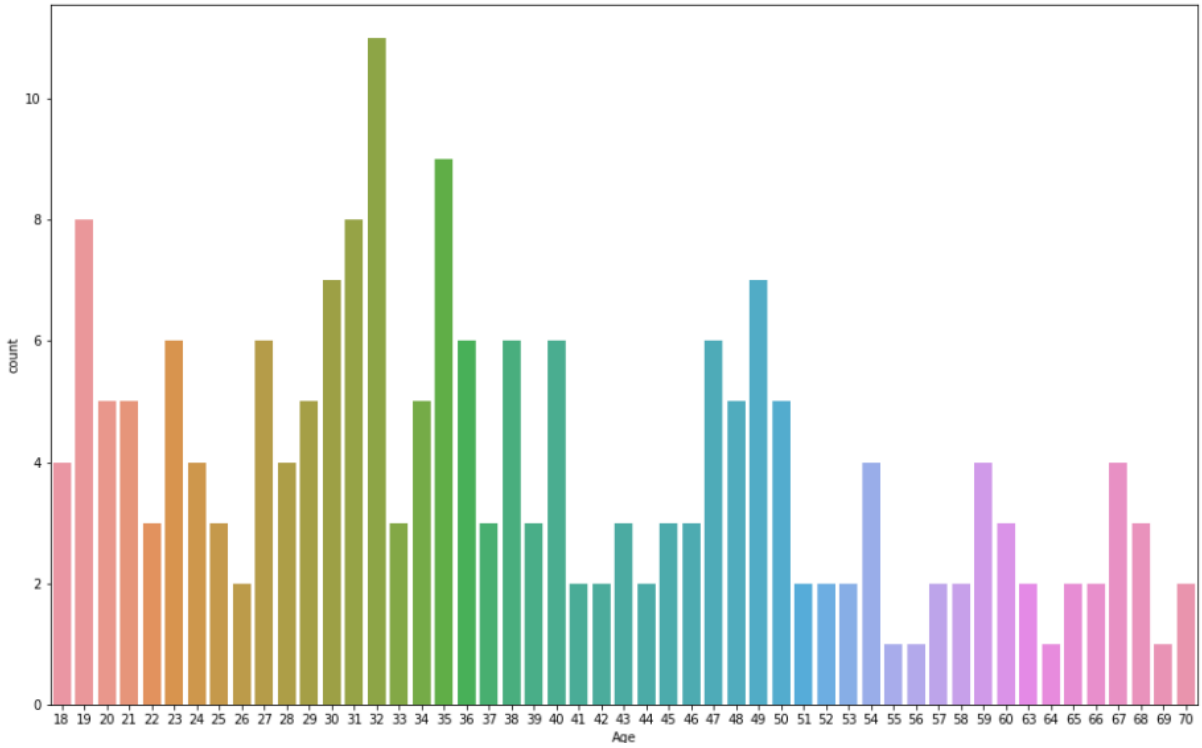
U cilju boljeg poslovanja trgovačkog centra, baza podataka je bitna kako bi se odredila i razumjela ciljana skupina kupaca koja pospješuje poslovanje. Tako se može bolje dati smisao i plan marketinškom timu da smisli efikasniju strategiju u skladu s navikama ciljane skupine.

Na slici 6. je prikazana podjela kupaca prema rodu. Grafikon obojen plavom bojom označava broj muškaraca, a grafikon obojen narančastom broj žena. Sada se iz slike može primijetiti da žene imaju veću tendenciju kupovanja od muškaraca.



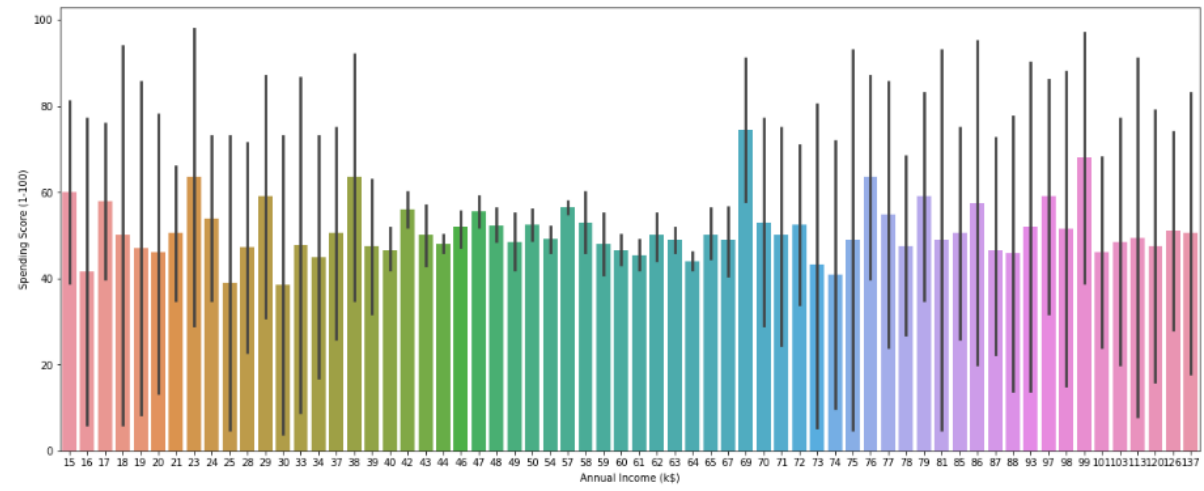
Slika 5. Rodna podjela kupaca[6]

Isto tako na slici 7. je prikazana dob grafičkom razdiobom gdje je na osi ordinata broj ljudi, a na os apscisa broj godina. Sada se da zaključiti da ljudi od trideset godina do srednjih tridesetih imaju najveću tendenciju kupovanja u odnosu na ostale. Također se može primijetiti kako porastom broja godina navika kupovanja opada.



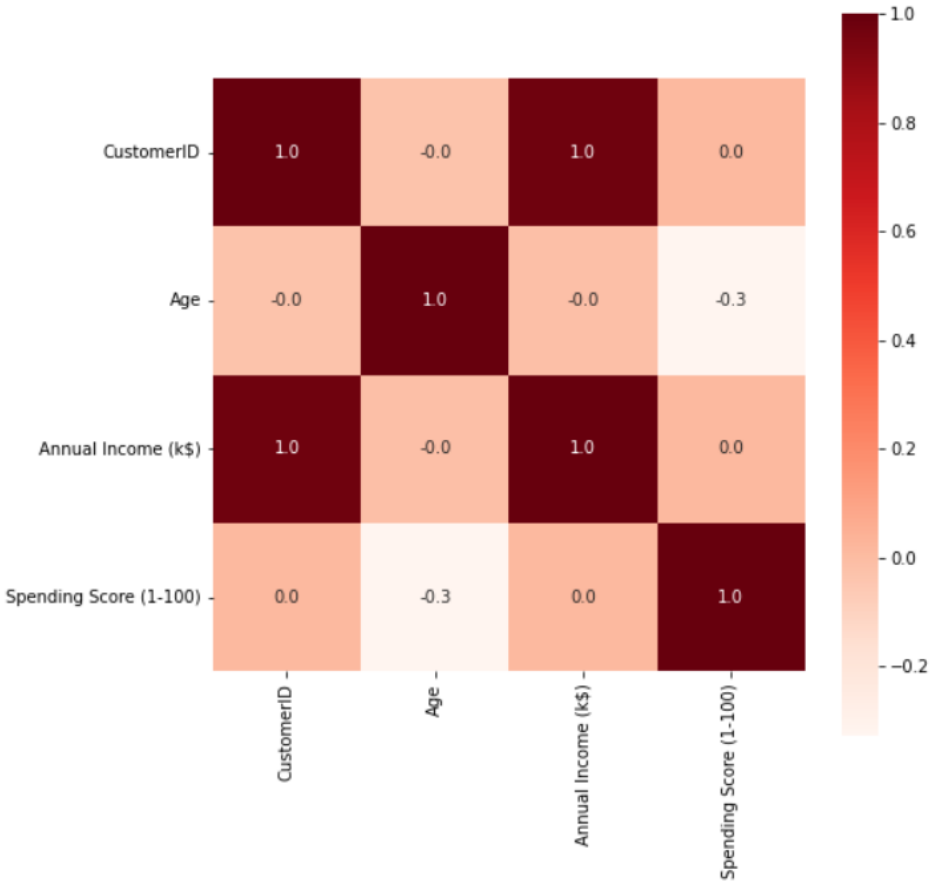
Slika 6. Dobna razdioba kupaca[6]

Zadnje dvije karakteristike, navika potrošnje i godišnja primanja, prikazani su grafikonima na slici 8. u ovisnosti jedan o drugom. Na osi ordinata su vrijednosti koji označavaju naviku potrošnje(definirana od 1 do 100), a na osi apscisa su označena godišnja primanja. Ono zanimljivo što se može zaključiti iz slike jest da ljudi sa niskim primanjima imaju sličnu naviku kao i ljudi s visokim primanjima, što govori da primanja nisu odlučujući faktor u navikama kupovanja. Ljudi sa srednjim primanjima se razlikuju od ova dva ekstrema na način da imaju izraženo manju tendenciju kupovanja i potrošnje.



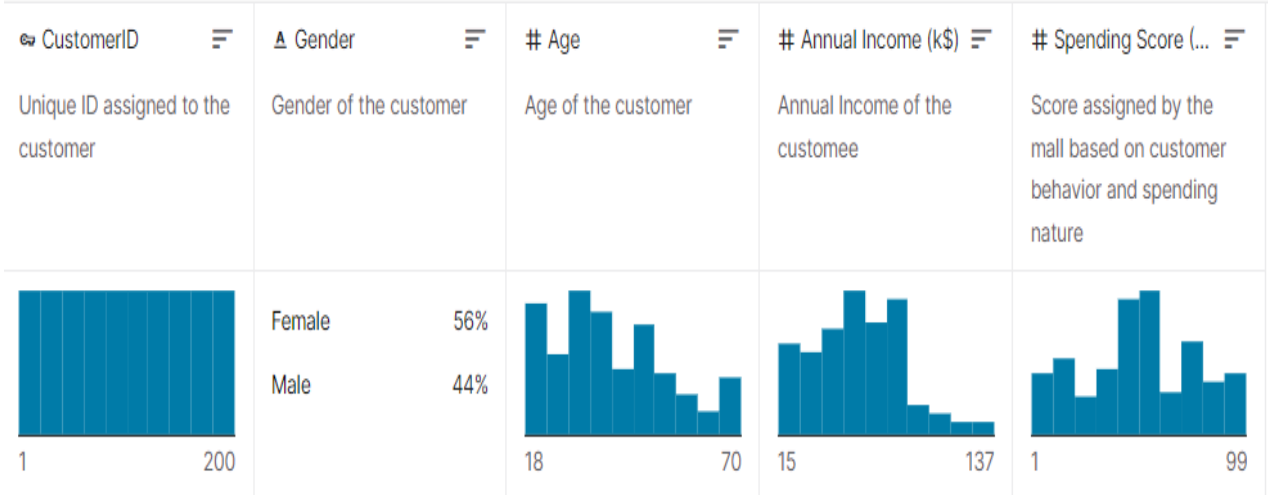
Slika 7. Navika potrošnje u ovisnosti o primanjima[6]

Na slijedećoj slici (slika 9.) je prikazana analiza korelacije svih karakteristika. Ovise li karakteristike jedna o drugoj je stavka koja se uzima u obzir jer je svaka značajka u algoritmu predstavljena kao varijabla, a ako su varijable međusobno zavisne, rješenje, odnosno model neće dati optimalno rješenje. Provodi se analiza kako bi bilo utvrđeno koliko je jaka međuzavisnost. Na konkretno ovom primjeru se može primijetiti da značajke nisu međusobno zavisne, tek se vidi negativna korelacija između dobi i navike potrošnje. Međutim taj iznos je dovoljno malen da se može zanemariti jer neće utjecati na model ako nema izražene korelacije.



Slika 8. Korelacija karakteristika[6]

Nakon eksplorativne analize podataka, sve dosad navedeno može se sažeti u jednu sliku i prikazati zajedno, broj ispitanika, postotak rodne skupine, dob ispitanika, godišnja primanja i navika potrošnje(slika 10.) .[6]



Slika 9. Podaci[6]

3. IZRADA MODELA

S obzirom da baza podataka pruža četiri značajke, nije nužno da se sve uzmu u obzir već je moguće odabrati.

3.1. K-means model s dvije značajke

U ovom modelu će se prikazati metoda i način rada algoritma s dvije značajke u ovom slučaju to su godišnja primanja i navika potrošnje.

Kako je spomenuto parametar K(broj grupa) se zadaje unaprijed. U ovom slučaju je to deset grupa.

Kako bi bilo sigurno da algoritam uvijek može dati iste izlaze za iste ulaze, unutar algoritma treba odabrati početna središta grupa.

Postoji više pristupa odabira početnih središta. Slučajno odabrati jedno početno središte μ_k , a zatim svako iduće središte odabrati tako da je što dalje od ostalih središta. Algoritam koji implementira ovakav pristup poznat je pod nazivom k-means++ i taj je algoritam korišten u ovom modelu. Kod tog algoritma je vjerojatnost da primjer $x^{(i)}$ bude odabran kao novo središte μ_{k+1} proporcionalna kvadratu udaljenosti tog primjera od njemu najbližeg, već odabranog središta μ_k :

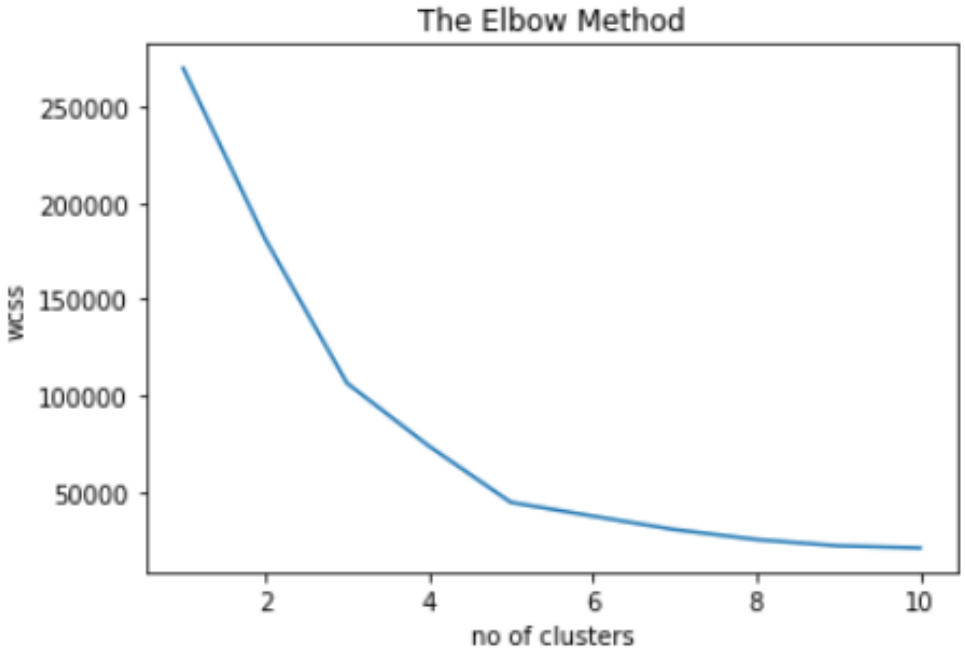
$$P\left(\mu_{k+1} = x^{(i)} \middle| D, \mu_{1,}, \ldots, \mu_k\right) = \frac{min_k \left\|\mu_k - x^{(i)}\right\|^2}{\sum_j min_k \left\|\mu_k - x^{(j)}\right\|^2} \qquad (7.)$$

Premda na ovaj način vrijednosti koje odskaču imaju veću vjerojatnost da budu odabrane za središte, njih je u pravilu manje, pa je ipak vjerojatniji odabir nekog od prosječnih primjera, koji su brojniji. Pokazano je da ovaj način odabira početnih središta znatno smanjuje pogrešku grupiranja, a također ubrzava konvergenciju algoritma. U slučajevima kada se početna središta određuju nedeterministički, npr. algoritam k-means++, običaj je algoritam K-srednjih vrijdenosti pokrenuti više puta i uzeti rezultat sa što manjom pogreškom grupiranja *J*.

Kod algoritma K-srednjih vrijednosti broj grupa, odnosno hiperparametar K, potrebno je odrediti unaprijed. To je slučaj kod mnogih algoritama grupiranja. Odabir optimalnog broja grupa dio je većeg problema koji se naziva provjera grupa (engl. cluster validation), a ono definira koliko je dobro grupiranje Odabir grupa jedan je od glavnih problema kod grupiranja. Idealno, broj grupa odgovarat će broju "prirodnih grupa" u skupu podataka, no taj broj je najčešće nepoznat.

Ovdje se za određivanje broja grupa koristi metoda koljena.

Metoda koljena (engl. elbow method). Grafički prikažemo ovisnost kriterijske funkcije o parametru K i tražimo "koljeno" krivulje (mjesto gdje funkcija najprije naglo pada i zatim stagnira ili pada vrlo sporo). S porastom broja grupa K, vrijednost kriterijske funkcije će padati, međutim taj pad općenito neće biti ujednačen. Naime, za neke vrijednosti K algoritam će početi razdjeljivati prirodne grupe. Ako povećanjem broja grupa kriterijska funkcija brzo opadne i onda neko vrijeme stagnira, to je signal da smo upravo "uhvatili" neko prirodno grupiranje u podatcima. Naime, ako s porastom K vrijednost *J* naglo opada, to znači da sa malim povećanjem broja grupa dobivamo znatno bolje grupiranje. U trenutku kada s povećanjem broja grupa K više ne dobivamo znatno bolje grupiranje, tj. kada *J* počne vrlo sporo opadati, to onda znači da novo dodane grupe ne hvataju baš neke prirodno grupirane primjere. Takvih koljena funkcije *J*(na slici *wcss*) može biti više, pogotovo ako grupe unutar sebe imaju podgrupe. Dakle, dobar odabir za broj grupa K jesu one vrijednosti koje odgovaraju točkama neposredno nakon koljena funkcije *J*, jer smo na tim mjestima upravo pogodili neki "prirodni" broj grupa. Npr., na slici 11. imamo jedno takvo koljeno iznad broja pet pa odabiremo broj pet kao optimalni broj grupa.[3]



Slika 10. Metoda koljena[7]

Vizualizacijom rezultata na slici 12. model se može objasniti.

- Grupa jedan označena crvenom bojom predstavlja skupinu s visokim primanjima, ali niskom navikom potrošnje

- Grupa dva označena plavom bojom obuhvaća skupinu s prosječnim primanjima i prosječnom navikom potrošnje
- Grupa tri zelene boje označava skupinu s visokim primanjima i visokom navikom potrošnje
- Grupa četiri svijetloplave boje predstavlja skupinu s niskim primanjima, ali visokom navikom potrošnje
- Grupa pet koja je ljubičaste boje predstavlja skupinu s niskim primanjima kao i niskom navikom potrošnje

Zaključak je da je ciljana skupina grupa tri. To je skupina koja ima visoka primanja i visoku naviku potrošnje, a jedna od strategija jest da takvoj skupini sustav za obavijesti o proizvodima šalje elektroničku poštu na dnevnoj bazi, dok ostalim grupama nije nužno tako često nego već na tjednoj ili mjesečnoj.[7]

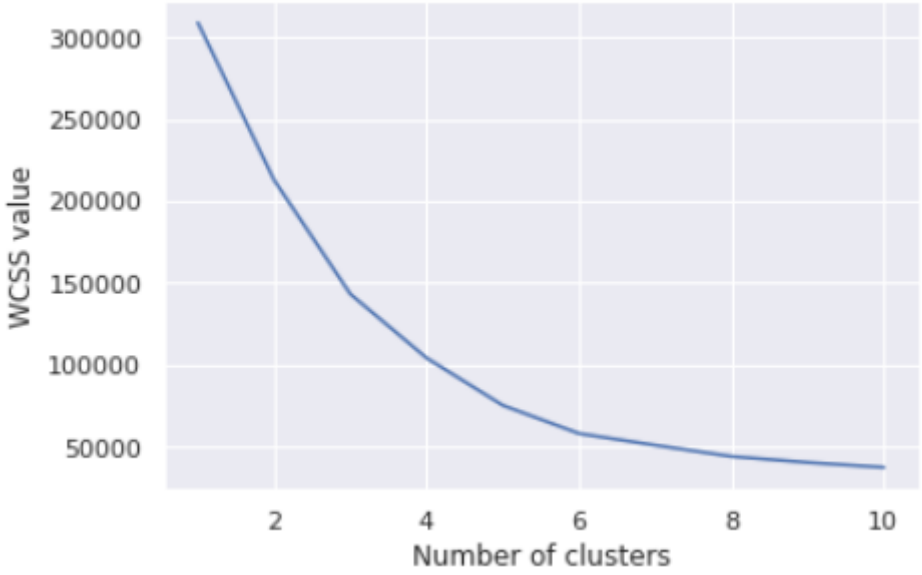
Slika 11. Vizualizacija grupa kupaca s dvije značajke[7]

3.2. K-means model s tri značajke

U modelu s tri značajke se može koristiti ista logika kao i s dvije značajke, ali se može očekivati detaljniji prikaz i konačne grupe u 3D prikazu. Primjenjuju isti principi kao u prethodnom samo sa dodatnom značajkom, a to je dob kupaca. Provođenjem iste metode nailazi se na određivanje broja grupa istom metodom kao u prethodnom primjeru, a to je metoda koljena(slika 13.)

S više korištenih značajki, slika grafa je ponešto grublja pa kad bi se uzela još jedna značajka bilo bi poželjno smanjiti unaprijed određeni hiperparametar K koji je u ovom slučaju kao i u

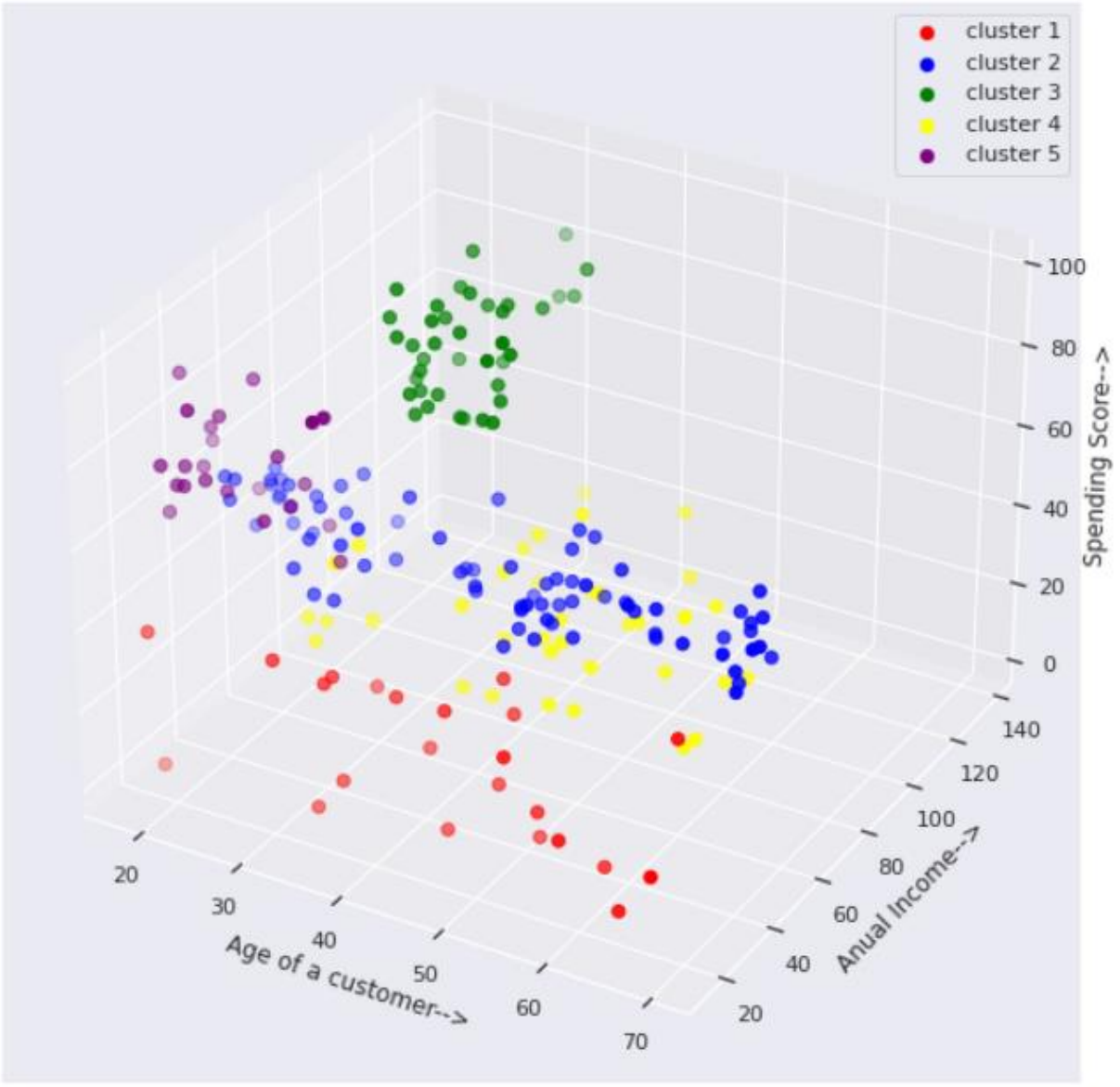
prethodnom deset. Kad bi se smanjio, graf bi bio izraženiji i lakše bi se uočilo „koljeno“ koje definira broj grupa modela.



Slika 12. Metoda koljena kod modela s tri značajke[6]

Može se primijetiti koljeno iznad broja grupa tri i pet. Međutim nakon koljena iznad broj a tri i dalje slijedi drastičnije opadanje greške(WCSS value) dok nakon broja grupe 5 nije istaknuto).

Vizualizacija grupa prikazana je na slici 14. u 3D prostoru jer u ovoj varijaciji modela imamo tri značajke. Sada se jasno vide grupe i karakteristike na temelju kojih se može objasniti, odnosno donijeti zaključak u interesu trgovačkog centra.



Slika 13. Vizualizacija grupa kupaca s tri značajke[6]

Iz slike grupa se može zaključiti da grupe tri i pet imaju visoku naviku potrošnje u odnosu na ostale grupe. Grupa tri predstavlja ljude mlađima od 30 godina s visokim primanjima i visokom navikom potrošnje. Kako bi se održalo ovakvo ponašanje skupine, mogu im se davati bolje ponude kako bi i dalje bili privučeni.

Grupa pet predstavlja ljude također mlađe od 30 godina, ali s niskim primanjima. Za održavanje ove skupine kupaca, mogu biti ponuđeni popustima i akcijama kako bi ih se više privuklo. [6]

3.3. Usporedba modela

Oba računalna modela su kao rješenje odredili optimalan broj grupa pet. U oba slučaja je hiperparametar bio deset. Dobivene su informacije o predviđanju navika određenih grupa kupaca. Kod modela s dvije značajke je na kraju dobivena jedna ciljana skupina koja bi predstavila zadatak za marketinški tim u svrhu boljeg poslovanja i vođenja trgovačkog centra. Kod modela s tri značajke je detaljnija vizualizacija rezultata i iako je ciljana skupina sa većim primanjima i većom navikom potrošnje, također se može obratiti pozornost na skupinu s niskim primanjima, ali isto visokom navikom potrošnje. To omogućuje bolji, detaljniji i spremniji pristup prema strategiji poboljšanja poslovanja. Uz više informacija ima više i posla, ali i lakše je smisliti i razraditi taj plan. Zaključak jest da model daje bolje rezultate, odnosno detaljnije što ima više značajki.

4. EVALUACIJA REZULTATA RADA MODELA

Da bismo što bolje procijenili tu točnost algoritma odnosno pogrešku, moramo znati kako ispravno vrednovati (evaluirati) algoritam. Tri su stvari bitne za vrednovanje algoritama strojnog učenja: koju mjeru vrednovanja koristiti, kako pravedno (realistično) procijeniti pogrešku modela i kako napraviti statističku analizu rezultata kako bismo bili donekle sigurni da naš rezultat nije puka slučajnost.

Problem u ovom slučaju jest što je evaluacija modela moguća na algoritmima metoda nadziranog učenja. U nadziranom učenju uz predviđene oznake postoje i stvarne oznake te je moguće vrednovati model, ako nenadziranog učenja, odnosno kod konkretnog primjera s brojem grupa koji je tražen, ne postoji stvaran ili točan broj već se nadamo da je model dao optimalan.

Ono što je moguće kod nenadziranog učenja evaluirati, točnije provjeriti, to su grupe, odnosno broj grupa. Postoji više tehnika provjera grupa, ali u ovom radu je korištena metoda koljena koja je detaljnije objašnjena u prethodnom poglavlju.

Postoji još pristupa kao npr. točnost na podskupu primjera. Pristup koji se temelji na usporedbi dobivenog grupiranja s grupiranjem koje je smatrano točnim. Tada bi unaprijed trebali imati referentno grupiranje na manjem uzorku primjera što na primjeru u ovom radu nije slučaj.

Za mjeru pogreške može biti korištena bilo koja mjera za točnost grupiranja: npr. Randov indeks, mjera normalizirane uzajamne informacije ili mjera F_1.

To je mjera koja izračunava u kojoj mjeri dobiveno grupiranje odgovara referentnom grupiranju (grupiranju koje smatramo točnim). Randov indeks zapravo nije ništa drugo nego mjera točnosti izračunata na razini parova primjera. To znači da za svaki mogući par iz skupa primjera gledamo jesu li ta dva primjera završila u istoj grupi ili nisu. Ako su završili u istoj grupi, onda par primjera je smatran pozitivnim, inače je smatran negativnim. Ako su oba primjera iz para završila u istoj grupi, i ako su doista trebala završiti u istoj grupi, onda je taj par primjera istinito pozitivan (engl. true positive, TP). Obrnuto, ako su primjeri iz para razdvojeni, tj. svaki je završio u drugoj grupi, a doista nisu trebala završiti u istoj grupi, onda je taj par primjera istinito lažan (engl. true negative, TN).

Randov indeks definiran je ovako:

$$R = \frac{\mathrm{a+b}}{\binom{N}{2}} \tag{8.}$$

gdje je a broj jednako označenih parova u istim grupama (tj. broj istinito pozitivnih parova, TP), a b je broj različito označenih parova u različitim grupama (tj. broj istinito negativnih parova, TN). [8]

5. ZAKLJUČAK

Grane umjetne inteligencije sve više zadiru u život čovjeka olakšavajući neke procese koji su nekad zahtijevali enormnu količinu vremena ili izgledali nemoguće. U ovo slučaju dotakli smo se samo jedne podgrupe, odnosno jedan način korištenja algoritma u svrhu čovjeka. Nenadzirano učenje ima široku primjenu, grupiranje je tek jedna od, a K-srednjih vrijednosti jedan od alata. Sve je zastupljenije u više područja kao što je biologija, knjižnice, osiguranja, društvene mreže, planiranje gradnje gradova, marketing… u ovom radu smo prikazali kako se može poboljšati i unaprijediti upravo marketinška strategija pomoću K-srednjih vrijednosti algoritma

U ovom je radu konkretni naglasak bio na algoritmu K-srednjih vrijednosti za grupiranje kupaca trgovačkog centra s određenim značajkama u svrhu predviđanja navika za poboljšanje rada trgovačkog centra. Algoritam je riješio problem za oba slučaja, s dvije značajke i s tri značajke. Iz modela s tri značajke se dalo zaključiti da je algoritam dao preciznije grupe, odnosno da povećanjem značajki se dobije precizniji model. Za ovaj problem je algoritam poslužio, ali uvijek se može poboljšati, npr. drugačijim brojem grupa ili značajki, te isprobavanjem i eksperimentiranjem modela. To je moguće na način da se proširi bazu podataka, uvede još značajki što bi rezultiralo sofisticiranijim algoritmom.

Cijeli zadatak zahtijeva relativno malo vremena zbog čega ova grana i zauzima dio prostora u našem životu jer pruža mogućnosti obavljanja procesa efektivnije i efikasnije.

LITERATURA

[1] https://www.fer.unizg.hr/predmet/struce1, pristupljeno 23.8.2022.

[2] Clustering in Machine Learning - GeeksforGeeks, pristupljeno 25.8.2022.

[3] https://www.fer.unizg.hr/_download/repository/SU-2020-19-Grupiranje[1].pdf, pristupljeno 23.8.2022.

[4] [https://medium.com/analytics-vidhya/intuition-behind-k-means-clustering-f1ef6006479], pristupljeno 25.8.2022.

[5] https://repozitorij.mathos.hr/en/islandora/object/mathos%3A491/datastream/PDF/view, pristupljeno 25.8.2022.

[6] https://www.kaggle.com/code/azmainmorshed/clustering-for-beginners/notebook#Exploratory-Data-Analysis, pristupljeno 25.8.2022.

[7] https://www.kaggle.com/code/vjchoudhary7/kmeans-clustering-in-customer-segmentation/notebook, pristupljeno 26.8.2022.

[8] SU1-2021-P21-VrednovanjeModela.pdf, pristupljeno 26.8.2022.

PRILOZI

I. Python kod

```
#import the libraries
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt #Data Visualization
import seaborn as sns  #Python library for Vidualization


# Input data files are available in the "../input/" directory.
# For example, running this (by clicking run or pressing Shift+Enter) will list the files in the input directory

import os
print(os.listdir("../input"))

#Import the dataset

dataset = pd.read_csv('../input/Mall_Customers.csv')

#Exploratory Data Analysis
dataset.head(10) #Printing first 10 rows of the dataset
#total rows and colums in the dataset
dataset.shape
dataset.info() # there are no missing values as all the columns has 200 entries properly
#Missing values computation
dataset.isnull().sum()
### Feature sleection for the model
#Considering only 2 features (Annual income and Spending Score) and no Label available
X= dataset.iloc[:, [3,4]].values

#Building the Model
#KMeans Algorithm to decide the optimum cluster number , KMeans++ using Elbow Mmethod
```

```
# ELBOW Method on KMEANS++ Calculation
from sklearn.cluster import KMeans
wcss=[]

for i in range(1,11):
    kmeans = KMeans(n_clusters= i, init='k-means++', random_state=0)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)

    #inertia_ is the formula used to segregate the data points into clusters

#Visualizing the ELBOW method to get the optimal value of K
plt.plot(range(1,11), wcss)
plt.title('The Elbow Method')
plt.xlabel('no of clusters')
plt.ylabel('wcss')
plt.show()
#Model Build
kmeansmodel = KMeans(n_clusters= 5, init='k-means++', random_state=0)
y_kmeans= kmeansmodel.fit_predict(X)

#For unsupervised learning we use "fit_predict()" wherein for supervised learning we use "fit_tranform()"
#y_kmeans is the final model

#Visualizing all the clusters

plt.scatter(X[y_kmeans == 0, 0], X[y_kmeans == 0, 1], s = 100, c = 'red', label = 'Cluster 1')
plt.scatter(X[y_kmeans == 1, 0], X[y_kmeans == 1, 1], s = 100, c = 'blue', label = 'Cluster 2')
plt.scatter(X[y_kmeans == 2, 0], X[y_kmeans == 2, 1], s = 100, c = 'green', label = 'Cluster 3')
plt.scatter(X[y_kmeans == 3, 0], X[y_kmeans == 3, 1], s = 100, c = 'cyan', label = 'Cluster 4')
plt.scatter(X[y_kmeans == 4, 0], X[y_kmeans == 4, 1], s = 100, c = 'magenta', label = 'Cluster 5')
```

```
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], s = 300, c = 'yellow',
label = 'Centroids')
plt.title('Clusters of customers')
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.legend()
plt.show()

###Model Interpretation
#Cluster 1 (Red Color) -> earning high but spending less
#cluster 2 (Blue Colr) -> average in terms of earning and spending
#cluster 3 (Green Color) -> earning high and also spending high [TARGET SET]
#cluster 4 (cyan Color) -> earning less but spending more
        #Cluster 5 (magenta Color) -> Earning less , spending less
```