

Analiza korisničkih podataka na internetskim platformama primjenom algoritama grupiranja

Šimundić, Lovre

Master's thesis / Diplomski rad

2024

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Electrical Engineering and Computing / Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva**

Permanent link / Trajna poveznica: <https://urn.nsk.hr/urn:nbn:hr:168:971661>

Rights / Prava: [In copyright](#) / [Zaštićeno autorskim pravom.](#)

Download date / Datum preuzimanja: **2025-03-21**



Repository / Repozitorij:

[FER Repository - University of Zagreb Faculty of Electrical Engineering and Computing repository](#)



SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 518

**ANALIZA KORISNIČKIH PODATAKA NA INTERNETSKIM
PLATFORMAMA PRIMJENOM ALGORITAMA GRUPIRANJA**

Lovre Šimundić

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 518

**ANALIZA KORISNIČKIH PODATAKA NA INTERNETSKIM
PLATFORMAMA PRIMJENOM ALGORITAMA GRUPIRANJA**

Lovre Šimundić

Zagreb, lipanj 2024.

DIPLOMSKI ZADATAK br. 518

Pristupnik: **Lovre Šimundić (0036526055)**

Studij: Računarstvo

Profil: Znanost o podacima

Mentor: izv. prof. dr. sc. Goran Delač

Zadatak: **Analiza korisničkih podataka na internetskim platformama primjenom algoritama grupiranja**

Opis zadatka:

Motivirati i opisati problem analize korisničkih podataka na internetskim platformama. Odabrati, proučiti i opisati algoritme za grupiranje primjerene obradi velikih skupova podataka. Opisati obilježja algoritama i objasniti njihov princip rada nad pokaznim primjerima. Proučiti postojeće metrike za vrednovanje uspješnosti algoritama grupiranja. Odabrati primjeren skup podataka za analizu korisničkih podataka. Programski ostvariti i provesti vrednovanje primjerenog skupa algoritama nad odabranim skupom podataka. Opisati programsko ostvarenje sustava i rezultate vrednovanja algoritama.

Rok za predaju rada: 28. lipnja 2024.

„Ovim putem zahvaljujem se svom mentoru izv. prof. dr. sc. Goranu Delaču na profesionalnom vodstvu i stručnim savjetima za ostvarenje ovog rada.

Zahvaljujem se ovim putem i svojoj ženi i obitelji, koji su mi bili istinska podrška i pravi suputnici tijekom cijelog studija.

Za kraj, posebnu zahvalu dajem svom Gospodinu, jer bez Njega zaista ništa nije moguće.“

Sadržaj

| | |
|--|----|
| Skraćenice..... | 1 |
| Uvod..... | 2 |
| 1. Analitika internetskih platformi..... | 3 |
| 1.1. Web analitika..... | 3 |
| 1.1.2. Povijest web analitike..... | 4 |
| 1.1.3. Važnost web analitike..... | 5 |
| 1.1.4. Ključni indikatori uspješnosti u web analitici..... | 6 |
| 1.1.5. Proces web analitike..... | 7 |
| 1.1.6. Najpoznatiji alati za web analitiku..... | 8 |
| 1.1.7. Aktualni problemi i budućnost web analitike..... | 8 |
| 1.2. Analitika platformi financijskog sektora..... | 9 |
| 1.2.1. Kratka povijest i potencijalna budućnost financijskih tehnologija.... | 10 |
| 1.2.2. <i>Big Data</i> i analitika u financijskom sektoru..... | 11 |
| 1.2.3. Analitika usluga kreditnih kartica..... | 14 |
| 2. Algoritmi grupiranja..... | 16 |
| 2.1. Podjela algoritama grupiranja..... | 16 |
| 2.2. Particijsko grupiranje..... | 17 |
| 2.2.1. Algoritam k-srednjih vrijednosti..... | 18 |
| 2.2.2. Algoritam k-medoida..... | 19 |
| 2.3. Grupiranje temeljeno na gustoći..... | 21 |
| 2.3.1. DBSCAN algoritam..... | 21 |
| 2.4. Hijerarhijsko grupiranje..... | 23 |
| 2.4.1. Hijerarhijsko aglomerativno grupiranje..... | 23 |
| 2.5. Grupiranje po distribuciji..... | 24 |

| | | |
|--------|--|----|
| 2.5.1. | Model Gaussovih mješavina..... | 25 |
| 2.6. | Metode za provjeru grupiranja..... | 26 |
| 3. | Programsko ostvarenje..... | 30 |
| 3.1. | Analiza korisničkih podataka internet trgovine | 30 |
| 3.2. | Analiza podataka korisnika kreditnih kartica | 37 |
| 3.2.1. | Učitavanje, pregled i vizualizacija podataka | 37 |
| 3.2.2. | Obrada podataka i skaliranje..... | 41 |
| 3.2.3. | Analiza glavnih komponenti (PCA) | 41 |
| 3.2.4. | Grupiranje korisnika kreditnih kartica | 43 |
| 3.2.5. | Rezultati grupiranja..... | 46 |
| 4. | Zaključak | 47 |
| | Literatura..... | 48 |
| | Sažetak | 54 |
| | Summary..... | 55 |

Skraćenice

UX – user experience

GA – *Google Analytics*

KPI – *Key Performance Indicators*

PAM – *Partitioning around medoids*

GMM – *Gaussian Mixture Model*

RFM – *Recency Frequency Monetary*

PCA – *Principal Component Analysis*

DB – Davies-Bouldin

CB – Calinski-Harabasz

Uvod

Nagli razvoj tehnologije i digitalizacija poslova uzrokuju potrebu za prikupljanjem ogromne količine podataka putem internetskih platformi. Podaci koje korisnici ostavljaju tijekom svojih aktivnosti na internetu predstavljaju bogat izvor informacija koji, pravilno analiziran, može donijeti značajne koristi u različitim poslovnim procesima.

Analitika internetskih platformi široka je disciplina koja se bavi prikupljanjem, analizom i interpretacijom korisničkim podataka na internetskim platformama.

Jedan od ključnih izazova u analizi korisničkih podataka je obrada velikih skupova podataka. U tu svrhu, algoritmi grupiranja predstavljaju moćan alat koji omogućava segmentaciju korisnika na temelju njihovih zajedničkih karakteristika. Grupiranje podataka omogućava prepoznavanje obrazaca i skrivenih struktura unutar podataka, što je ključno za donošenje informiranih poslovnih odluka.

Cilj ovog diplomskog rada je istražiti i primijeniti algoritme grupiranja u analizi korisničkih podataka na internetskim platformama. Rad se posebno usredotočuje na web analitiku i analitiku financijskih platformi. U prvom dijelu rada definirana je i objašnjena važnost web analitike, uključujući osnovne pojmove, povijest razvoja te važnost za poslovanje. Prikazani su najčešće korišteni alati i tehnike za prikupljanje i analizu korisničkih podataka. Analogno, obrađena je i analitika platformi financijskog sektora.

Nadalje, rad pruža pregled osnovnih pojmova i principa grupiranja podataka. Opisani su najvažniji algoritmi grupiranja, kao što su K-means, DBSCAN, hijerarhijsko grupiranje i model Gaussovih mješavina, te će se analizirati njihove prednosti i nedostaci. Posebna pažnja bit će posvećena usporedbi algoritama i kriterijima za njihov odabir ovisno o specifičnom problemu.

Nakon teorijskog dijela, odabrana su dva primjerena skupa korisničkih podataka. Na ovim skupovima podataka provedena je analiza i programska implementacija odabranih algoritama grupiranja. Prvi skup podataka su događaji prikupljeni s internet trgovine, a drugi skup podataka su korisnički podaci internet bankarstva. Opisani su korišteni alati i tehnologije potrebne za ostvarenje analize. Implementacija algoritama je detaljno opisana, a rezultati grupiranja prikazani su i analizirani uz pomoć vizualizacija. Posebna pažnja posvećena je postojećim metrikama za vrednovanje uspješnosti algoritama grupiranja.

1. Analitika internetskih platformi

Internetske platforme predstavljaju različite vrste internet stranica i digitalnih sustava koji omogućuju korisnicima razne vrste interakcija, od objave sadržaja do kupovine i financijskih transakcija. One uključuju društvene mreže, e-trgovinske platforme, blogove, financijske servise i druge digitalne sustave. Primjeri internetskih platformi su društvene mreže poput Facebooka i Instagrama, e-trgovinske platforme kao što su Amazon i eBay, blogove poput WordPress-a i Blogger-a te financijski servisi poput PayPal-a i Stripe-a.

Podaci imaju velik utjecaj na uspješnost internetskih platformi jer omogućuju razvoj strategija za njihovo poboljšanje. Analitika internetskih platformi bavi se proučavanjem i analizom podataka prikupljenih tijekom korisničke interakcije s istima. Ova analitika omogućuje razvoj strategija za poboljšanje uspješnosti platformi, poboljšanje korisničkog iskustva, marketinga i prodaje te sigurnosti i zaštite korisnika. Analitički alati omogućuju dubinsko razumijevanje korisničkih obrazaca ponašanja, identifikaciju trendova i optimizaciju sadržaja koji se objavljuje na platformama.

Zbog složenosti teme i odabranih skupova podataka, ovaj rad će se fokusirati na web analitiku, posebno analitiku internetskih trgovina, te analitiku internetskih platformi financijskog sektora. Analitika internetskih trgovina obuhvaća prikupljanje i analizu podataka o kupovnim navikama korisnika, učinkovitosti marketinških kampanja i optimizaciji procesa kupovine. S druge strane, analitika financijskih internetskih platformi uključuje praćenje transakcija, zaštitu podataka i prevenciju prijevara.

1.1. Web analitika

Web analitika je disciplina koja se bavi prikupljanjem, obradom i analizom korisničkih podataka na internetskim stranicama. To je proces praćenja i analize ponašanja posjetitelja na web stranici s ciljem optimizacije performansi web stranice i poboljšanja korisničkog iskustva (engl. *User experience*, skraćeno *UX*) [1]. Korištenjem alata za web analitiku prikupljaju se i analiziraju podaci koji odgovaraju na ključna pitanja poput: koliko ljudi posjećuje web stranicu (engl. *website*), odakle dolaze, koje stranice (engl. *web pages*) posjećuju, koliko vremena provode na najvažnijim stranicama te koliko ih odlazi nakon posjeta samo jedne stranice [2]. Detaljnijom analizom i optimizacijom gore navedenih

faktora indirektno se povećava uspješnost stranice, odnosno poslovanja kojeg stranica reprezentira.

1.1.2. Povijest web analitike

Web analitika započela je zajedno s internetom, u ranim 1990-ima. U to vrijeme, internet je bio u svojim počecima, a web stranice su bile jednostavne, sastavljene uglavnom od teksta i poveznica (engl. *links*). Analiza podataka započela je s jednostavnim brojačima posjeta (engl. *hit counters*) koji su pratili koliko puta je neka stranica posjećena. Ovi rani oblici analize internetskog prometa bili su relativno primitivni i nisu pružali mnogo informacija o stvarnom ponašanju korisnika na web stranicama [3].

Godine 1993. osnovana je prva komercijalna web analitička tvrtka, WebTrends. Ova tvrtka omogućila je analizu log datoteka koje su bilježile sve zahtjeve za HTML elementima, poput teksta, slika ili zvučnih datoteka. Analiza log datoteka postala je prvi korak prema sofisticiranijem razumijevanju web prometa [4].

S razvojem interneta i povećanjem složenosti web stranica, pojavila se potreba za naprednijim alatima za analizu. Dr. Stephen Turner stvorio je 1995. godine *Analog*, prvi besplatni program za analizu log datoteka. *Analog* je omogućio vlasnicima web stranica i marketinškim stručnjacima da bolje razumiju podatke o prometu, pružajući jasne grafikone i dokumentaciju [3].

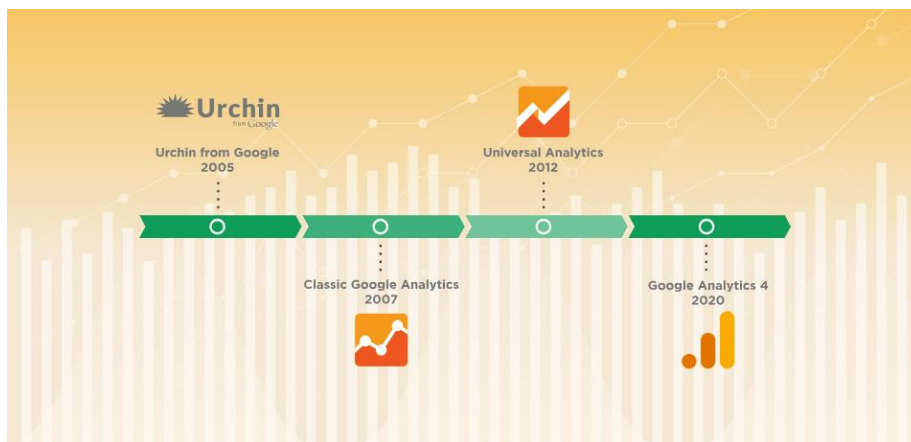
Dodavanjem sve većeg broja multimedijskog sadržaja na internetskim stranicama, brojanje posjetitelja je postalo nedovoljno kvalitetna metrika. Rješenje se pojavilo 1997., a to su bile JavaScript oznake (engl. *tags*) koje su omogućile precizniju analizu prometa i ponašanja korisnika na web stranicama. Oznake su isječak koda koji, jednom umetnut na stranicu, omogućuje slanje podataka na poslužitelja treće strane, na kojem se prikupljaju podaci za analitiku. Ova metoda prikupljanja podataka postala je standard u industriji sve do danas [3][4].

Godine 2004. osnovana je udruga *Web Analytics Association* (WAA), koja je formalizirala praksu web analitike i razvila opsežno znanje za optimizaciju web stranica. U međuvremenu je WAA promijenila ime u *Digital Analytics Association*, a 1. svibnja 2024. udruga je prestala postojati. Godine 2005., *Google* je kupio *Urchin Software Corporation* i lansirao *Google Analytics*, koji je ubrzo postao najkorišteniji alat za web analitiku na tržištu, kao i danas [5].

Google-ov *Universal Analytics* lansiran je 2012., omogućujući praćenje korisnika preko više uređaja i platformi. Ovo je razdoblje obilježeno i porastom mobilne analitike, prilagodbe za praćenje aplikacija i mobilnog web prometa. UA je 2016. integrirao strojno učenje, pružajući korisnicima pametnije uvide i mogućnost praćenja u stvarnom vremenu [3].

Razvojem globalne oznake (engl. *Global tag*; skraćeno GTAG) 2017., Google je omogućio integraciju GA platforme s raznim drugim *Google* platformama, od kojih je najpoznatija *Google Ads* [4].

UA je 2020. zamjenjen naprednijom, GA4 verzijom, koja je u analitičkoj industriji i danas standard.



Slika 1.1.2.1 Kronološki prikaz razvoja GA platforme [4]

Bitno je naglasiti da posljednjih nekoliko godina sve više dolaze do izražaja analitički softveri otvorenog koga (engl. *open source*), kao što su Matomo, Plausible, i mnogi drugi.

1.1.3. Važnost web analitike

Web analitika je ključan alat koji pruža uvid u podatke koji olakšavaju donošenje poslovnih odluka vezanih za neku internetsku stranicu. Neki od primjera kako web analitika pomaže poslovanju [1]:

- Personalizacija sadržaja za već postojeće korisnike ili kupce – poboljšava se UX same stranice, a stječe se i povjerenje korisnika prema tvrtki
- Predviđanje ponovne kupnje u internet trgovini – vjerojatnost da će kupac ponovno kupiti neki proizvod
- Praćenje potrošnje korisničkih skupina – pogodno za primjenu strategija prodaje npr. povećana (engl. *upselling*) i unakrsna (engl. *cross-selling*)

- Geografska analiza – praćenje broja korisnika/kupaca s obzirom na geografsku regiju, prilagođavanje promotivnih strategija na ciljane geografske regije
- Optimizacija sadržaja web stranice s obzirom na najfrekventnije upite u tražilici [1]

1.1.4. Ključni indikatori uspješnosti u web analitici

Ključni indikatori uspješnosti (engl. *Key Performance Indicators*; skraćeno KPI) su kvantitativne mjere koje daju koristan uvid u uspješnost poslovanja u nekom trenutku. Oni ukazuju na napredak (ili nedostatak) u područjima koja su ključna za uspješnost web stranice. KPI-evi su zahvalni jer su precizni i jasni. Osim što daju uvid u trenutno poslovno stanje, također ističu potencijalne probleme i pomažu u pronalaženju rješenja istih [6]. Sada će se prikazati najvažniji KPI-evi.

Prvi, vjerojatno najvažniji KPI je stopa konverzije (engl. *conversion rate*). To je omjer posjetitelja koji su izvršili neku ciljanu akciju na stranici i ukupnog broja posjetitelja. Formula (1) daje kvantitativnu definiciju stope konverzije:

$$\text{stopa konverzije} = \frac{\text{broj konverzija}}{\text{broj posjetitelja}} * 100 \quad (1)$$

Konverzija može biti bilo što što se smatra bitnim za poslovanje i stranicu. Ako se određena tvrtka bavi uslugom asfaltiranja, onda bi konverzija na njihovoj web stranici mogla biti poziv potencijalnog klijenta koji želi asfaltiranje. Bitno je naglasiti da konverzije mogu biti i ulančane. Na internetskoj trgovini, konverzija prvo može biti dodavanje proizvoda u košaricu. Potom se proizvodi pregledavaju u košarici i slijedi opcija plaćanja. Zadnji korak je potvrda plaćanja. Ovo je primjer konverzije u tri koraka. Ulančane konverzije marketinški se nazivaju „lijevak“ (engl. *funnel*).

Sljedeći bitan KPI je stopa odstupanja (engl. *bounce rate*). Stopa odstupanja je broj sjednica (engl. *sessions*) u kojima je posjetitelj posjetio samo jednu stranicu (*page*), i napustio web stranicu. Stopa odstupanja dana je formulom (2):

$$\text{stopa odstupanja} = \frac{\text{broj posjeta samo jedne stranice po sjednici}}{\text{broj sjednica}} * 100 \quad (2)$$

Stopa odstupanja relevanta je metrika zainteresiranosti korisnika za sadržaj na određenoj web stranici. Ako je stopa odstupanja niska, korisnici tj. posjetitelji posjećuju više stranica i duže se zadržavaju gledajući sadržaj istih.

Prosječno vrijeme zadržavanja na pojedinoj stranici također je metrika zainteresiranosti korisnika za sadržaj na istoj. Ova metrika ovisi još i o količini sadržaja na pojedinoj stranici.

Stopa povratka posjetitelja (engl. *return visitor rate*) je KPI koji govori koliki postotak posjetitelja se vrati na web stranicu. Formula za stopu povratka posjetitelja (3) je:

$$\text{stopa povratka posjetitelja} = \frac{\text{broj posjetitelja koji su se vratili}}{\text{broj jedinstvenih posjetitelja}} * 100 \quad (3)$$

Iz priloženog je očito da su KPI-evi neizostavni u web analitici, a ovi pokazani samo su primjeri osnovnih indikatora. Svaka tvrtka prilagođava i razvija svoje KPI-eve u skladu s poslovnim ciljevima i specifičnim situacijama u kojima se nalazi.

1.1.5. Proces web analitike

Proces web analitike obično se sastoji od šest koraka. Prvi korak je postavljanje ciljeva, što se odnosi na definiranje ciljeva i rezultata koje tvrtka želi postići analitikom, poput povećanja prodaje, poboljšanja korisničkog iskustva (UX) i sličnih ciljeva. Ti ciljevi mogu biti kvantitativni i kvalitativni.

Drugi korak je prikupljanje podataka, što je nužan korak za koji su potrebni alati kao što su Google Analytics (GA), Matomo, Adobe Analytics i Usermaven. Korisnički podaci se ne prikupljaju samo na web stranici, već i na drugim platformama koje korisnik posjećuje.

Obrada podataka je treća faza u kojoj se prikupljeni podaci vizualiziraju kako bi se olakšalo donošenje poslovnih odluka.

Četvrti korak uključuje određivanje ključnih pokazatelja uspješnosti (KPI), koji su objašnjeni u prethodnom poglavlju. Tvrtka prilagođava KPI-eve prema vlastitim potrebama.

Razvoj poslovne strategije temeljen na analitici čini peti korak, pri čemu tvrtka, ovisno o analitičkim uvidima, osmišljava strategiju za ostvarenje već definiranih poslovnih ciljeva.

Konačno, šesti korak je eksperimentiranje za validaciju i implementaciju strategije. Čest primjer toga je A/B testiranje, vrsta statističkog testa u kojem se uspoređuju dvije verzije neke strategije, a verzija koja pokaže bolje rezultate primjenjuje se u poslovnoj kampanji tvrtke [1].

Valja istaknuti da je proces web analitike često cikličan i funkcionira na principu pokušaja i pogrešaka, što zahtijeva kontinuiranu prilagodbu i optimizaciju. Ovakav pristup osigurava da strategije ostanu relevantne i uspješne, unatoč dinamičnim promjenama tržišta.

1.1.6. Najpoznatiji alati za web analitiku

U ovom djelu rada navedeni su i ukratko objašnjeni najpoznatiji softveri za web analitiku. Prvi, najzastupljeniji među njima je već spomenuti *Google Analytics*.

GA pruža detaljne informacije o prometu na web stranici i korisničkom ponašanju. Iako je besplatan, omogućava korisnicima uvid u razne informacije: iz kojeg geografskog područja dolaze posjetitelji, kako „navigiraju“ kroz stranicu, koje sadržaje najviše pregledavaju, i koliko vremena provode na pojedinim segmentima stranice. GA nudi informacije o izvorima prometa, učinkovitosti marketinških kampanja i stopama konverzije, što pogoduje optimizaciji marketinških strategija. GA također nudi integraciju s drugim *Google* servisima, poput *Google Ads* i *Google Search Console*-a.

Matomo, ranije poznat kao *Piwik*, predstavlja vodeću *open-source* platformu za web analitiku koja naglasak stavlja na privatnost i vlasništvo podataka korisnika. Ovaj alat omogućava korisnicima da zadrže potpunu kontrolu nad svojim podacima bez dijeljenja s trećim stranama, čime se izbjegavaju problemi povezani s privatnošću i zaštiti podataka. *Matomo* nudi raznovrsne funkcionalnosti, a jednostavniji je za korištenje nego GA. Sve veći broj analitičara se opredjeljuje za *Matomo*.

Hotjar je analitički softver za web analitiku koji se posebno fokusira na UX. Pruža funkcionalnosti kao što su mape vrućine (engl. *heat maps*), snimke sjednica korisnika, ankete i sl. Ovaj alat nastao je primarno za dizajnere.

1.1.7. Aktualni problemi i budućnost web analitike

Web analitika suočava se s nizom aktualnih problema koji utječu na način prikupljanja, obrade i korištenja podataka o korisnicima. Među najznačajnijim izazovima su pitanja privatnosti, prestanak korištenja kolačića trećih strana te rastuća važnost podataka prve strane.

Vjerojatno najveći problem u web analitici danas je pitanje privatnosti podataka. Porast svijesti potrošača o njihovim pravima na privatnost rezultirao je regulativama poput Opće uredbe o zaštiti osobnih podataka (engl. *General Data Protection Regulation*; skraćeno

GDPR) u Europi i Zakona o zaštiti podataka američke savezne države Kalifornija (engl. *California Consumer Privacy Act*; skraćeno *CCPA*). Ove regulative postavljaju stroge zahtjeve na način prikupljanja i obrade podataka. Tvrtke moraju zadobiti suglasnost korisnika za prikupljanje podataka i osigurati da podaci budu korišteni u skladu s važećim zakonima. Ove regulative povećavaju složenost upravljanja podacima i zahtijevaju implementaciju naprednih sustava za upravljanje suglasnostima korisnika [10].

Problem koji se veže na prethodni je kraj kolačića trećih strana. Preglednici kao što su *Safari* i *Firefox* već blokiraju kolačiće trećih strana, a *Google Chrome* planira to učiniti do kraja 2024. godine. Ovo ograničenje utječe na sposobnost tvrtki da prate korisnike preko različitih web stranica i pružaju personalizirane oglase [11].

Smanjenjem upotrebe kolačića trećih strana i postroženjem regulacijskih zakona, budućnost web analitike sve više se oslanja na metode prikupljanja korisničkih podataka bez ugrožavanja privatnosti korisnika. Jedan od ključnih aspekata ove tranzicije je rastuća važnost podataka prve strane. Ovi podaci su izuzetno vrijedni ih tvrtke prikupljaju izravno od korisnika putem interakcija na vlastitim web stranicama, što im daje na preciznosti. U izvještaju iz 2022. tvrtke *Acquia*, 88% marketingaša govore da su im podaci prve strane važniji nego ikad [10]. Prikupljanje podataka prve strane također pomaže u izgradnji povjerenja korisnika jer se podaci prikupljaju uz njihov pristanak. Tvrtke koriste različite alate za prikupljanje ovih podataka, kao što su CRM sustavi, analitičke platforme i alati za upravljanje suglasnostima [10].

Korištenje anonimnih podataka u analitici također postaje sve popularnije. Metode anonimnog praćenja, koje ne koriste kolačiće (engl. *cookieless*) ili prateće identifikatore, omogućuju prikupljanje podataka o korisničkim sjedincama bez identifikacije pojedinačnih korisnika. Ove metode pomažu u usklađivanju s propisima o privatnosti, kao što su *GDPR* i *CCPA*, i omogućuju prikupljanje korisnih podataka o posjetiteljima web stranica na način koji poštuje njihovu privatnost. Iako ove metode mogu biti manje precizne u razlikovanju posjetitelja, one i dalje pružaju vrijedne analitičke uvide [12].

1.2. Analitika platformi financijskog sektora

Financijski sektor prati trendove tehnološkog napretka. Financijska tehnologija (engl. *fintech*) oblikuje načine na koje korisnici upravljaju novcem. Prema istraživanjima *McKinsey & Company*, fintech tvrtke koriste tehnologiju i usluge oblaka (engl. *cloud*) kako

bi omogućile različite financijske usluge bez potrebe za fizičkim lokacijama. Ove platforme olakšavaju korisnicima prebacivanje novca između računa, osoba, država i organizacija. Razvoj fintech-a od 2010. godine donio je revoluciju u platnom prometu, investiranju i zaštiti novca, pružajući jednostavnije i pristupačnije usluge korisnicima širom svijeta [13].

Analitika igra ključnu ulogu u fintech sektoru, omogućujući tvrtkama da koriste podatke za poboljšanje svojih usluga. Kroz analizu velikih skupova podataka, fintech kompanije mogu otkriti obrasce i trendove koji im pomažu donositi informirane odluke, poboljšati korisničko iskustvo i sl. Na primjer, upotrebom prediktivnog modeliranja i analitike u stvarnom vremenu, fintech tvrtke mogu unaprijediti upravljanje rizikom i detekciju prevara, što je ključno za njihovu sigurnost i učinkovitost [14].

Ovo potpoglavlje će se posebno usredotočiti na analitiku kreditnih kartica i online transakcija. Obradit će ključne faze razvoja financijskih tehnologija, ulogu velikih podataka u financijskom sektoru te primjenu analitike u otkrivanju prijevara, procjeni kreditne sposobnosti, personalizaciji ponuda i optimizaciji poslovnih procesa.

1.2.1. Kratka povijest i potencijalna budućnost financijskih tehnologija

Financijska tehnologija razvija se gotovo 150 godina. Prema izvoru [15], povijest fintecha može se podijeliti u nekoliko ključnih faza.

Prvom fazom smatra se razdoblje od 1886. – 1967. U ovom periodu izgrađivale su se infrastrukture za podržavanje globalizirane financijske usluge. Prvi transatlantski kabel postavljen 1866. i *Fedwire* 1918. u SAD-u omogućili su prvi sustav elektroničkog prijenosa sredstava. Ovi sustavi koristili su tehnologije poput telegrafa i Morseove abecede.

Druga faza trajala je od 1967. – 2008. Početak ove faze obilježila je instalacija prvog bankomata od strane britanske banke *Barclays*. Ovo je zapravo značilo početak prelaska s analognog na digitalno financiranje. Osnivanje NASDAQ-a 1970-ih, prve digitalne burze dionica, te SWIFT-a (skraćeno od *Society For Worldwide Interbank Financial Telecommunications*), komunikacijskog protokola između financijskih institucija, omogućilo je veliki obujam prekograničnih plaćanja. Ova era se nastavila kroz 1980-e s rastom bankovnih glavnih (engl. *mainframe*) računala. Rast online bankarstva kroz 1980-e promijenio je način poslovanja ljudi, zajedno s internet revolucijom koja je radikalno promijenila pogled na financijske institucije. Digitalno bankarstvo započelo je 1990-ih,

povezivanjem korisnika koji su počeli upravljati svojim novcem na različite načine. *PayPal* je lansiran 1998. godine, što je nagovijestilo nove sustave plaćanja dok je svijet sustavno postajao sve više online. Ekonomija se naizgled činila stabilnom, što je čak potaknulo tadašnjeg britanskog kancelara Gordona Browna da proglasi "kraj buma i pada". Međutim, upravo je pad – globalna financijska kriza 2008. godine, stavio točku na kraj ove ere fintech-a i potaknuo inovacije koje će se vidjeti u narednoj eri [15].

Trenutna, treća faza fintech-a, traje od 2008. Nakon globalne financijske krize 2008. godine, smanjeno povjerenje u banke i regulatorne promjene otvorile su tržište za nove pružatelje usluga. *Bitcoin* je stvoren 2009. godine, a pametni telefoni postali su primarno sredstvo pristupa internetu i drugim financijskim uslugama. Era *start-up*-ova donijela je val novih proizvoda i usluga, a čak su i postojeće banke počele djelovati i prezentirati se kao *start-up*-ovi [15].

Dok se svijet oporavlja od pandemijske krize, predviđanje budućnosti fintech industrije predstavlja izazovan problem. Iako su investicije rizičnog kapitala u fintech znatno pale, temeljne tehnologije poput blockchaina i otvorenog bankarstva vjerojatno će nastaviti poticati inovacije. Kina i Indija, koje su među zemljama s najvećim korištenjem financijskih tehnologija, brzo su usvojile nova rješenja bez naslijeđene bankovne infrastrukture Zapada, što im je omogućilo brži napredak u digitalnom bankarstvu [15].

Razvoj fintech-a od kasnog 19. stoljeća do danas pokazuje kako tehnologija kontinuirano poboljšava način na koji ljudi upravljaju svojim financijama, oblikujući budućnost financijskih usluga. Osim u bankarstvu, strojno učenje ubrzava obradu osiguravateljskih zahtjeva, kao što se već događa u Kini. Uvođenje integriranih sustava plaćanja u različite sektore omogućit će tvrtkama učinkovitije upravljanje poslovanjem. Personalizacija i automatizacija bit će ključne za budući razvoj sustava za upravljanje financijama, poboljšanje korisničkog iskustva i omogućavanje tvrtkama dubljeg uvida u poslovne performanse [15].

1.2.2. *Big Data* i analitika u financijskom sektoru

Jedan od najznačajnijih aspekata rasta fintech industrije u posljednjem desetljeću je digitalizacija sektora bankarstva, financija i osiguranja (engl. *Banking, Financial Services and Insurance*). Posebno se ističe Indija, koja je postigla izvanredan napredak u digitalnim financijskim uslugama. Na primjer, prema izvještaju *Ernst & Young*-a, digitalno kreditiranje

je činilo samo 1% svih odobrenih kredita u 2017. godini, dok je taj postotak porastao na 12% u 2022. godini. Također, Indija je zabilježila najveći broj digitalnih platnih transakcija na svijetu, s impresivnih 8,5 milijuna. Nadalje, 46% globalnih plaćanja u stvarnom vremenu potječe upravo iz Indije, što naglašava njen utjecaj i važnost u oblikovanju budućnosti digitalnih financijskih ekosustava [16].

Korištenje big data u fintechu omogućava financijskim institucijama bolje razumijevanje ponašanja i potreba korisnika. Svaka osoba s digitalnim otiskom ostavlja trag podataka koje tvrtke mogu koristiti za donošenje preciznih poslovnih odluka. Ovi podaci, bilo strukturirani ili nestrukturirani, prikupljaju se i analiziraju kako bi se identificirali trendovi, obrasci i korelacije. Financijske institucije koriste big data za segmentaciju korisnika, unapređenje upravljanja odnosima s korisnicima, optimizaciju opskrbnog lanca i procjenu rizika. Na primjer, u digitalnom kreditiranju, analitika velikih podataka koristi se za procjenu kreditne sposobnosti korisnika, omogućujući brže i točnije donošenje odluka.

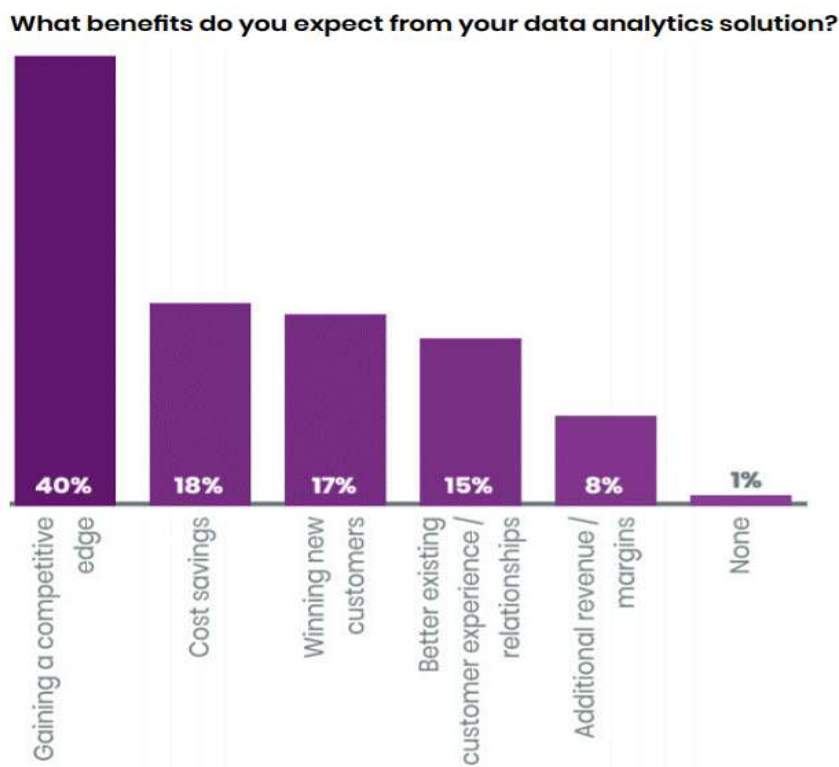
Analitika u fintech industriji postaje neizostavna, što potvrđuje izvještaj ankete tvrtke Fintech Future [18]. U anketi je čak 86% ispitanika priznalo da analitika igra važnu ulogu u oblikovanju poslovanja i da je neizostavna za bilo kakav oblik tehnološkog rješenja. Analiza velikih podataka primjenjuje se u financijskom sektoru u različite svrhe.

Velika količina podataka omogućava segmentaciju korisnika u vrlo specifične svrhe, što fintech tvrtkama omogućava pružanje ciljanih proizvoda. Ovo ciljanje korisnika temelji se na ponašanju, potrebama i demografskim karakteristikama korisnika, čime se smanjuju troškovi akvizicije i zadržavanja korisnika.

Big data pomaže pružateljima financijskih usluga da ponude personalizirane proizvode i poboljšaju korisničko iskustvo. Na primjer, zajmodavci mogu upotrijebiti osnovne podatke za čišćenje kako bi došli do pojedinosti poput imena i brojeva mobitela, koji se kasnije mogu koristiti za automatsko popunjavanje digitalnih aplikacija.

Automatizacija i optimizacija kroz analitiku podataka pojednostavljaju operacije, smanjujući ručni rad i minimizirajući greške. Analitika podataka također pomaže u prepoznavanju tržišnih trendova i preferencija korisnika, vodeći razvoj novih proizvoda koji bolje zadovoljavaju tržište. Podaci o korisnicima koriste se za mapiranje njihovog financijskog ponašanja, što omogućava sveobuhvatan pregled financijskog zdravlja

korisnika. Ova se analiza koristi za prodaju dodatnih proizvoda, unakrsnu prodaju i procjenu rizika, čime se poboljšava.



Slika 1.2.2.1 Anketa koja govori što tvrtke očekuju primjenom analize podataka [18]

Primjeri fintech tvrtki koje učinkovito koriste analitiku podataka za postizanje uspjeha su *PayPal* i *Square*. *PayPal* koristi analitiku podataka za različite aspekte poslovanja kako bi poboljšao sigurnost, korisničko iskustvo i poslovne odluke. Korištenjem big data tehnologija poput Hadoop-a i strojnog učenja, *PayPal* može analizirati ogromne količine podataka u stvarnom vremenu. Ova analitika omogućava *PayPal*-u prepoznavanje obrazaca ponašanja korisnika, što pomaže u personalizaciji usluga i detekciji prijevara.

Na primjer, analitički alati omogućuju brzu identifikaciju sumnjivih transakcija, čime se smanjuje rizik od prijevara i povećava sigurnost korisnika. *PayPal* koristi podatke o transakcijama za kreiranje prediktivnih modela koji analiziraju ponašanje korisnika i predviđaju buduće aktivnosti. Ova analitika omogućava personalizirano oglašavanje i prilagođene ponude temeljene na povijesti kupovine korisnika.

Nadalje, *PayPal* koristi podatke za preporuku proizvoda i usluga koje bi mogle zanimati korisnike na temelju njihovih prethodnih kupovina i aktivnosti na mreži. Hadoop se koristi za pohranu i obradu velikih količina podataka, omogućujući *PayPalu* da kombinira podatke

iz različitih izvora i brzo izvlači korisne uvide. Ovo uključuje analizu podataka u stvarnom vremenu kako bi se otkrile prijevare i optimizirali poslovni procesi. *PayPal* također koristi algoritme strojnog učenja za procjenu kreditne sposobnosti korisnika, omogućavajući brže i točnije donošenje odluka o kreditima [18].

Square platforma za analitiku podataka pruža trgovcima vrijedne uvide u njihove poslovne operacije. Analizom podataka o transakcijama, *Square* pomaže tvrtkama donositi informirane odluke o upravljanju inventarom, cijenama i marketingu. Osim toga, *Square* koristi podatke za prepoznavanje obrazaca ponašanja korisnika, poboljšavajući time svoje usluge [17].

Spajanje financijskih tehnologija i velikih podataka drastično je promijenilo financijske platforme, poboljšavajući interne procese unutar financijskih institucija i korisničko iskustvo. Međutim, s rastućom količinom prikupljenih podataka, ključno je uspostaviti stroge mjere zaštite privatnosti kako bi se održalo povjerenje korisnika. Iako sama etika podataka nije krajnje rješenje, usvajanje etičkog pristupa obradi osobnih podataka ključno je za povjerenje i poticanje angažmana kupaca, povećanje prihvaćanja modernih kreditnih proizvoda i u konačnici poticanje financijske uključenosti [16].

1.2.3. Analitika usluga kreditnih kartica

Analitika usluga kreditnih kartica postala je ključni alat za financijske institucije u razumijevanju ponašanja korisnika, otkrivanju prijevara i donošenju informiranih odluka o kreditima.

Jedan od najvažnijih aspekata analitike kreditnih kartica je otkrivanje i prevencija prijevara. Financijske institucije koriste algoritme strojnog učenja za analizu transakcija u stvarnom vremenu kako bi identificirale sumnjive aktivnosti. Na primjer, algoritmi mogu prepoznati neuobičajene obrasce potrošnje ili transakcije iz neuobičajenih lokacija, što može ukazivati na potencijalnu prijevaru. Ovi sustavi mogu raditi na nadziranom i nenadziranom učenju, gdje se analiziraju povijesni podaci za identificiranje anomalija te se kontinuirano uče iz novih podataka kako bi se poboljšala točnost predviđanja (Miquido) (DataToBiz).

Analitika kreditnih kartica također omogućava detaljnu procjenu kreditne sposobnosti korisnika. Korištenjem prediktivne analitike, banke mogu analizirati povijesne podatke o kreditnim karticama, financijskom ponašanju i drugim relevantnim faktorima kako bi

procijenile vjerojatnost otplate kredita. Ovaj pristup omogućava bankama da donesu informirane odluke o odobravanju kredita i smanji rizik od neispunjenja obveza [21].

Korištenjem analitike podataka, financijske institucije mogu personalizirati ponude kreditnih kartica prema specifičnim potrebama i preferencijama korisnika. Analizom podataka o potrošnji korisnika, banke mogu identificirati kategorije potrošnje i prilagoditi marketinške kampanje kako bi potaknule potrošnju u tim kategorijama. Na primjer, korisnicima koji često putuju mogu se ponuditi posebne ponude za putovanja, dok se korisnicima koji često kupuju u određenim trgovinama mogu ponuditi dodatni popusti ili pogodnosti [20].

2. Algoritmi grupiranja

U ovom poglavlju definirani su i objašnjeni algoritmi grupiranja. Pokazani su primjeri primjene ovih algoritama. Definirane su i mjere za procjenu kvalitete grupiranja.

Grupiranje (engl. *clustering*) je skup algoritama nenadziranog strojnog učenja koji dijeli primjere iz skupa podataka na grupe (engl. *clusters*) po nekoj mjeri sličnosti. Slični primjeri idu u istu grupu, što je cilj grupiranja – pronalazak „prirodnih“ grupa među primjerima [25].

2.1. Podjela algoritama grupiranja

Postoji nekoliko vrsta podjele algoritama grupiranja. Važno je znati da nijedna podjela nije konačna i da su sve podjele na neki način isprepletene. Prva je podjela na temelju jasnoće granica između grupa. Postoje čvrsto i meko grupiranje. U slučaju čvrstog grupiranja (eng. *hard clustering*), svaki primjer pripada samo jednoj grupi. Nasuprot tome, u mekom grupiranju (engl. *soft clustering*), pojedini primjeri mogu biti dio više grupa, pri čemu stupanj ili vjerojatnost pripadnosti može varirati [25].

Primjer čvrstog grupiranja na jednostavnom skupu podataka:

Tablica 2.1.1. Primjer čvrstog grupiranja

| Primjeri | Grupa |
|----------|-------|
| A | G1 |
| B | G2 |
| C | G2 |

Svaki primjer pripada točno jednoj grupi, kao što se vidi u tablici 2.1.1.

Primjer mekog grupiranja za isti skup podataka:

Tablica 2.1.2. Primjer mekog grupiranja

| Primjeri | P(G1) | P(G2) |
|----------|-------|-------|
| A | 0.75 | 0.25 |
| B | 0.12 | 0.88 |
| C | 0.34 | 0.66 |

Svaki primjer pripada grupi G1 ili G2 s određenom vjerojatnošću.

Druga podjela grupiranja je na particijsko (engl. *partitional*) i hijerarhijsko (engl. *hierarchical*) grupiranje. Particijsko grupiranje primjere particionira u grupe sličnih primjera. Hijerarhijsko grupiranje primjere razdjeljuje u **ugniždene** (engl. *nested*) grupe koje čine hijerarhiju grupa.

Treća, uobičajena i najvažnija podjela je na ove četiri vrste algoritama:

1. Grupiranje temeljeno na centroidima (particijsko grupiranje)
2. Grupiranje temeljeno na gustoći (algoritmi temeljeni na modelu)
3. Grupiranje temeljeno na povezanosti (hijerarhijsko grupiranje)
4. Grupiranje temeljeno na distribuciji

Sljedeća potpoglavlja detaljno će obraditi svaku vrstu grupiranja, navesti najpoznatiji algoritme svake vrste i pokazati primjer primjene tih istih algoritama.

2.2. Particijsko grupiranje

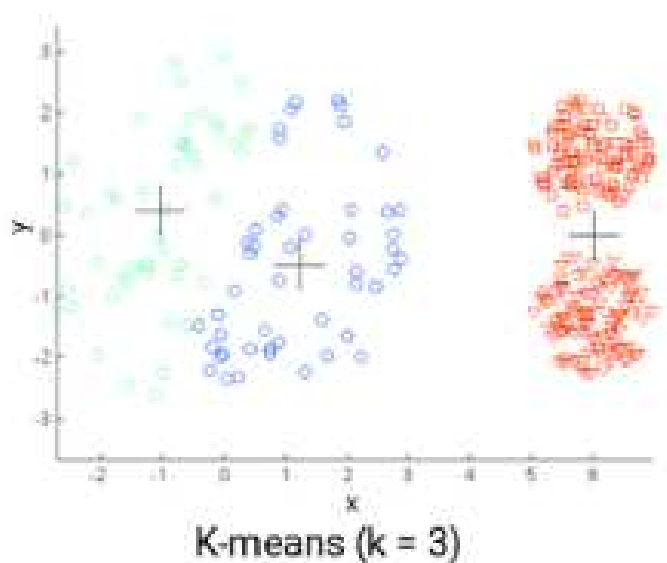
Particijsko grupiranje je, kao što je u prethodnom potpoglavlju objašnjeno, vrsta grupiranja u kojem se primjeri iz skupa podataka raspodjeljuju u grupe prema udaljenosti od centra grupe - centroida. Ovo je najjednostavnija vrsta grupiranja, a primjeri se grupiranju na temelju njihove sličnosti tj. bliskosti. Sličnost se mjeri mjerama kao što su euklidska udaljenost, Manhattan udaljenost i Minkowski udaljenost. Skupovi podataka dijele se na unaprijed određeni broj grupa, a svaka grupa predstavljena je vektorom vrijednosti. Sljedeća dva potpoglavlja opisuju dva najpoznatija ovakva algoritma.

2.2.1. Algoritam k-srednjih vrijednosti

Algoritam k-srednjih vrijednosti (engl. *K-means*) je vrsta partijskog grupiranja u kojem se svaki primjerak iz skupa podataka smješta u točno jednu od k disjunktne grupa primjeraka.

Pseudokod algoritma glasi:

1. Odaberi nasumično k točaka iz skupa podataka. Ove točke su početni centroidi grupa.
2. Za svaku točku (primjerak) iz skupa podataka izračunaj udaljenost između te točke i svakog od k centroida. Dodijeli točku onoj grupi čiji je centroid najbliži. Ovaj korak efektivno formira k grupa.
3. Nakon što su sve točke dodijeljene grupama, ponovno izračunaj centroide grupa tako da uzmeš srednju vrijednost svih točaka dodijeljenih svakoj grupi.
4. Ponavljaj korake 2 i 3 dok ne dođe do konvergencije. Konvergencija se događa kada se centroidi više ne mijenjaju značajno ili kada je dosegnut određeni broj iteracija.
5. Nakon što se postigne konvergencija, algoritam ispisuje konačne centroide grupa i dodjelu svake točke grupi [28].



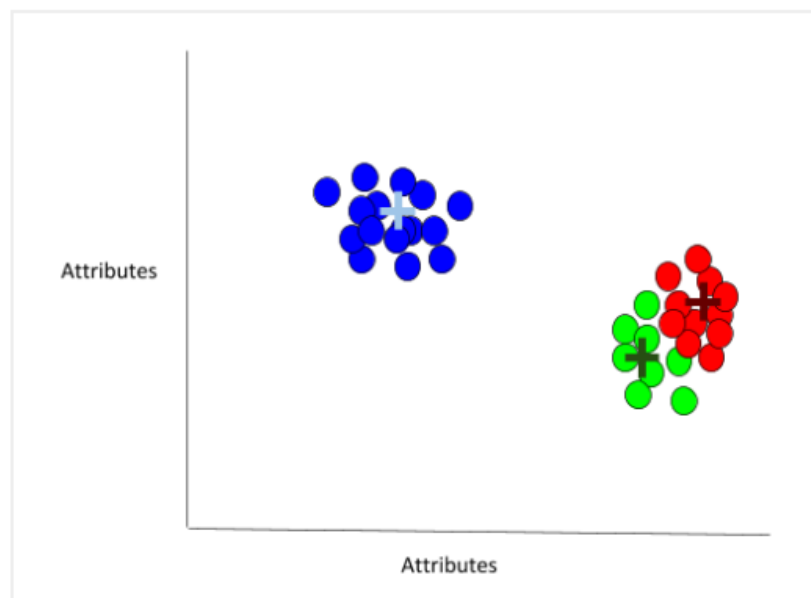
Slika 2.2.1.1 Primjer grupiranja podataka u $k=3$ grupe [28]

Cilj *K-means* algoritma je smanjiti zbroj kvadrata pogreške (engl. *sum of squared error*; skraćeno SSE), a to je kvadrirana suma udaljenosti svakog primjera iz grupe od centroida te grupe [27]. Zbroj suma kvadrata dan je formulom (4) [27]:

$$SSE = \sum_{k=1}^K \sum_{i=1}^{n_k} \|x_{k_i} - \mu_k\|^2 \quad (4)$$

Bitno je naglasiti da algoritam konvergira kad suma kvadrata teži minimumu.

K-means je vjerojatno najpopularniji algoritam grupiranja, jednostavan je za implementaciju i široko primjenjiv. Ovaj algoritam je intuitivan jer optimizira sličnost unutar svake grupe. Mane ovog algoritma su što loše podnosi stršeće vrijednosti (engl. *outliers*) tj. iznimke u podacima [27]. Nadalje, povećanjem dimenzionalnosti svaka točka postaje sve sličnija drugoj pa se gubi intuitivno grupiranje. Isto tako, nije primjenjiv na kategorijske varijable te često završava u lokalnom, a ne u globalom minimumu.



Slika 2.2.1.2 Problem lokalnog minimuma k-means algoritma [32]

Iz gornje slike se jasno vidi problem lokalnog minimuma k-means algoritma. Intuitivno je jasno da su podaci „djeljivi“ u dvije grupe, ali pogrešnim inicijalnim odabirom parametra k i postavljanja centroida dolazi do ovakve pojave.

2.2.2. Algoritam k-medoida

Algoritam k-medoida (engl. *K-medoids*) je poopćeni algoritam k-srednjih vrijednosti. Još se naziva i particioniranje oko medoida (engl. *Partitioning around medoids*; skraćeno *PAM*), gdje je medoida definirana kao primjer koji je najmanje različit od svih ostalih primjera u

grupi [29]. PAM algoritam nije ograničen samo na vektorski prostor i može se koristiti u situacijama gdje je dana samo mjera sličnosti ili različitosti između primjera. Za razliku od toga, K-means obično koristi euklidsku udaljenost kao specifičan slučaj mjere različitosti, što PAM čini općenitijim algoritmom koji omogućuje korištenje različitih mjera sličnosti, odnosno različitosti. Primjer korištenja ovog algoritma je grupiranje riječi na temelju sličnosti nizova znakova [25].

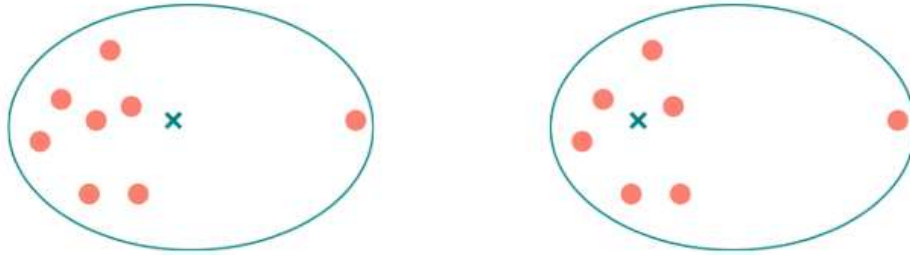
Formula funkcije gubitka PAM algoritma (5) definirana je kao [29]:

$$c = \sum_{\{C_i\}} \sum_{\{P_i \in C_i\}} |P_i - C_i| \quad (5)$$

$E = |P_i - C_i|$ u formuli je zapravo mjera različitosti između primjerka iz grupe i i medoida grupe. Funkcija gubitka jednaka je sumi svih različitosti u svim grupama.

Pseudokod algoritma glasi:

1. Odaberi nasumično k početnih medioda iz skupa podataka.
2. Za svaku točku iz skupa podataka izračunaj različitost između te točke i svakog od k -medoida. Dodijeli točku onom medoidu s kojim ima najmanji različitost. Ovaj korak formira K grupa.
3. Za svaki medoid m , i za svaku točku t koja nije medoid:
 - a. Privremeno zamijeni medoid m s točkom t .
 - b. Ponovno dodijeli sve točke najbližim medoidima koristeći novi skup medoida.
 - c. Izračunaj ukupni trošak (sumu različitosti svih točaka s njihovim najbližim medoidom).
 - d. Ako je novi ukupni trošak manji, prihvati zamjenu; inače, poništi zamjenu.
4. Ponavljaj korak 3 dok se ne postigne stabilnost u pozicijama medoida i dok se ukupni trošak ne prestane smanjivati
5. Algoritam završava kada daljnje zamjene ne smanjuju ukupni trošak ili kada se medoidi više ne mijenjaju između iteracija.
6. Nakon što se postigne konvergencija, algoritam ispisuje konačne medoide i dodjelu svake točke svojoj grupi [28].



Slika 2.2.2.1 Grafički prikaz razlike *k-means* (s lijeva) i *PAM* algoritma (s desna) [31]

Na slici 2.2.2.1 valja primjetiti da je najdesnija točka zapravo iznimka, koja pomiče centroidu kod *K-means*-a, ali ne i kod *PAM*-a. Razlog tome je što je centroida tj. medoida stvarni primjer iz skupa podataka, a ne srednja vrijednost.

PAM je otporniji (robustniji) prema iznimkama (outlierima) u usporedbi s *K-means* jer su medoidi, kao stvarne točke skupa podataka, manje podložni ekstremnim vrijednostima. Međutim, *PAM* je, zbog potrebe za provjerom svih mogućih zamjena medoida, često računalno zahtjevan, što ga čini sporijim od *K-means*, posebno za velike skupove podataka.

2.3. Grupiranje temeljeno na gustoći

Grupiranje temeljeno na gustoći je vrsta grupiranja koja grupe definira kao regije visoke gustoće podataka, a te grupe razdvojene su s regijama male gustoće podataka [31]. Ovi algoritmi lako prepoznaju grupe bilo kakvih vizualnih oblika, ali se oslanjaju na činjenicu da su podaci unutar iste grupe gusto raspoređeni.

2.3.1. DBSCAN algoritam

DBSCAN (skraćeno od engl. *Density-Based Spatial Clustering of Applications with Noise*) je najpoznatiji i najkorišteniji algoritam grupiranja temeljen na gustoći. Ovaj algoritam grupira točke koje su gusto raspoređene u jednoj regiji, dok točke u područjima niske gustoće označava kao stršeće vrijednosti. DBSCAN ima skup relacija po kojima primjere svrstaje u grupe ili ih definira kao šumove. Primjer q je direktno dostupan po gustoći (engl. *direct density-reachable*) primjeru p ako je na najviše ϵ udaljen od njega, i ako postoji dovoljno primjera oko p da se može tvoriti grupa oko p i q . Ova relacija nije simetrična, odnosno ne mora vrijediti da je p direktno dostupan q [31].

Primjer q je dostupan po gustoći (engl. *density-reachable*) primjeru p ako postoji n primjera p_1, \dots, p_n , gdje je svaki p_{i+1} direktno dostupan po gustoći p_i . Bitno je primjetiti da je p_1 zapravo q , a p_n je p [31].

Treća relacija je povezanost po gustoći (engl. *density connectivity*), a analogna je drugoj relaciji. Primjeri p i q su povezani ako postoji n primjera p_1, \dots, p_n , gdje je svaki p_{i+1} dostupan po gustoći p_i , gdje su p i p_1 dostupni, kao i p_n i q [31].

Algoritam ima dva parametra – ε i m . ε je udaljenost prema kojoj se računa direktna dostupnost po gustoći, a m je minimalan broj podataka koji tvore grupu [31]. Svaki primjer unutar grupe moraju biti međusobno povezani, a ako je primjer povezan s drugim primjerom, onda je i taj drugi primjer također uključen u grupu.

Pseudokod algoritma izgleda ovako:

1. Odaberi nasumično početni primjer iz skupa podataka.
2. Formiraj ε -susjedstvo za početni podatak i provjeri njegovu veličinu (broj susjeda).
3. Odluči o statusu početnog podatka:
 - a. Ako susjedstvo ima dovoljno susjeda ($\geq m$): Formiraj grupu koja uključuje sve podatke iz tog susjedstva.
 - b. Ako susjedstvo nema dovoljno susjeda: Označi početni primjer kao stršeću vrijednost
4. Proširi grupu:
 - a. Za svaki novi primjer u grupi, formiraj njegovo ε -susjedstvo.
 - b. Ako podatak ima dovoljno susjeda, uključi sve njegove susjede u grupu koristeći povezanost po gustoći.
 - c. Ponavljaj ovaj proces sve dok se ne uključe svi dostupni podaci koji ispunjavaju kriterije za uključivanje u grupu.
5. Ponovi proces odabira početnog primjera i formiranja grupa sve dok postoje neposjećeni primjeri ili dok se grupa ne može više proširiti.
6. Algoritam završava kada svi podaci budu ili uključeni u neku grupu ili označeni kao stršeće vrijednosti.
7. Algoritam ispisuje grupe i pripadnost svakog podatka odgovarajućoj grupi [31].



Slika 2.3.1.1 Grafički prikaz razlike K-means (s lijeva) i DBSCAN (s desna) grupiranja [31]

Iz slike 2.3.1.1 jasno se vide prednosti DBSCAN algoritma naspram K-means-a.

2.4. Hijerarhijsko grupiranje

Hijerarhijsko grupiranje je vrsta grupiranja koja stvara hijerarhiju grupa. Ova hijerarhija se grafički prikazuje kao stablo, odnosno dendrogram. Lišće dendrograma su svi primjeri iz skupa podataka, dok je u korijenu samo jedna grupa. Hijerarhijsko grupiranje se dijeli na aglomerativno i divizno. Aglomerativno hijerarhijsko grupiranje je „bottom-up“, dok je divizno hijerarhijsko grupiranje „top-down“ metoda [31]. Više o ovome u sljedećem potpoglavlju.

Divizno grupiranje je u osnovi računalno sporije i zahtjevnije, stoga će se sljedeće potpoglavlje baviti samo aglomerativnim grupiranjem.

2.4.1. Hijerarhijsko aglomerativno grupiranje

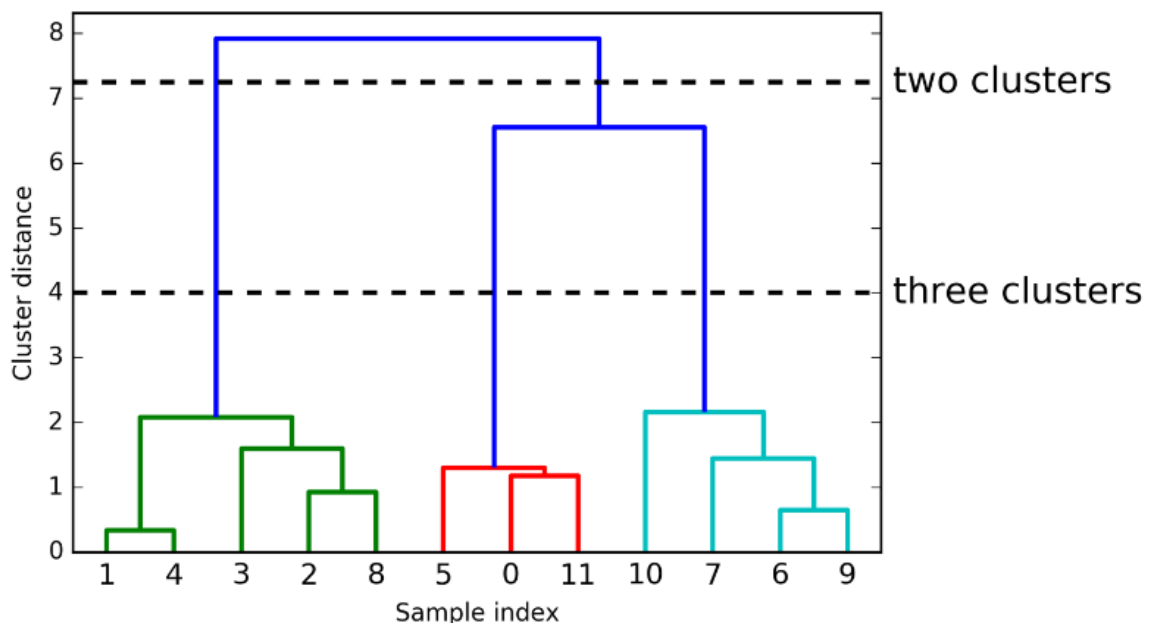
Aglomerativno grupiranje je, već spomenuto, „bottom-up“ metoda hijerarhijskog grupiranja. To znači da svaki primjer iz skupa podataka na početku pripada svojoj grupu, onda se iterativnom procedurom primjeri stapaju u grupe, dok svi primjeri ne postanu dio jedne iste grupe [25].

Pseudokod aglomerativnog grupiranja:

1. Inicijaliziraj matricu udaljenosti za pohranu udaljenosti između svakog para točaka podataka.
2. Inicijaliziraj popis grupa gdje je svaka točka podataka zasebna grupa.
3. Inicijaliziraj matricu udaljenosti grupa gdje je svaki red i stupac grupa.

4. Ponavlja sljedeće korake dok ne ostane samo jedna grupa:
 - a. Pronađi dvije najbliže grupe u popisu grupa i spoji ih u jednu grupu.
 - b. Ažuriraj matricu udaljenosti grupa kako bi odražavala udaljenosti nove grupe prema ostalim grupama.
 - c. Ukloni dvije grupe koje su upravo spojene i dodaj novu jedinstvenu grupu na popis grupa.
5. Vrati konačnu jedinstvenu grupu, koja predstavlja hijerarhiju grupa [35].

Udaljenost, iz drugog koraka algoritma, između grupa može biti definirana na više načina, uključujući najmanju udaljenost, najveću udaljenost, prosječnu udaljenost ili udaljenost između centroida (ako su grupe veće od jedne točke).



Slika 2.4.1.1 Dendrogram aglomerativnog grupiranja [36]

Iz slike 2.4.1.1 jasno se može iščitati hijerarhijsko grupiranje uzoraka, gdje na dnu dendrograma svaki primjer pripada svojoj grupi, a na vrhu postoji samo jedna grupa.

2.5. Grupiranje po distribuciji

Grupiranje po distribuciji je probabilističko, meko grupiranje. Kod grupiranja po distribuciji, primjeri pripadaju grupi s određenom vjerojatnošću. Glavni problem ovakvih modela je modeliranje funkcije gustoće vjerojatnosti $p(x)$, koja je najčešće složena funkcija.

Složena je iz razloga što je ona zapravo linearna kombinacija K osnovnih razdioba, s nepoznatim parametrima. Cilj ovih modela je utvrđivanje tih nepoznatih parametara i određivanje vjerojatnosti pripadnosti svakog primjera određenoj grupi. Ovakvi modeli nazivaju se miješani (engl. *mixture*) modeli, jer su to „mješavine“ već spomenutih osnovnih razdioba [25]. Sljedeće potpoglavlje razradit će model Gaussovih mješavina.

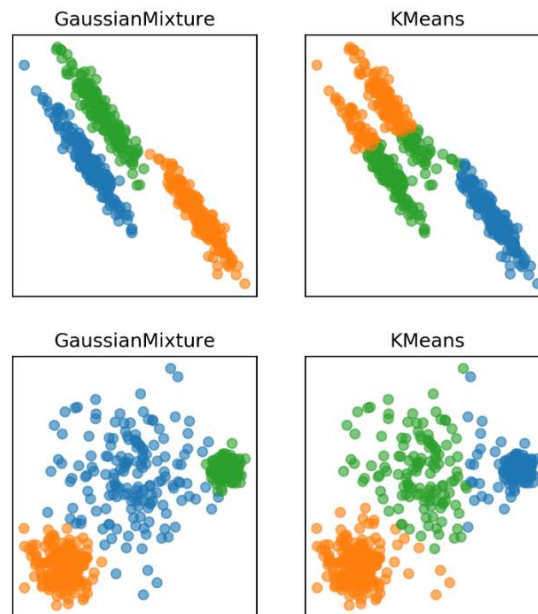
2.5.1. Model Gaussovih mješavina

Model Gaussovih mješavina (engl. *Gaussian Mixture Model*; skraćeno *GMM*) je algoritam mekog grupiranja koji svakom primjeru iz skupa podataka daje određuje vjerojatnost pripadanja određenoj grupi. GMM je funkcija sastavljena od zadanog broja normalnih (Gaussovih) razdioba, od kojih svaka razdioba odgovara jednoj od K grupa iz skupa podataka.

Svaka razdioba ima tri parametra:

- srednju vrijednost μ – srednja vrijednost Gaussove distribucije k -te razdiobe
- kovarijancu Σ – matrica kovarijacija u multi-dimenzionalnom prostoru
- mješovitu vjerojatnost π – vjerojatnost da bilo koji primjer pripada toj grupi

Postupak grupiranja zapravo je algoritam iterativnog ažuriranja ovih parametara u svrhu optimizacije. Ovaj rad neće ulaziti u srž algoritma, već će ga ukratko objasniti. Algoritam se naziva očekivanje-maksimizacija (engl. *Expectation-maximization* skraćeno *EM*). EM algoritam ima dva koraka. Prvi, E-korak, je očekivanje. Ovaj korak procjenjuje s kojom vjerojatnošću tj. „odgovornošću“ svaki primjer pripada svakog grupi. Ova procjena odgovornosti temelji se na trenutnim parametrima $\{\mu, \Sigma, \pi\}$. Drugi, M-korak, naziva se maksimizacija. U ovom koraku provodi se postupak ažuriranja parametara, tako da se maksimizira log-izglednost (engl. *log-likelihood*) da su podaci proizašli iz distribucija sa novim, ažuriranim parametrima. Ova dva koraka se ponavljaju dok ne dođe do konvergencije parametara ili log-izglednosti [45].



Slika 2.5.1.1 Grafički prikaz razlike GMM (lijeve slike) i K-means (desne slike) grupiranja [36]

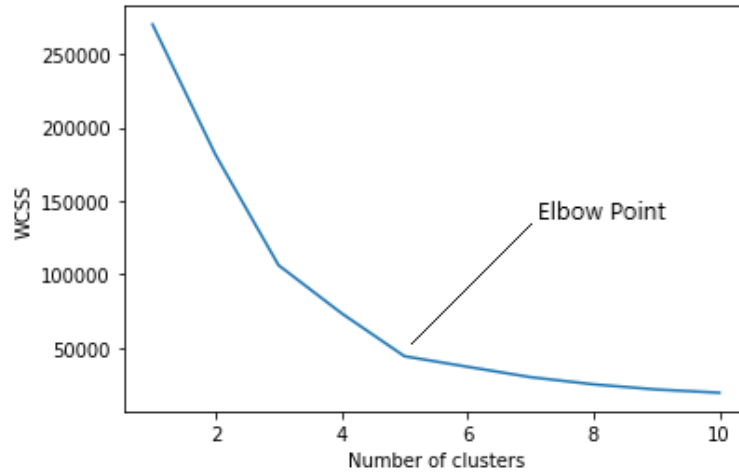
GMM se u nekim situacijama primjenjuje umjesto K-means-a, upravo iz razloga što koristi meko grupiranje. Još jedna prednost GMM-a je što ne pretpostavlja da su grupe otprilike jednake veličine.

2.6. Metode za provjeru grupiranja

Metode za provjeru grupiranja su skup metoda i metrike koje opisuju kvalitetu algoritama grupiranja. Odgovaraju na pitanje „Koliko je dobro grupiranje?“.

Prve ovakve su metode za odabir optimalnog broja grupa. Ovo su metode za određivanje hiperparametra k , kojeg je potrebno odrediti prije samog grupiranja. Ovakve metode primjenjuju se kod raznih algoritama, od kojih su iznimka neparametarski algoritmi (npr. DBSCAN) [30].

Dvije tipične ovakve metode su metoda lakta (engl. *elbow method*) i metoda siluete (engl. *silhouette*). Metoda lakta je grafička metoda u kojoj se traži „lakat“ kriterijske funkcije ovisne o parametru k [25]. Kriterijska funkcija najčešće se računa kao suma kvadrata unutar grupe (engl. *Within-Cluster Sum of Squares*, skraćeno *WCSS*) [30]. Lakat ukazuje da se povećanjem broja grupa neće postići bolje grupiranje.



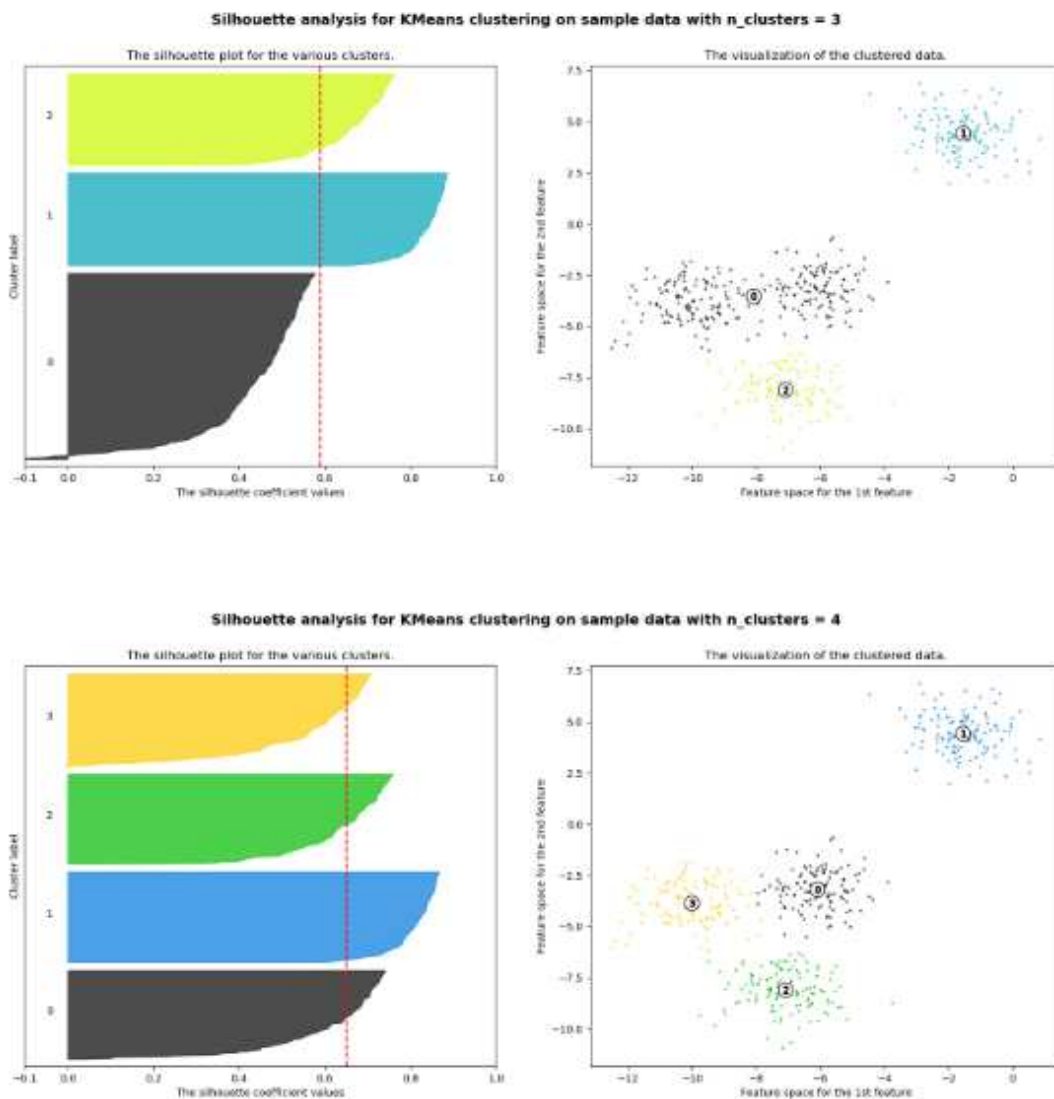
Slika 2.4.1.1 Metoda lakta na grafu kriterijske funkcije u ovisnosti o parametru k [37]

Sa [slike 2.4.1.1](#) se primjećuje da je za ovaj skup podataka optimalan k jednak 5. Povećanjem parametra k neće se znatno smanjiti WCSS, a izgubit će se „prirodno“ grupiranje.

Metoda siluete također je grafička metoda određivanja optimalnog broja grupa. Za svaki primjer i iz skupa podataka računa se vrijednost siluete. Formula siluete je:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (6)$$

Ovdje, $a(i)$ predstavlja prosječnu udaljenost i -tog primjera do ostalih članova iste grupe, dok $b(i)$ označava minimalnu prosječnu udaljenost do primjera u drugim grupama. Vrijednosti siluete su između -1 i 1, gdje vrijednost blizu 1 ukazuje na dobro grupiranje. Analiza siluete uključuje grafički prikaz vrijednosti siluete za sve primjere i i izračunavanje prosječne vrijednosti siluete [30].



Slika 2.4.1.2 Metoda siluete za grupiranje podataka na tri (gornji grafovi) i četiri (donji grafovi) grupe [37]

Iz gornje slike, u analizi siluete za $k=3$, primijećuje se da grupiranje ne rezultira jasno odvojenim grupama, što se očituje u niskim vrijednostima siluete za neke primjere koji su bliži drugim grupama nego svojoj. To ukazuje na preklapanje grupa i potrebu za finijim podešavanjem broja grupa. Za $k=4$, grupe su uniformnije veličine i bolje razdvojene, što rezultira visokim vrijednostima siluete i ukazuje na prirodnije grupiranje. Na ovom skupu podataka, podjela na za četiri grupe izgleda kao bolji izbor zbog bolje ravnoteže među grupama i većom prosječnom vrijednošću siluete.

Treća metrika koju ovaj rad kratko obrađuje je Davies-Bouldin (skraćeno DB) indeks. DB indeks mjeri kvalitetu grupiranja uspoređujući omjer unutargrupne udaljenosti (raspršenost unutar klastera) s međugrupnom udaljenosti (razlika između klastera). Niže vrijednosti ovog

indeksa ukazuju na bolje grupiranje jer sugeriraju da su klasteri dobro odvojeni i da su elementi unutar klastera tijesno grupirani [38].

Četvrta metrika koju ovaj rad kratko obrađuje je Calinski-Harabasz (CH) indeks. CH indeks, također poznat i kao varijacijski omjerni kriterij, mjeri kvalitetu grupiranja uspoređujući disperziju (raspršenost) unutar grupe s međugrupnom disperzijom (razlika između grupa). Više vrijednosti ovog indeksa ukazuju na bolje grupiranje jer sugeriraju da su klasteri dobro odvojeni i da su elementi unutar klastera tijesno grupirani. Indeks se računa kao omjer između zbroja kvadrata udaljenosti centara grupa i zbroja kvadrata udaljenosti primjera unutar svake grupe, normaliziran s brojem primjera i grupa [39].

3. Programsko ostvarenje

Programsko rješenje ovog rada ostvareno je u programskom jeziku *Python*, u okruženju *Jupyter*. Odabrana su dva skupa podataka korisničkih podataka internetskih platformi na kojima su pokazane metode grupiranja. Prvi skup podataka su korisnički podaci internet trgovine, a drugi skup podataka su podaci korisnika kreditnih kartica i njihovih transakcija. U sljedeća dva potpoglavlja detaljnije je pokazana analiza ova dva skupa podataka.

Nužno je istaknuti da se analizom prvog skupa podataka više provela istraživačka analiza podataka, a pokazan je jedan algoritam grupiranja - K-means. S druge strane, analiza korisnika kreditnih kartica više je usredotočena na usporedbu različitih algoritama grupiranja. Glavni razlog ovakve raspodjele analiza je što je prvi skup podataka prikladniji za web analitiku i njene metrike, dok je drugi skup podataka idealan za primjenu i usporedbu algoritama grupiranja.

3.1. Analiza korisničkih podataka internet trgovine

Prvi skup podataka koji je analiziran preuzet je s Kaggle platforme. Kaggle je popularna internet zajednica znanstvenika o podacima i onima koji to žele postati. Skup podataka su podatci ponašanja korisnika na multi kategorijskoj internet trgovini (engl. *E-commerce behavior data from multi category store*) [40]. Ovaj rad se referencira na rad [46]. Prikupljeni podaci su iz mjeseca studenog 2019 godine.

| | event_time | event_type | product_id | category_id | category_code | brand | price | user_id | user_session | |
|----------|----------------------------|------------|------------|---------------------|----------------------------------|---------|---------|-----------|--|--|
| 0 | 2019-11-01 00:00:00 UTC | view | 1003461 | 2053013555631882655 | electronics.smartphone | xiaomi | 489.07 | 520088904 | 4d3b30da-a5e4-49df- b1a8-ba5943f1dd33 | |
| 1 | 2019-11-01 00:00:00 UTC | view | 5000088 | 2053013566100866035 | appliances.sewing_machine | janome | 293.65 | 530496790 | 8e5f4f83-366c-4f70- 860e-ca7417414283 | |
| 2 | 2019-11-01 00:00:01 UTC | view | 17302664 | 2053013553853497655 | | NaN | creed | 28.31 | 561587266 | 755422e7-9040-477b- 9bd2-6a6e8fd97387 |
| 3 | 2019-11-01 00:00:01 UTC | view | 3601530 | 2053013563810775923 | appliances.kitchen.washer | lg | 712.87 | 518085591 | 3bfb58cd-7892-48cc- 8020-2f17e6de6e7f | |
| 4 | 2019-11-01 00:00:01 UTC | view | 1004775 | 2053013555631882655 | electronics.smartphone | xiaomi | 183.27 | 558856683 | 313628f1-68b8-460d- 84f6-cec7a8796ef2 | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| 67501974 | 2019-11-30 23:59:58 UTC | view | 15700137 | 2053013559733912211 | | NaN | NaN | 277.74 | 532714000 | 02b4131c-0112-4231- aafa-ceaa08e77c1b |
| 67501975 | 2019-11-30 23:59:58 UTC | view | 28719425 | 2053013565639492569 | apparel.shoes | baden | 62.81 | 545223467 | 734c5eef-0742-4f8b- 9d22-48f75b0bc359 | |
| 67501976 | 2019-11-30 23:59:59 UTC | view | 1004833 | 2053013555631882655 | electronics.smartphone | samsung | 167.03 | 557794415 | 6fecf566-ebb0-4e70- a243-cdc13ce044cb | |
| 67501977 | 2019-11-30 23:59:59 UTC | view | 2701706 | 2053013563911439225 | appliances.kitchen.refrigerators | samsung | 566.27 | 531607492 | 368ddc8b-5db9-40fb- b7ff-b6582a1192c0 | |
| 67501978 | 2019-11-30 23:59:59 UTC | view | 1004233 | 2053013555631882655 | electronics.smartphone | apple | 1312.52 | 579969851 | 90aca71c-ed8a-4670- 866a-761ebacb732d | |

67501979 rows × 9 columns

Slika 3.1.1 Učitavanje podataka internet trgovine

Podaci su učitani pomoću Pandas biblioteke, koja služi za rad sa skupovima podataka. Skup podataka zauzima 8.38 gigabajta diskovnog prostora, a sadržava oko 67 milijuna redova i 9 stupaca.

```
▶ data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 67501979 entries, 0 to 67501978
Data columns (total 9 columns):
#   Column          Dtype
---  -
0   event_time      object
1   event_type      object
2   product_id      int64
3   category_id     int64
4   category_code   object
5   brand           object
6   price           float64
7   user_id         int64
8   user_session    object
dtypes: float64(1), int64(3), object(5)
memory usage: 4.5+ GB
```

Slika 3.1.2 Informacije o stupcima skupa podataka internetske trgovine

Skup podataka, što se vidi i na slici 3.1.2, sadrži sljedećih devet stupaca odnosno značajki:

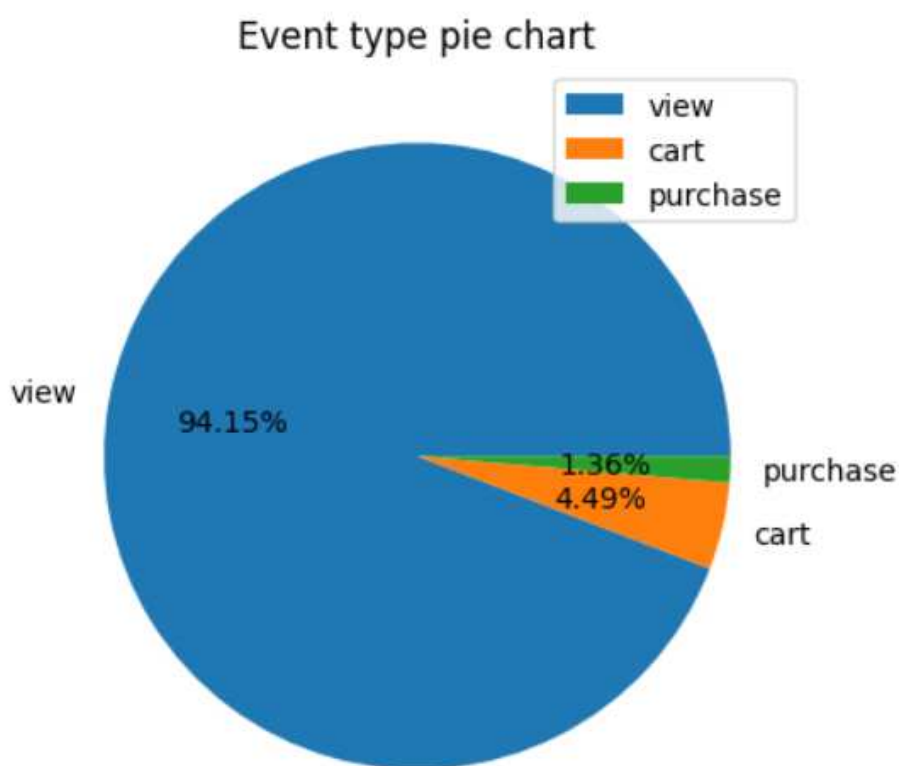
- event_time – vremenska oznaka u kojoj se dogodio događaj (engl. *event*)
- event_type – vrsta događaja, kategorička varijabla koja može biti „view“, „cart“ i „purchase“.
 - view – pogled tj. pregled, korisnik je pogledao određeni proizvod
 - cart – košarica, korisnik je dodao određeni proizvod u košaricu
 - purchase – korisnik je kupio određeni proizvod
- product_id – id proizvoda
- category_id – id kategorije proizvoda
- category_code – kod kategorije proizvoda
- brand – marka proizvođača proizvoda
- price – cijena proizvoda
- user_id – id korisnika koji je ostvario događaj
- user_session – oznaka sjednice za čije je vrijeme korisnik ostvario događaj

Ovakav skup podataka može se smatrati velikim podacima. Također je i računalno zahtjevan. Stoga je za daljnu obradu počišćen i skraćen. Značajke category_code i brand

sadržavaju redom 21 i 9 milijuna nedostajućih vrijednosti, što se može interpretirati na način da neki proizvodi nisu svrstani u nijednu kategoriju, a dosta njih nema ni marku. Budući da gotovo trećina redaka ima ovakvu pojavu, ne bi imalo smisla obrisati ovakve podatke, već su oni preimenovani u „nepoznato“ (engl. *unknown*).

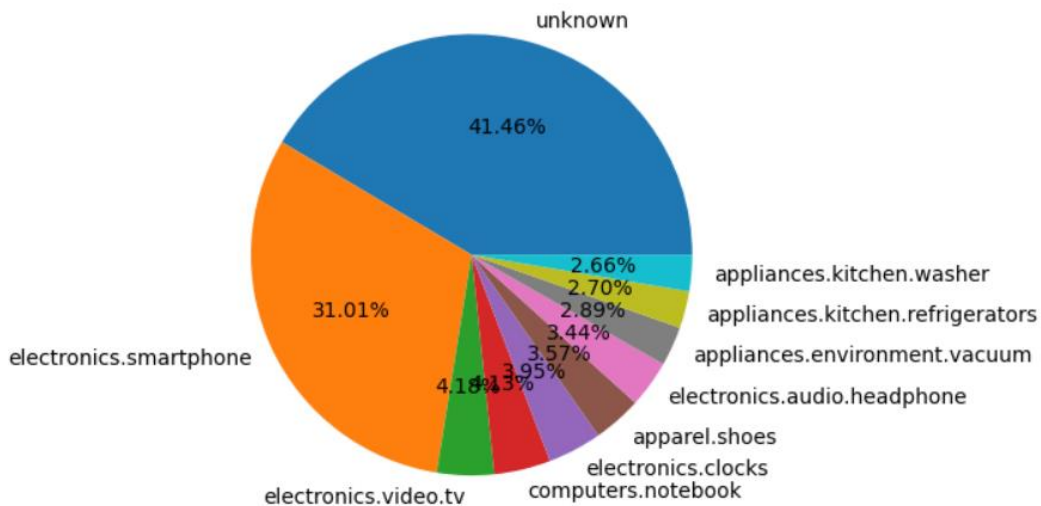
Sljedeći korak čišćenja podataka je promjena tipova podataka značajki `product_id`, `price`, `user_id` i `user_session`. Zatim je iz skupa podataka izbrisana značajka `category_id`. Razlog tome je što je za svrhu ovog rada i analize redundantna.

U koraku istraživačke analize podataka, izvađene su određene metrike web analitike, kao i druge zanimljivosti među podacima.



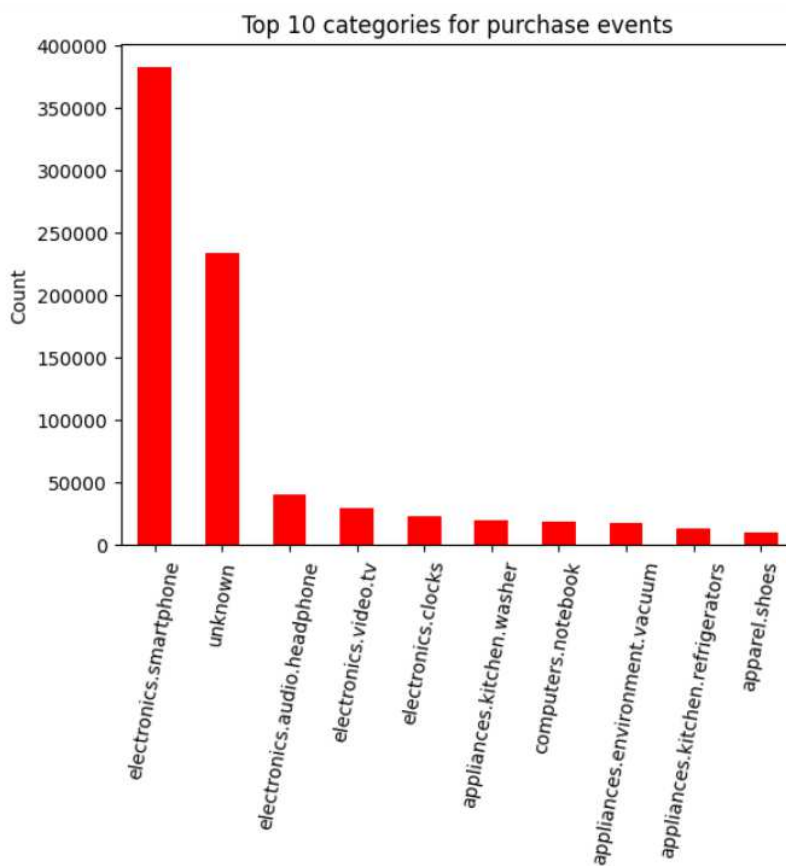
Slika 3.1.3 Strukturni krug (engl. *pie chart*) vrsta događaja (značajke `event_type`)

Slika 3.1.3 prikazuje udio svake vrste događaja u ukupnom broju događaja. Od približno 67 milijuna zabilježenih događaja, 1.36% sačinjava događaj kupnje. Može se reći da je ovo vrijednost stope glavne konverzije ove internet trgovine jer je ciljani događaj kupovina.

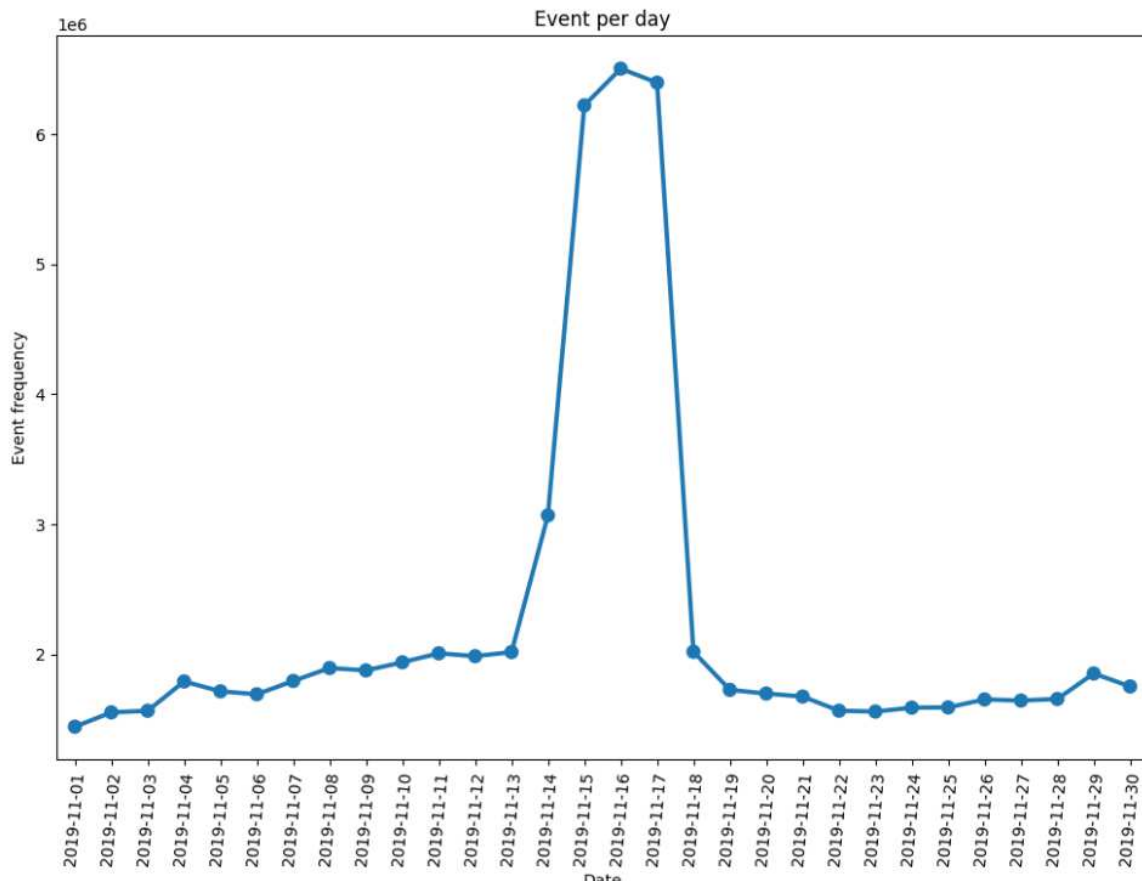


Slika 3.1.4 Strukturalni krug kategorija proizvoda

Slika 3.1.4 prikazuje udio kategorija proizvoda. Iz slike se da iščitati da 41.46% proizvoda nema svoju kategoriju, dok je „electronics.smartphone“ najčešća kategorija kod kategoriziranih proizvoda. Ovo znači da je gotovo trećina svih zabilježenih događaja zabilježena na kategoriji pametnih telefona.



Slika 3.1.5 Deset najkupovanijih kategorija, od kojih su najkupovaniji pametni telefoni



Slika 3.1.6 Broj događaja po danu

Iz slike 3.1.6 se primjećuje neravnomjeran broj događaja, gdje je od 15. – 17.11 značajno povećan broj događaja. Pretpostavka je da je u pitanju crni petak (engl. *black friday*), ili nekakva druga vrsta promotivne ponude na web trgovini.

Nakon istraživačke analize podataka, slijedi grupiranje. Nije praktično u grupiranje uključiti sve značajke iz ovog skupa podataka. Isto tako, u sklopu rada nije računalno izvedivo uključiti sve retke. Stoga je ovaj skup podataka prikladan za RFM (skraćeno od *Recency Frequency Monetary*) analizu.

- Recency – kad je korisnik zadnji put obavio kupnju
- Frequency – koliko često korisnik kupuje
- Monetary – koliko novaca je korisnik potrošio u narudžbama

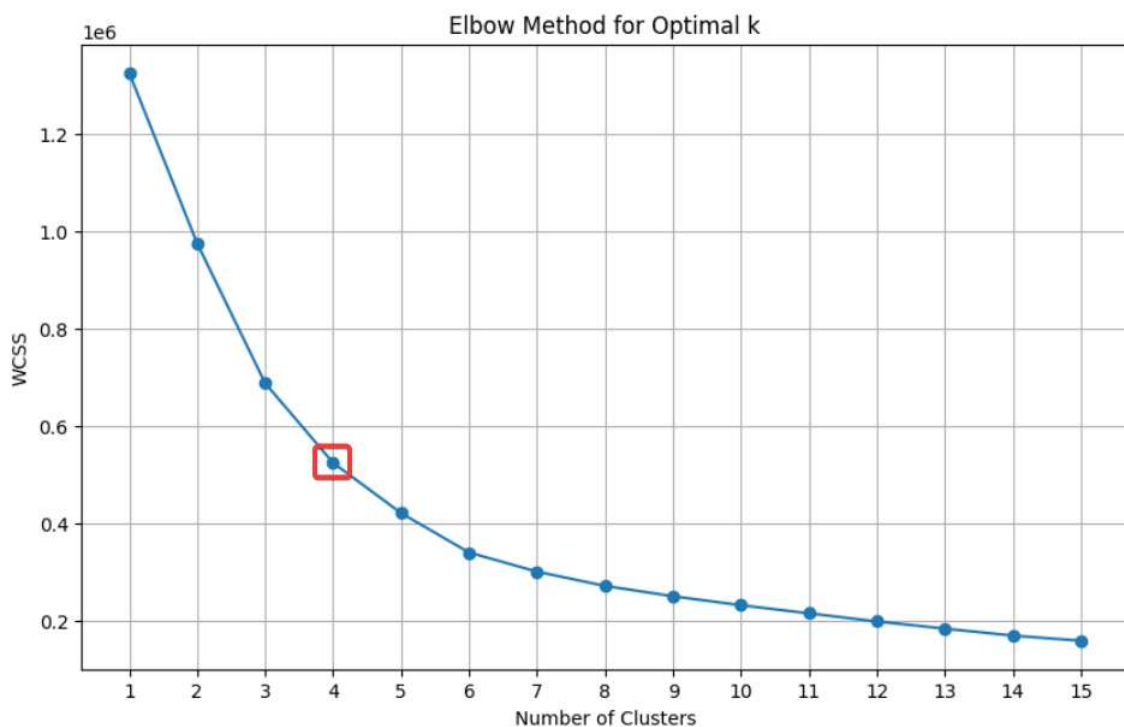
RFM analiza je vrsta segmentacije korisnika uz pomoć koje se pronalazi grupa optimalnih kupaca. Za RFM analizu potrebno je napraviti derivat početnog skupa podataka na način da indeks značajka postane id korisnika, a ostala tri stupca RFM stupci.

| | user_id | recency | frequency | monetary |
|--------|---------|---------|-----------|------------|
| 0 | 2 | 24 | 2 | 224.559998 |
| 1 | 24 | 29 | 1 | 154.160004 |
| 2 | 26 | 17 | 3 | 395.850006 |
| 3 | 32 | 8 | 2 | 84.130005 |
| 4 | 35 | 29 | 1 | 531.260010 |
| ... | ... | ... | ... | ... |
| 441633 | 3695972 | 0 | 2 | 401.849976 |
| 441634 | 3695994 | 0 | 1 | 261.269989 |
| 441635 | 3695998 | 0 | 1 | 334.600006 |
| 441636 | 3696037 | 0 | 1 | 124.110001 |
| 441637 | 3696060 | 0 | 1 | 124.110001 |

441638 rows × 4 columns

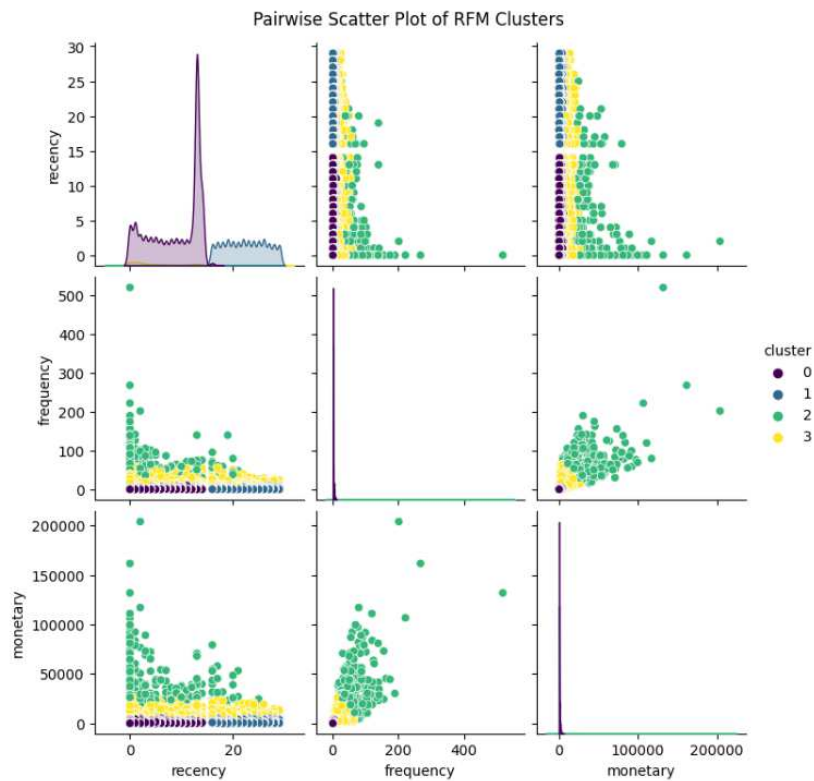
Slika 3.1.7 Derivirani skup podataka za RFM analizu

Novi, derivirani skup podataka ima 441638 redaka i 4 stupca. Retci su svi jedinstveni korisnici, a ostali stupci su njihove pripadne RFM metrike. Potom je potrebno ovaj skup podataka skalirati. Skaliranje je obavljeno klasom *StandardScaler* iz *sklearn.preprocessing Python* biblioteke. *StandardScaler* standardizira podatke. Od svakog podatka oduzme se srednja vrijednost, a onda ga se podjeli s varijancom. Tako se podaci ponašaju prema normalnoj distribuciji. Ovakvi skalirani podatci dani su na grupiranje. Za ovaj skup podataka korišten je algoritam K-means. Metodom lakta određen je optimalan broj grupa k .



Slika 3.1.8 Metoda lakta za određivanje optimalnog broja grupa

Optimalan broj grupa je $k=4$, što se vidi iz slike 3.1.8. Sljedeći korak je prikaz grafova grupa.



Slika 3.1.9 Točkasti graf (engl. scatter plot) parova RFM skupa podataka za $k=4$

Iz slike 3.1.9. vidi se prikaz svih parova RFM značajki. Važno je istaknuti da je „recency“ svim korisnicima u intervalu od trideset dana.



Slika 3.1.10 Točkasti graf (engl. scatter plot) parova RFM skupa podataka za $k=4$

Slika 3.1.10 prikazuje dijagram raspršenja značajki „monetary“ i „frequency“. Opis grupa je sljedeći:

- Grupa 0 – korisnici koji najrijeđe kupuju, najdavnije su zadnji put kupili i ne troše puno
- Grupa 1 – korisnici koji često kupuju, relativno nedavno su kupili i puno troše
- Grupa 2 – korisnici koji najčešće kupuju, najrecentnije su zadnji put kupili i najviše troše
- Grupa 3 – korisnici koji ne kupuju često, nisu odavno kupili i ne troše puno, ali više nego grupa 0

Naravno, postoje primjeri npr. iz grupe 1 koji češće i više troše nego neki primjeri iz grupe 2. Razlog tome je što u grupiranje ulazi i treća značajka „recency“, a ovo daje prednost kupcima čiji je „recency“ najmanji.

3.2. Analiza podataka korisnika kreditnih kartica

Drugi skup podataka koji je analiziran također je preuzet s Kaggle platforme [41]. Skup podataka sadrži podatke o ponašanju korisnika kreditnih kartica, a prikupljen je na razini korisnika i obuhvaća razdoblje od šest mjeseci. Podaci uključuju 18 varijabli ponašanja korisnika. Detaljna analiza ovog skupa podataka uključuje primjenu različitih algoritama grupiranja s ciljem usporedbe njihove učinkovitosti. U nastavku je prikazana analiza i rezultati grupiranja. Ova analiza referencira se na analizu [42] i [43].

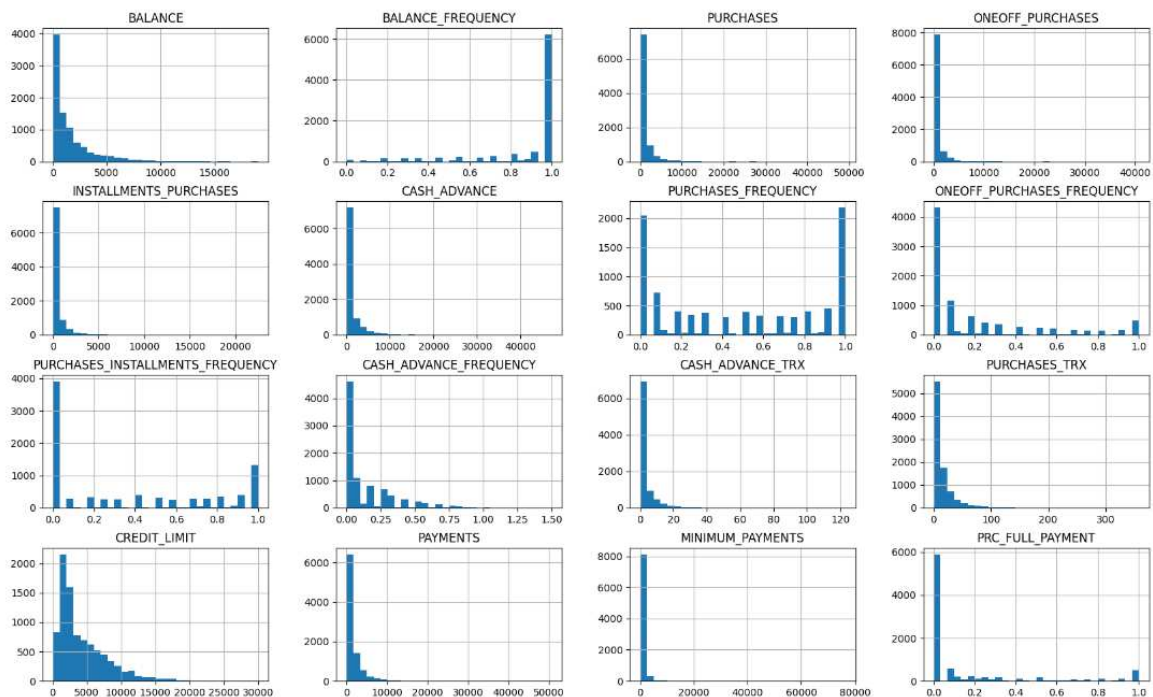
3.2.1. Učitavanje, pregled i vizualizacija podataka

Podaci su učitani pomoću Pandas biblioteke. Skup podataka sadržava 8950 redova i 18 stupaca. Skup podataka sadrži sljedeće značajke:

- CUST_ID - id korisnika kreditne kartice (kategorijska varijabla)
- BALANCE – iznos preostalog stanja na računu
- BALANCE_FREQUENCY – koliko često se stanje ažurira, ocjena između 0 i 1
- PURCHASES – iznos obavljenih kupnji s računa
- ONEOFF_PURCHASES – najveći iznos jednokratne kupnje
- INSTALLMENTS_PURCHASES – iznos kupnji obavljenih na rate
- CASH_ADVANCE – gotovina koju korisnik unaprijed koristi

- PURCHASES_FREQUENCY – koliko često se obavljaju kupnje, ocjena između 0 i 1
- ONEOFF_PURCHASES_FREQUENCY – koliko često se obavljaju jednokratne kupnje, ocjena između 0 i 1
- PURCHASES_INSTALLMENTS_FREQUENCY – koliko često se obavljaju kupnje na rate, ocjena između 0 i 1
- CASH_ADVANCE_FREQUENCY – koliko često se koristi gotovina unaprijed
- CASH_ADVANCE_TRX – broj transakcija gotovine unaprijed
- PURCHASES_TRX – broj kupovnih transakcija
- CREDIT_LIMIT – limit kreditne kartice za korisnika
- PAYMENTS – iznos plaćanja obavljenih od strane korisnika
- MINIMUM_PAYMENTS – minimalni iznos plaćanja obavljenih od strane korisnika
- PRC_FULL_PAYMENT – postotak potpunih plaćanja obavljenih od strane korisnika
- TENURE – trajanje korištenja usluge kreditne kartice od strane korisnika

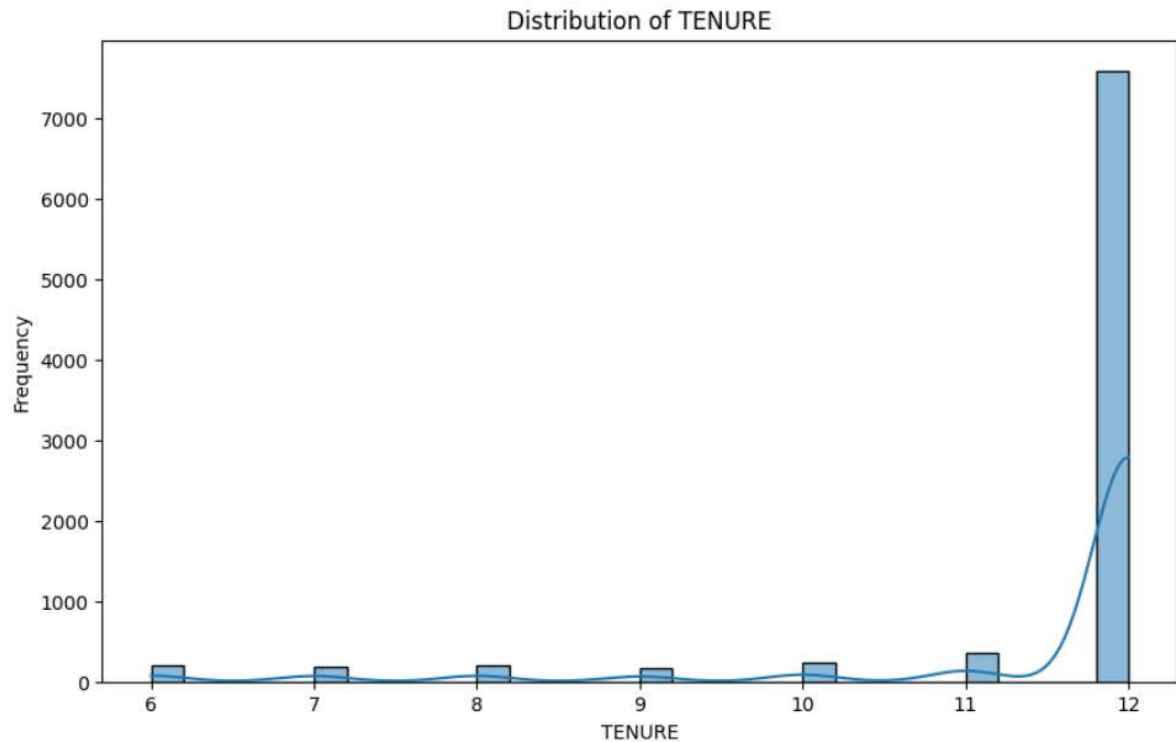
U svrhu daljnjeg pregleda podataka, prikazane su distribucije svih numeričkih varijabli, što su sve značajke osim CUST_ID.



Slika 3.2.1.1 Distribucije gotovo svih numeričkih varijabli

Iz slike 3.2.1.1 očituje se da su distribucije odstupaju od normalne i imaju pozitivnu asimetričnost (engl. *skewness*). Ova pojava vjerojatno je posljedica činjenice da su većina značajki mjere frekvencije.

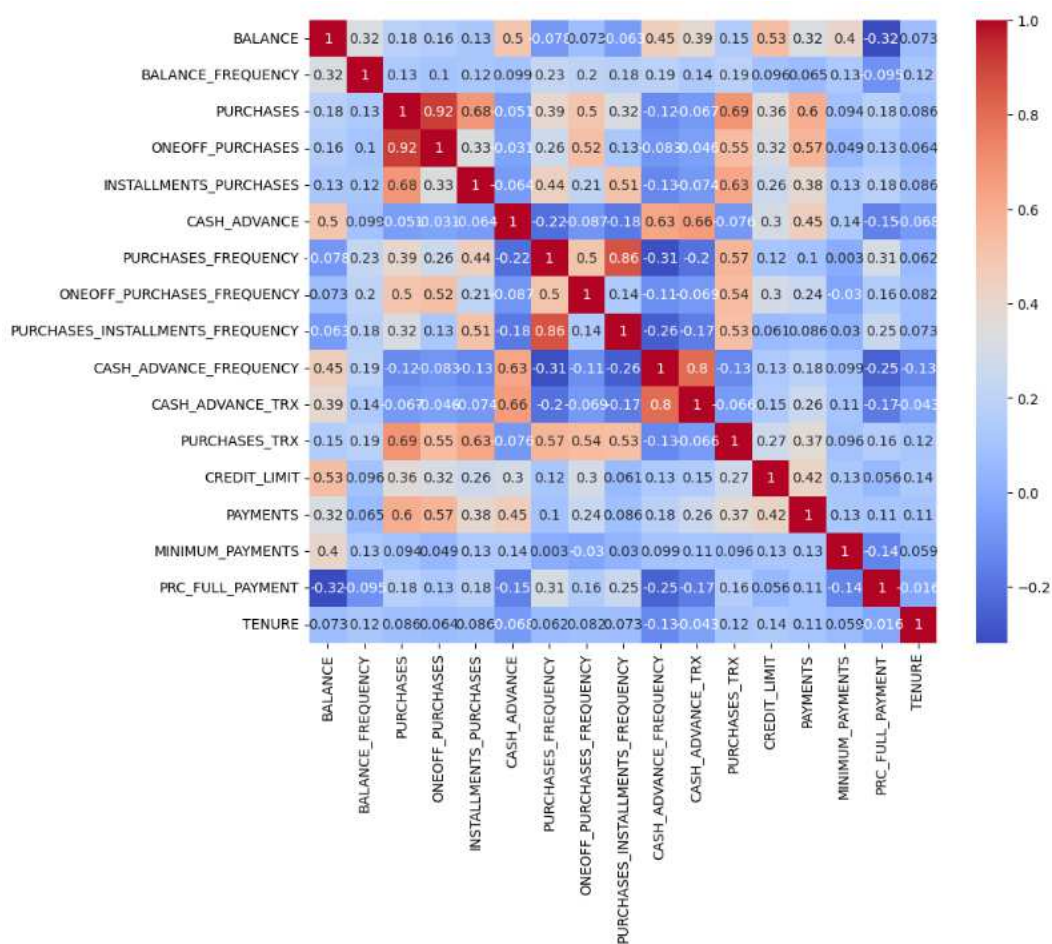
Od svih distribucija, izdvojene su neke posebno zanimljive.



Slika 3.2.1.2 Distribucije značajke TENURE

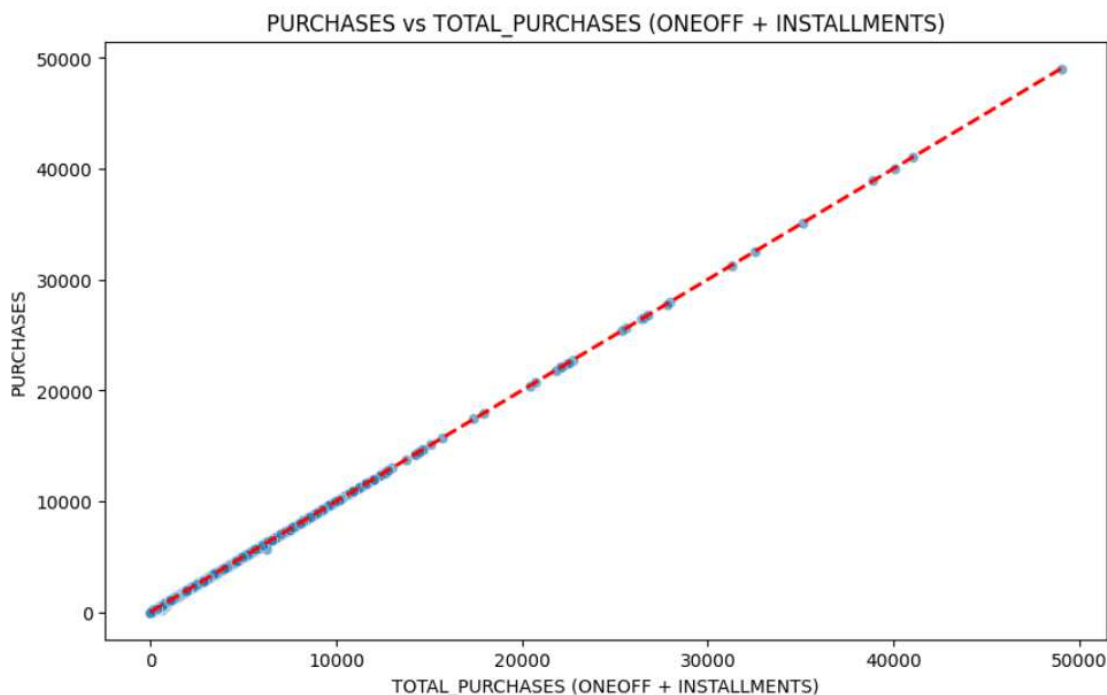
Na slici 3.2.1.2 prikazan je distribucija značajke TENURE, koja govori koliko godina korisnici koriste uslugu kreditnih kartica. Primjećuje se da su više od 80% korisnika korisnici najmanje 12 godina.

Nakon vizualizacije distribucije, prikazana je matrica korelacija numeričkih značajki. Matrica korelacija prikazuje linearnu zavisnost među varijablama iskazanu u terminima Pearsonovog koeficijenta korelacije. Snaga povezanosti mjeri se apsolutnom vrijednošću koeficijenta, dok predznak ukazuje na smjer povezanosti.



Slika 3.2.1.4 Matrica korelacija numeričkih varijabli

Iz slike 3.2.1.4 primjećuje se određena korelacija među varijablama. Primjer ovome su CREDIT_LIMIT i BALANCE, s korelacijom od 0.53. Razumno je da su ove dvije značajke u korelaciji jer osobe s većim kreditnim limitom češće koriste kreditne kartice. Također, korisnici s većim balansom na računu su pouzdaniji, pa im se tolerira i veći kreditni limit. Drugi primjer su PURCHASES_TRX i PURCHASES, sa korelacijom jednakom 0.69. PURCHASES_TRX je, već spomenuto, broj transakcija, a PURCHASES je ukupan iznos kupnji. Skroz je razumno da korisnici koji frekventno kupuju i potroše više. Značajka PURCHASES je zbroj značajki ONEOFF_PURCHASES i INSTALLMENTS_PURCHASES. Ovo ima smisla jer je iznos obavljenih kupnji jednak zbroju jednokratnih kupnji i kupnji na rate.



Slika 3.2.1.5 Linearna zavisnost značajki vezanih za kupnju

3.2.2. Obrada podataka i skaliranje

Sljedeća izazovna stvar u ovom skupu podataka je popunjavanje nedostajućih vrijednosti. Jedino značajka `PRC_FULL_PAYMENT` ima nedostajuće vrijednosti, njih 313. `PRC_FULL_PAYMENT` je, gore već navedeno, postotak potpunih plaćanja iznos dugovanja na kartici. Nedostajuće vrijednosti popunjene su sa srednjom vrijednošću značajke.

Iz skupa podataka izbačena je kategorička značajka `CUST_ID`, iz razloga što je nepotrebna za daljnju analizu.

Potom su podatci standardizirani klasom *StandardScaler*, u svrhu daljnje obrade i analize.

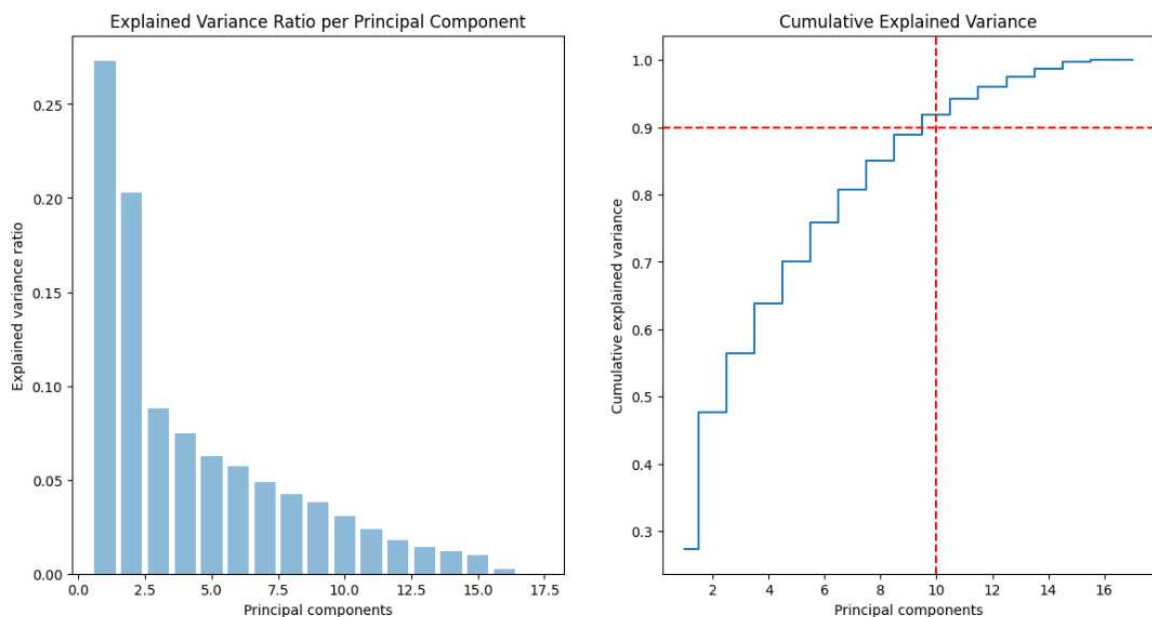
3.2.3. Analiza glavnih komponenti (PCA)

Analiza glavnih komponenti (engl. *Principal component analysis*) je tehnika strojnog učenja koja služi za smanjenje dimenzionalnosti skupa podataka. Dimenzionalnost skupa podataka se smanji na način da se izračunaju glavne komponente – linearne kombinacije trenutnih značajki koje objašnjavaju što je više varijance moguće. Ovaj rad ne ulazi u dubinu PCA tehnike, ali objašnjava njegove glavne točke. Svojstva i pretpostavke primjene PCA algoritma su:

1. Standardizacija podataka – podaci moraju biti normalizirani jer PCA prije obrade pretpostavlja normalnost
2. Računanje matrice kovarijance – ova matrica pokazuje linearnu zavisnost značajki
3. Izračun svojstvenih vektora i svojstvenih vrijednosti – svojstveni vektori i svojstvene vrijednosti matrice kovarijance određuju smjer i veličinu glavnih komponenti. Vlastiti vektori predstavljaju smjer, dok vlastite vrijednosti pokazuju veličinu varijance duž tih smjerova.
4. Izbor glavnih komponenti – glavne komponente postaje novi skup varijabli koji je linearna kombinacija postojećih. Biraju se po veličini njihovih svojstvenih vrijednosti, kako bi se zadržala što veća varijanca.
5. Transformacija početnog skupa podataka – početni skup podataka zamjenjen je s novim skupom, smanjene dimenzionalnosti [44]

PCA je koristan u grupiranju. Razlog tome je što se povećanjem dimenzionalnosti kod algoritama koji koriste metrike kao što su euklidska udaljenost smanjuje korisnost grupiranja. U ovom radu proveden je PCA sa 90% očuvane varijance.

Number of components to retain 90% variance: 10

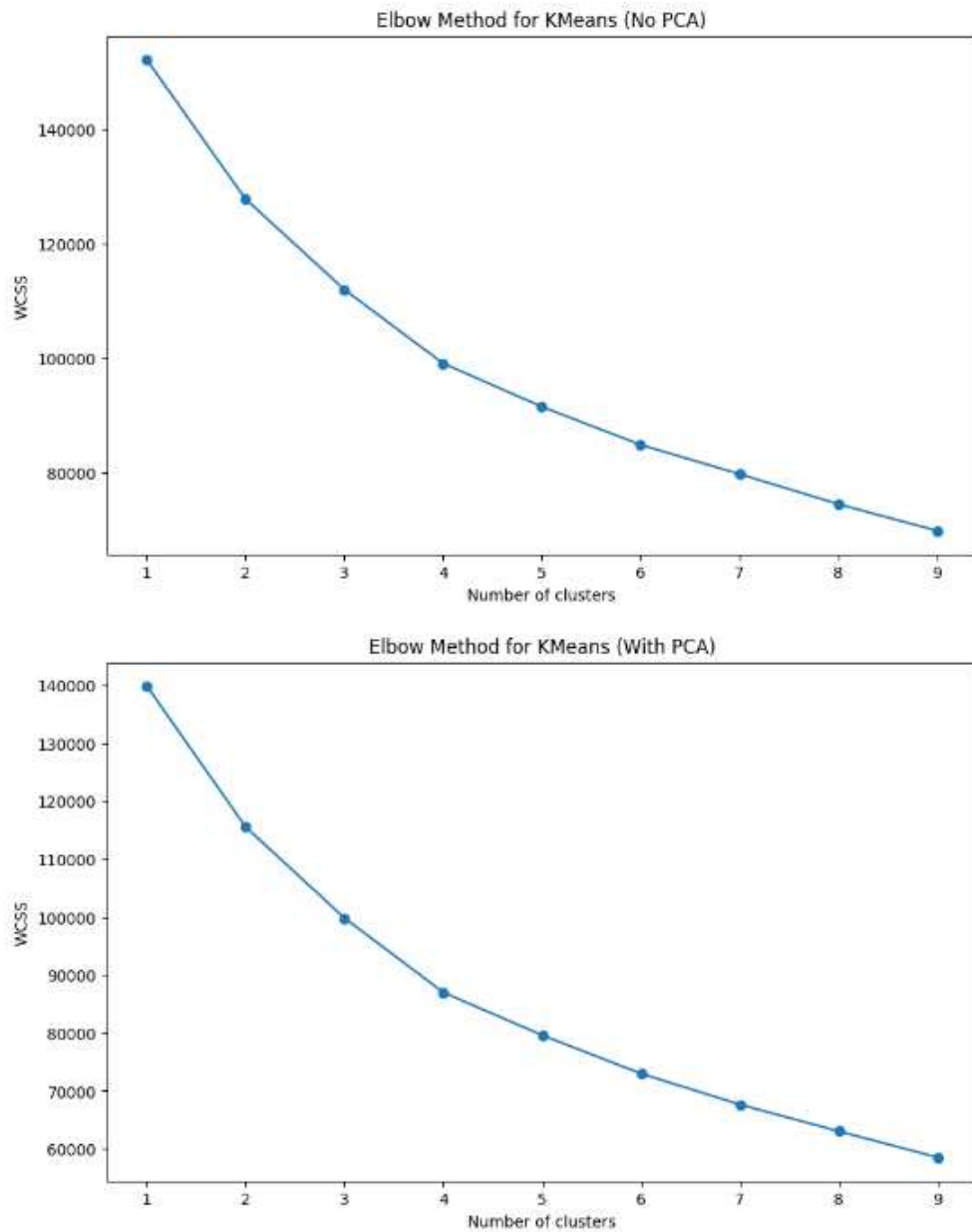


Slika 3.2.3.1 Primjena PCA na skup podataka

Iz slike 3.2.3.1 primjećuje se da je skup podataka smanjen s 18 na 10 značajki, uz 90% očuvanje varijance.

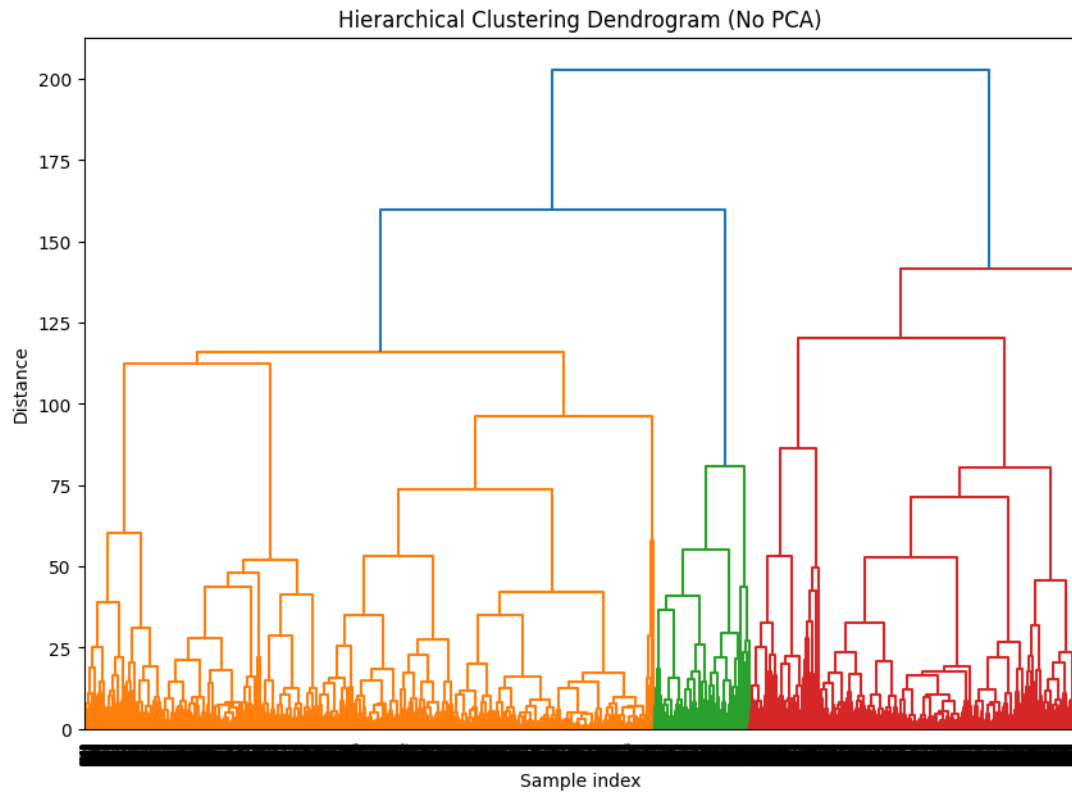
3.2.4. Grupiranje korisnika kreditnih kartica

U fazi grupiranja, provedena su četiri algoritma: K-means, hijerarhijsko aglomerativno grupiranje, DBSCAN i GMM. Svaki od ovih algoritama proveden je kroz standardizirani skup podataka sa i bez PCA. Ispitana je učinkovitost PCA na ovom skupu podataka. Svaki algoritam testiran je na četiri grupe. Također je svaki algoritam testiran sa metrikama siluete, DB indeksa i CB indeksa. Konačno, napravljena je tablica koja prikazuje i uspoređuje rezultate uspješnosti grupiranja svih algoritama.

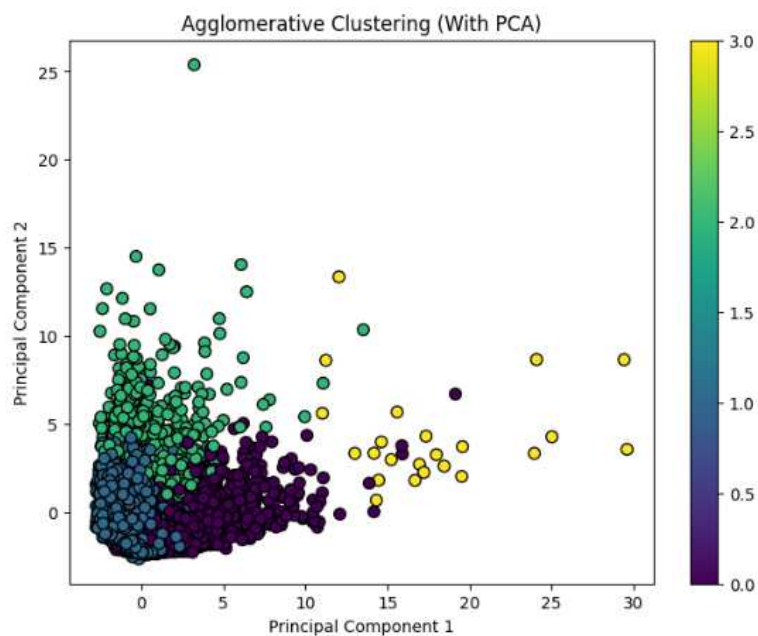


Slika 3.2.4.1 Metoda lakta K-means algoritma za skupove podataka sa i bez PCA

Metodom lakta određen je optimalan $k=4$ za K-means, za skupove podataka sa i bez PCA. Kako bi usporedba bila konzistentna, svaki sljedeći algoritam koristio je grupaciju na 4 grupe. Grafički prikaz aglomerativnog grupiranja prikazan je na sljedeće dvije slike.

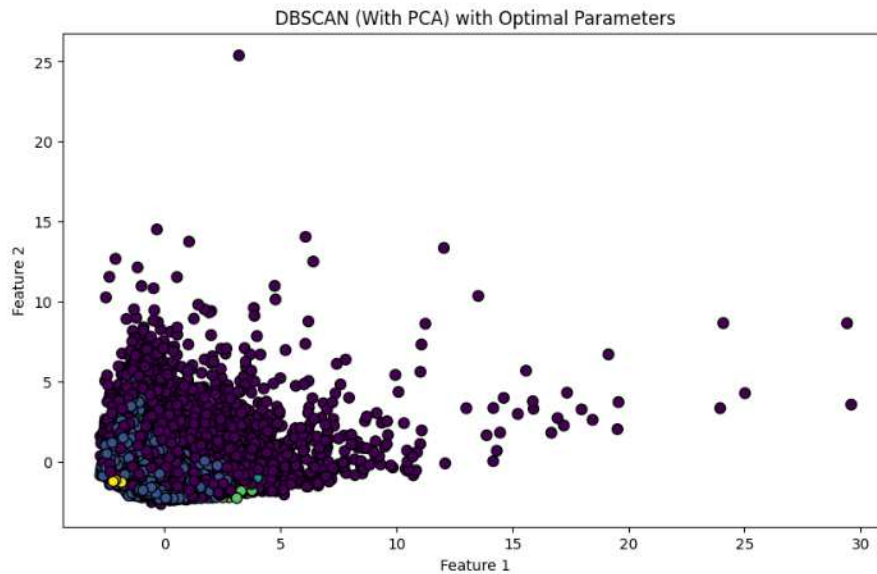


Slika 3.2.4.3 Dendrogram aglomerativnog grupiranja na skupu podataka bez PCA



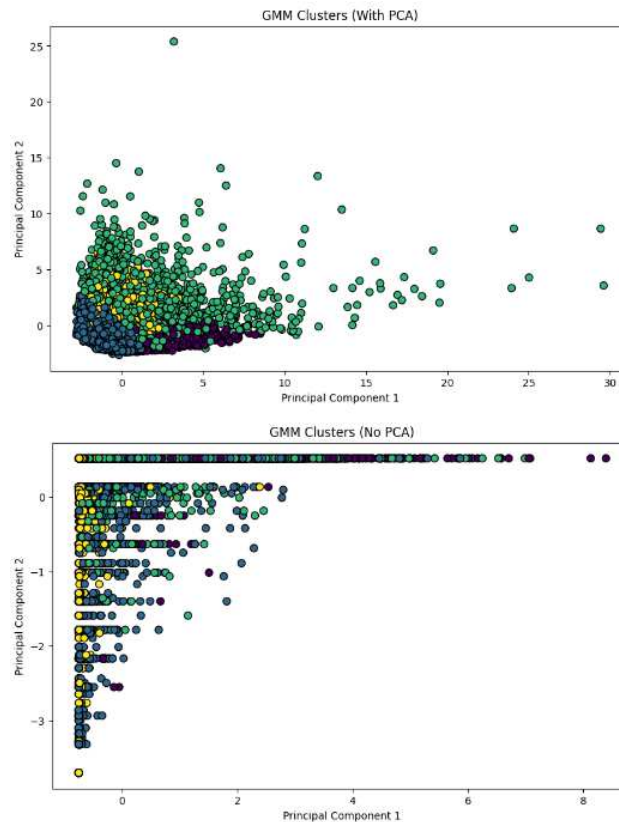
Slika 3.2.4.4 Dendrogram aglomerativnog grupiranja na skupu podataka bez PCA

Sljedeći algoritam, DBSCAN, neparametarska je metoda koja ne pretpostavlja razdiobu podataka, stoga je potrebno izračunati optimalne parametre ϵ i m , umjesto da ima hiperparametar k .



Slika 3.2.4.5 DBSCAN grupiranje na PCA skupu podataka

Posljedni algoritam grupiranja koji je analiziran je GMM.



Slika 3.2.4.5 GMM grupiranje na oba skupa podataka

3.2.5. Rezultati grupiranja

Rezultati grupiranja su, kako je prethodno objašnjeno, uspoređeni na 3 metrike. Rezultati su prikazani slikom (tablicom).

| | Algorithm | Silhouette Score (No PCA) | Davies-Bouldin Score (No PCA) | Calinski-Harabasz Score (No PCA) | Silhouette Score (PCA) | Davies-Bouldin Score (PCA) | Calinski-Harabasz Score (PCA) |
|---|--------------------------|---------------------------|-------------------------------|----------------------------------|------------------------|----------------------------|-------------------------------|
| 0 | KMeans | 0.197587 | 1.575238 | 1597.207111 | 0.217021 | 1.479424 | 1812.726854 |
| 1 | Agglomerative Clustering | 0.182381 | 1.501720 | 1186.266148 | 0.204841 | 1.393242 | 1354.755575 |
| 2 | DBSCAN | 0.085045 | 2.035785 | 308.698167 | 0.047635 | 1.836761 | 255.446772 |
| 3 | GMM | 0.102092 | 2.407619 | 841.134840 | 0.073387 | 2.331728 | 819.323417 |

Slika 3.2.5.1 Rezultati i usporedba algoritama

Od dva skupa podataka, iz rezultata se vidi da PCA većinski pridonosi boljim rezultatima grupiranja za sve algoritme i metrike osim siluete i CH za DBSCAN i GMM.

Što se samih algoritama tiče, K-means pokazuje najbolje rezultate, kako za prvi tako i za drugi skup podataka.

Sljedeći, po rezultatima, blizu K-meansu, aglomerativni je algoritam.

Treći po rezultatima je GMM, a najgori je DBSCAN.

4. Zaključak

Zaključak ovog rada temelji se na praktičnoj primjeni algoritama grupiranja za analizu korisničkih podataka internetskih platformi, s posebnim naglaskom na web analitiku i analitiku financijskih platformi.

Teorijski dio rada pružio je temeljito objašnjenje važnosti web analitike i analitike financijskih platformi, uključujući povijesni razvoj, ključne pojmove i najčešće korištene alate. Također su detaljno opisani algoritmi grupiranja, kao što su K-means, DBSCAN, hijerarhijsko grupiranje i model Gaussovih mješavina, zajedno s njihovim prednostima, nedostacima i kriterijima za odabir.

Praktični dio uključivao je analizu dvaju skupova podataka – korisničke podatke internetske trgovine i podatke korisnika kreditnih kartica. Kroz analizu podataka internetske trgovine korišten je algoritam K-means, što je omogućilo segmentaciju korisnika na temelju RFM analize. Rezultati su pokazali identificiranje različitih korisničkih skupina koje su se razlikovale po frekvenciji kupnje, potrošnji i vremenskoj recentnosti.

Analiza korisničkih podataka kreditnih kartica obuhvatila je usporedbu četiri algoritma grupiranja: K-means, hijerarhijsko aglomerativno grupiranje, DBSCAN i GMM. Korištenjem različitih metrika za procjenu kvalitete grupiranja, kao što su metoda siluete, DB indeks i CH indeks, utvrđeno je da algoritam K-means daje najbolje rezultate u oba skupa podataka, dok je DBSCAN pokazao najmanje zadovoljavajuće rezultate.

Primjena analize glavnih komponenti (PCA) dodatno je poboljšala rezultate grupiranja za većinu algoritama, smanjujući dimenzionalnost podataka i povećavajući učinkovitost grupiranja. Rezultati praktičnog dijela pokazuju da pravilna primjena algoritama grupiranja može poboljšati segmentaciju korisnika, što je ključno za donošenje informiranih poslovnih odluka i optimizaciju poslovnih procesa u različitim industrijama.

Literatura

- [1] Web analytics, TechTarget. Poveznica: <https://www.techtarget.com/searchbusinessanalytics/definition/Web-analytics>; pristupljeno 24. svibnja 2024.
- [2] Hotjar. (n.d.). Web analytics. Poveznica: <https://www.hotjar.com/web-analytics/>; pristupljeno 24. svibnja 2024.
- [3] Contentsquare. (n.d.). A Brief History of Web Analytics. Poveznica: <https://contentsquare.com/blog/a-brief-history-of-web-analytics>; pristupljeno 25. svibnja 2024.
- [4] Franco, L., & Valdés, M. (2021, February 23). History of Google Analytics. Justia Onward. Poveznica: <https://onward.justia.com/history-of-google-analytics/>; pristupljeno 25. svibnja 2024.
- [5] Leady.com. (n.d.). A brief history of website analytics. Poveznica: <https://leady.com/blog/a-brief-history-of-website-analytics>; pristupljeno 25. svibnja 2024.
- [6] Ryan, D. (2014, May 4). Understanding Digital Marketing. Kogan Page. Poveznica: <https://www.koganpage.com/marketing-communications/understanding-web-analytics-and-key-performance-indicators>; pristupljeno 26. svibnja 2024.
- [7] Author Unknown. (n.d.). Web Analytics: Essential KPIs for Measuring Website Performance. LinkedIn. Poveznica: <https://www.linkedin.com/pulse/web-analytics-essential-kpis-measuring-website-performance->; pristupljeno 26. svibnja 2024.
- [10] Natalia Chronowskam, What is first-party data and how does it benefit your marketing. 30.1.2024., URL: <https://piwik.pro/blog/first-party-data-value/>; pristupljeno 27. svibnja 2024.

[11] Małgorzata Poddębniak, “Cookieless future” is just a buzzword – Here is all you need to know about the end of third-party cookies, 22.4.2024., Poveznica: <https://piwik.pro/blog/the-end-of-third-party-cookies/>; pristupljeno 1. lipnja 2024.

[12] David Street, Karolina Lubowicka, Sebastian Synowiec, Małgorzata Poddębniak, Anonymous tracking: How to do useful analytics without personal data, 4.7.2023., Poveznica: <https://piwik.pro/blog/how-to-do-useful-analytics-without-personal-data/>; pristupljeno 1. lipnja 2024.

[13] McKinsey & Company, What is Fintech?, 16.1.2024. URL <https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-fintech>, pristupano: 1. lipnja 2024.

[14] Max Amies, Harnessing the Power of Data Analytics in Fintech: Strategies for Success, 16.1.2024. URL <https://datasearchconsulting.com/harnessing-the-power-of-data-analytics-in-fintech-strategies-for-success>, pristupano: 2. lipnja 2024.

[15] Payments Association, Fintech: The History and Future of Financial Technology, 12.10.2020. URL <https://thepaymentsassociation.org/article/fintech-the-history-and-future-of-financial-technology/>, pristupano: 3. lipnja 2024.

[16] Anant Deshpande, Big Data and its Role in FinTech Evolution, 31.1.2024. URL <https://www.expresscomputer.in/guest-blogs/big-data-and-its-role-in-fintech-evolution/108709/>, pristupano: 3. lipnja 2024.

[17] Max Amies, Harnessing the Power of Data Analytics in Fintech: Strategies for Success, 16.1.2024. URL <https://datasearchconsulting.com/harnessing-the-power-of-data-analytics-in-fintech-strategies-for-success>, pristupano: 4. lipnja 2024.

[18] Lvivity, Fintech Statistics & Facts: Growth, Market Share, Trends and Forecasts, 19.8.2021. URL <https://lvivity.com/fintech-statistics-trends-and-forecasts>, pristupano: 5. lipnja 2024.

- [19] ProjectPro, Big Data Use Cases: How PayPal leverages Big Data Analytics, 14.4.2024. URL <https://www.projectpro.io/article/big-data-use-cases-how-paypal-leverages-big-data-analytics/231>, pristupano: 6. lipnja 2024.
- [20] Databricks, Hyper-Personalization Accelerator for Banks and Fintechs Using Credit Card Transactions, 16.1.2024. URL <https://www.databricks.com/blog/2024/01/16/hyper-personalization-accelerator-for-banks-and-fintechs-using-credit-card-transactions>, pristupano: 7. lipnja 2024.
- [21] McKinsey, Designing next-generation credit-decisioning models, 15.1.2024. URL <https://www.mckinsey.com/industries/financial-services/our-insights/designing-next-generation-credit-decisioning-models>, pristupano: 7. lipnja 2024.
- [22] Miquido, Predictive Analytics in Fintech: Benefits, Use Cases, 14.1.2024. URL <https://www.miquido.com/blog/predictive-analytics-in-fintech-benefits-use-cases>, pristupano: 8. lipnja 2024.
- [23] Just Analytics, Credit Card Analysis: Leverage Your Cardholders Insights, 13.1.2024. URL <https://www.justanalytics.com/blog/credit-card-analysis-leverage-your-cardholders-insights>, pristupano: 9. lipnja 2024.
- [24] DataToBiz, 5 Applications of Data Science in FinTech: FinTech Secrets, 14.1.2024. URL <https://www.datatobiz.com/blog/5-applications-of-data-science-in-fintech-fintech-secrets>, pristupano: 10. lipnja 2024.
- [25] Šnajder, J., & Dalbelo Bašić, B. (2014). Strojno učenje. Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva.
- [26] Shreya Joshi, Types of Clustering Algorithms in Machine Learning: With Examples, 5.8.2022. URL <https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms/>, pristupano: 11. lipnja 2024.
- [27] Ungaro, Tea. "KLAŠTERSKA ANALIZA." (2016), Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet

[28] Pulkit Sharma, "Comprehensive Guide to K-Means Clustering," Analytics Vidhya, 7.8.2019. URL: <https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/>, pristupano: 12. lipnja 2024.

[29] GeeksforGeeks, "ML | K-Medoids Clustering with Example," URL: <https://www.geeksforgeeks.org/ml-k-medoids-clustering-with-example/>, pristupano: 12. lipnja 2024.

[30] Jan Šnajder, "Grupiranje," Strojno učenje 1, UNIZG FER, ak. god. 2021./2022., predavanja, v2.1.

[31] Habijanić, Ana. "Grupiranje podataka." Master's thesis, Sveučilište Josipa Jurja Strossmayera u Osijeku, Odjel za matematiku, 2020.

[32] Automatic Addison, "K-Means Clustering and the Local Search Problem," 27.6.2019. URL: <https://automaticaddison.com/k-means-clustering-and-the-local-search-problem/>, pristupano: 13. lipnja 2024.

[33] Abhishek Sharma, "How Does DBSCAN Clustering Work?" Analytics Vidhya, 23.9.2020. URL: <https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/>, pristupano: 14. lipnja 2024.

[34] GeeksforGeeks, "Hierarchical Clustering," URL: <https://www.geeksforgeeks.org/hierarchical-clustering/>, pristupano: 15. lipnja 2024.

[35] Computing For All, "Hierarchical Agglomerative Clustering (HAC) Algorithm," 2021. URL: <https://computing4all.com/courses/introductory-data-science/lessons/hierarchical-agglomerative-clustering-hac-algorithm/>, pristupano: 16. lipnja 2024.

[36] Andreas C. Müller, "Clustering and Mixture Models," Applied Machine Learning in Python, 2020. URL: <https://amueller.github.io/aml/03-unsupervised-learning/02-clustering-mixture-models.html>, pristupano: 16.6.2024.

- [37] Harika Bonthu, "In-depth Intuition of K-Means Clustering Algorithm in Machine Learning," Analytics Vidhya, 6.1.2021. URL: <https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/>, pristupano: 17.6.2024.
- [38] GeeksforGeeks, "Davies-Bouldin Index," URL: <https://www.geeksforgeeks.org/davies-bouldin-index/>, pristupano: 20.6.2024.
- [39] GeeksforGeeks, "Calinski-Harabasz Index - Cluster Validity Indices | Set 3," URL: <https://www.geeksforgeeks.org/calinski-harabasz-index-cluster-validity-indices-set-3/>, pristupano: 22.6.2024.
- [40] Michael Kechinov, "eCommerce behavior data from multi category store" URL: <https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store>, pristupano: 22.6.2024.
- [41] Arjun Bhasin, "Credit Card Dataset for Clustering," Kaggle, 2020. URL: <https://www.kaggle.com/datasets/arjunbhasin2013/ccdata>, pristupano: 23.6.2024.
- [42] Caesar Mario, "Clustering for Effective Marketing Strategy," Kaggle, 2024. URL: <https://www.kaggle.com/code/caesarmario/clustering-for-effective-marketing-strategy>, pristupano: 24.6.2024.
- [43] Jay Shah, "Credit Card Customer Segmentation," LinkedIn, URL: <https://www.linkedin.com/pulse/credit-card-customer-segmentation-jay-shah>, pristupano: 24.6.2024.
- [44] Zakaria Jaadi, "A Step-by-Step Explanation of Principal Component Analysis (PCA)," Built In, 23.2.2024. URL: <https://builtin.com/data-science/step-step-explanation-principal-component-analysis>, pristupano: 25.6.2024.
- [45] Oscar Contreras Carrasco, "Gaussian Mixture Model Explained," Built In, 23.2.2024. URL: <https://builtin.com/articles/gaussian-mixture-model>, pristupano: 25.6.2024.

[46] Sylee0702, "Customer Marketing Analysis," Kaggle, 2024. URL:
<https://www.kaggle.com/code/sylee0702/customer-marketing-analysis>, pristupano:
26.6.2024.

Sažetak

Analiza korisničkih podataka na internetskim platformama primjenom algoritama grupiranja

Ovaj diplomski rad istražuje primjenu algoritama grupiranja za analizu korisničkih podataka na internetskim platformama, s naglaskom na web analitiku i analitiku financijskih platformi. Teorijski dio obuhvaća povijest razvoja ovih dviju analitika, ključne pojmove, najčešće korištene alate i sl. Isto tako, teorijski dio opisuje algoritme grupiranja kao što su K-means, DBSCAN, hijerarhijsko grupiranje i model Gaussovih mješavina. Praktični dio uključuje analizu podataka internetske trgovine i korisnika kreditnih kartica, korištenjem različitih algoritama grupiranja. Rezultati pokazuju da pravilna primjena ovih algoritama može poboljšati segmentaciju korisnika i optimizaciju poslovnih procesa.

Ključne riječi:

Grupiranje, algoritmi grupiranja, nenadzirano učenje, web analitika, analitika internet trgovina, financijska analitika, analitika kreditnih kartica, K-means, DBSCAN, aglomerativno grupiranje, model Gaussovih mješavina, PCA, korisnički podaci, segmentacija korisnika

Summary

User Data Analysis on Internet Platforms Using Clustering Algorithms

This thesis explores the application of clustering algorithms for analyzing user data on internet platforms, with a focus on web analytics and financial platform analytics. The theoretical section covers the history of the development of these analytics, key concepts, and the most commonly used tools. It also describes clustering algorithms such as K-means, DBSCAN, hierarchical clustering, and Gaussian mixture models. The practical section includes the analysis of e-commerce data and credit card users, using various clustering algorithms. The results show that proper application of these algorithms can enhance user segmentation and optimize business processes.

Keywords:

Clustering, clustering algorithms, unsupervised learning, web analytics, e-commerce analytics, financial analytics, credit card analytics, K-means, DBSCAN, agglomerative clustering, Gaussian mixture models, PCA, user data, user segmentation