SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Ivan Vlašić

Marko Kir

Razvoj deskriptivnih i prediktivnih modela na podacima iz poslovanja

PROJEKT

Varaždin, 2024.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Ivan Vlašić

Marko Kir

Studij: Informacijski i poslovni sustavi

Razvoj deskriptivnih i prediktivnih modela na podacima iz poslovanja

PROJEKT

Mentor/Mentorica:

Izv. Doc. dr. sc. Dijana Oreški

Varaždin, travanj 2024.

Ivan Vlašić i Marko Kir

Izjava o izvornosti

Izjavljujemo da je naš projekt izvorni rezultat našeg rada te da se u izradi istoga nismo koristili drugim izvorima osim onima koji su u njemu navedeni. Za izradu projekta su korištene etički prikladne i prihvatljive metode i tehnike rada.

Autor/Autorica potvrdio/potvrdila prihvaćanjem odredbi u sustavu FOI-radovi

Sadržaj

1. Uvod

U današnjem digitalnom dobu, poduzeća se suočavaju s izazovima sve veće količine podataka i potrebom za njihovom analizom radi donošenja informiranih poslovnih odluka. Analiza podataka u poslovanju postaje ključni faktor uspjeha, omogućujući organizacijama da stvore konkurentsku prednost, identificiraju trendove i predvide buduće događaje. Ovaj projekt će istražiti važnost deskriptivnih i prediktivnih modela u analizi podataka iz poslovanja te će pružiti uvid u primjenu CRISP-DM (Cross-Industry Standard Process for Data Mining) standarda kao strukture za vođenje projekata analize podataka.

U prvom dijelu projekta fokusirat ćemo se na razumijevanje domene, što uključuje istraživanje postojeće literature i identifikaciju ključnih istraživanja i trendova u području analize podataka u poslovanju. Nakon toga, detaljno ćemo analizirati skup podataka koji će se koristiti u projektu, istražujući njegovu strukturu, relevantne varijable i potencijalne izazove s podacima.

U drugom dijelu projekta fokusirat ćemo se na upoznavanje s podacima kako bismo sistematski proveli projekt analize podataka iz poslovanja. Kroz različite korake istražit će se kako deskriptivni i prediktivni modeli mogu pružiti vrijedne uvide i predikcije za organizacije u različitim sektorima poslovanja.

Kroz ovaj projekt, cilj nam je pružiti osnovno razumijevanje važnosti analize podataka u poslovanju te praktične uvide i alate za primjenu deskriptivnih i prediktivnih modela u organizacijama.

2. Razumijevanje domene

U današnjem digitalnom dobu poduzeća se suočavaju s izazovima sve većih količina podataka, često opisivanih kao "veliki podaci" (eng. big data), te se prepoznaje potreba za učinkovitom analizom tih podataka radi donošenja informiranih poslovnih odluka. Analiza podataka postaje ključni alat za organizacije koje žele iskoristiti potencijal svojih podataka kako bi stekle konkurentsku prednost, identificirale trendove i predvidjele buduće događaje. Međutim, prije nego što započnemo s analizom podataka, ključno je prvo razumjeti specifičnu domenu ili područje poslovanja u kojem se podaci generiraju i koriste. Tek nakon toga možemo razmotriti kako deskriptivni i prediktivni modeli mogu pružiti ključne uvide i pomoći u donošenju informiranih odluka. Prema autoru Greasley, (2019) modeliranje temeljeno na podacima ima za cilj izvući opis ponašanja sustava na temelju promatranja, opisujući kako se sustav ponaša pod različitim uvjetima ili scenarijima. Ovi modeli, nazvani deskriptivni, fokusiraju se na odnos između ulaza i izlaza. Deskriptivna analitika, koja koristi izvještaje i vizualne prikaze, pomaže u razumijevanju prošle i trenutačne poslovne izvedbe, pružajući statističke sažetke metrika poput prodaje i prihoda te pregled trendova u izvedbi.

Agyapong, Acquah i Asante, (2016) u svojem radu objašnjavaju da deskriptivna analitika obuhvaća proces pretvaranja podataka u korisne informacije za izvještavanje i analizu. Ona omogućuje detaljnu analizu kako bi se odgovorilo na pitanja:

- "što se dogodilo?
- što je trenutno važno?".(Agyapong, Acquah i Asante, (2016), str. 55)

Sumarizacija podataka je proces kompresije skupa, zadržavajući bitne informacije. Sekvencijski podaci su alternativni alat koji se koristi kao deskriptivni model. Deskriptivni modeli prepoznaju obrasce u podacima i otkrivaju karakteristike, ali ne predviđaju uvijek budućnost kao prediktivni modeli. Što nas uvodi u dio povezan s prediktivnim modelima.

Agyapong i kolege, (2016) i Greasley, (2019) u svojim radovima navode da se prediktivni modeli fokusiraju na predviđanje budućih ishoda umjesto trenutnog ponašanja. Njihova svrha je pružiti izlazne podatke koji mogu biti kategoričke ili numeričke vrijednosti. Primjerice, analizom transakcija kreditnim karticama putem prediktivnog modela, možemo procijeniti vjerojatnost prijevare u određenoj transakciji. Jedna od bitnih metoda je regresija, a koristi se kao tehnika nadziranog učenja za

analizu ovisnosti između atributa i razvijanje modela koji može predvidjeti te vrijednosti za nove slučajeve. Ovaj pristup omogućuje planiranje budućnosti putem otkrivanja uzoraka ili odnosa u povijesnim podacima i njihovog projiciranja u budućnost.

Agyapong i kolege, (2016) opisuje različite aspekte prediktivnog modeliranja, uključujući analizu vremenskih serija, klasifikaciju i njihove primjene. Navode da je analiza trendova važna u proučavanju vremenskih serija, poput dnevnih zatvaranja cijena dionica na burzi, koje se mogu prikazati kao funkcija vremena. Glavni ciljevi u analizi vremenskih serija su modeliranje vremenskih serija, kako bi se razumjeli mehanizmi ili temeljne sile koje ih generiraju, te prognoziranje budućih vrijednosti vremenskih serija. Klasifikacija je jedna vrsta prediktivnog modeliranja koja uključuje pridruživanje novih objekata unaprijed definiranim grupama. Primjeri primjena klasifikacija uključuju dijagnozu medicinskih bolesti, ciljani marketing, analizu bioloških podataka, analizu društvenih mreža te filtriranje dokumenata.

Dselen & Ram, (2018) u svojem istraživanju navode prednosti i mane analitike, također navode sljedeće. Poslovna analitika, kao relativno novo područje koje dobiva popularnost u poslovnim i akademskim krugovima, predstavlja umjetnost i znanost otkrivanja uvida korištenjem naprednih matematičkih, statističkih i mrežnih znanstvenih metoda, te strojnog učenja. Ova disciplina obuhvaća i deskriptivne metode, koje opisuju prošle događaje i prediktivne metode, koje predviđaju buduće trendove, omogućujući tvrtkama da donose bolje i brže odluke.

Glavni razlozi popularnosti analitike uključuju potrebu za boljim poslovnim odlukama, dostupnost i ekonomsku isplativost te promjenu kulture prema donošenju odluka temeljenih na podacima.

Nedavni tehnološki napreci omogućili su organizacijama prikupljanje ogromnih količina podataka putem automatiziranih sustava temeljenih na senzorima i RFID tehnologiji, kao i putem internetskih tehnologija poput društvenih mreža. Prednosti analitike u istraživanju uključuju razvoj novih uvida, povećanu relevantnost istraživanja, mogućnost korištenja inovativnih metoda rješavanja problema te sveprisutnost analitike u gotovo svim aspektima poslovanja. Međutim, postoje i izazovi kao što su naglasak na tradicionalnim modelima istraživanja, nedostatak obuke za korištenje analitičkih metoda, poteškoće u opravdavanju ulaganja, te pitanja sigurnosti i privatnosti podataka.

Sveukupno, analitika predstavlja ključni alat za donošenje poslovnih odluka u suvremenom poslovnom okruženju te je važno da istraživači i organizacije prepoznaju njezinu važnost i potencijal.

Zbog sve većeg značaja poslovne analitike za donošenje informiranih poslovnih odluka, posebno u kontekstu sve veće dostupnosti ogromnih količina podataka, važno je također istaknuti nužnost razumijevanja samih podataka koji se koriste u analizi. U sljedećem dijelu rada, fokusirat ćemo se na razumijevanje podataka iz trgovačkog lanca, istražujući njihovu strukturu, kvalitetu te moguće izvore i načine prikupljanja. Ovo razumijevanje će nam omogućiti dublji uvid u specifične karakteristike podataka te bolje pripremiti temelj za daljnju analizu i interpretaciju rezultata. Ciljani atribut koji će se koristiti kod prodiktivnih, a i deskriptivnih modela je TOT_SALES koji bude naknadno prikazan i objašnjen.

Za deskriptivni model je odabran DBSCAN dok za prediktivni model je odabran random forest.

Nadalje u radu će se provesti usporedba. Usporedba se odnosi na rad navedenih modela ovisno o pripremi podataka. Od jednog skupa podataka napravit će se kopija koja neće proći jednaki tretman u pripremi podataka te će se na kraju pokušati usporediti rješenja i pokušat dobit zaključak koji način pripreme podataka je bolji ako postoji bolji način.

3. Razumijevanje podataka

Uzimajući u obzir potrebe istraživanja i analize u kontekstu poslovne analitike, podaci korišteni za naredni dio rada su Quantium retail data, koji je dostupan putem Kaggle (MUKESH, 2022), popularne platforme za dijeljenje i pristupanje skupovima podataka iz različitih područja. Ovi podaci potječu iz trgovačkog lanca i obuhvaćaju širok spektar informacija. Ovi podaci pružaju dubok uvid u dinamiku poslovanja trgovačkog lanca, omogućujući nam analizu trendova, identifikaciju ključnih uzoraka potrošnje, predviđanje budućih potreba i optimizaciju poslovnih procesa.

U nastavku je slika koja prikazuje nekoliko prvih redaka podataka iz skupa. U ovim podacima možemo vidjeti osnovne informacije o proizvodima. Prikazana slika pruža

početni uvid u strukturu i sadržaj podataka koji će biti dalje analizirani i istraživani kako bismo izvukli korisne uvide i informacije.

	LYLTY_CARD_NBR	DATE	STORE_NBR	TXN_ID	PROD_NBR	PROD_NAME	PROD_QTY	TOT_SALES	PACK_SIZE	BRAND	LIFESTAGE	PREMIUM_CUSTOMER
0	1000	2018-10-17	1	1	5	Natural Chip Compny SeaSalt175g	2	6.0	175	NATURAL	YOUNG SINGLES/COUPLES	Premium
1	1002	2018-09-16	1	2	58	Red Rock Deli Chikn&Garlic Aioli 150g	1	2.7	150	RRD	YOUNG SINGLES/COUPLES	Mainstream
2	1003	2019-03-07	1	3	52	Grain Waves Sour Cream&Chives 210G	1	3.6	210	GRNWVES	YOUNG FAMILIES	Budget
3	1003	2019-03-08	1	4	106	Natural ChipCo Hony Soy Chckn175g	1	3.0	175	NATURAL	YOUNG FAMILIES	Budget
4	1004	2018-11-02	1	5	96	WW Original Stacked Chips 160g	1	1.9	160	WOOLWORTHS	OLDER SINGLES/COUPLES	Mainstream
...	...	...	...	...	...	...	...	...	...	...	...	...
264829	2370701	2018-12-08	88	240378	24	Grain Waves Sweet Chilli 210g	2	7.2	210	GRNWVES	YOUNG FAMILIES	Mainstream
264830	2370751	2018-10-01	88	240394	60	Kettle Tortilla ChpsFeta&Garlic 150g	2	9.2	150	KETTLE	YOUNG FAMILIES	Premium
264831	2370961	2018-10-24	88	240480	70	Tyrrells Crisps Lightly Salted 165g	2	8.4	165	TYRRELLS	OLDER FAMILIES	Budget
264832	2370961	2018-10-27	88	240481	65	Old El Paso Salsa Dip Chnky Tom Ht300g	2	10.2	300	OLD	OLDER FAMILIES	Budget
264833	2373711	2018-12-14	88	241815	16	Smiths Crinkle Chips Salt & Vinegar 330g	2	11.4	330	SMITHS	YOUNG SINGLES/COUPLES	Mainstream

264834 rows × 12 columns

Slika 1 Tablica prvih par i zadnjih par podataka

Skup sadrži sljedeće podatke:

Loyalty card number - odnosi se na broj kartice vjernosti
Date - datum prodaje
Store number - id trgovine gdje je prodaja napravljena
Transaction identification - id transakcije koju je kupac napravio kupnjom
Product number - odnosi se na id proizvoda
Product name - naziv proizvoda
Product quantity - odnosi se na količinu proizvoda koja je prodana tijekom kupnje
Total sales - mjera koja predstavlja ukupnu sumu svih prodaja koje je trgovina ostvari la tijekom određenog razdoblja
Pack size - veličina pakiranja, odnosno težina pakiranja proizvoda
Brand - odnosi se na marku proizvoda
Lifestage - dobno razdoblje kupca, navedeno je ima li kupac i obitelj
Premium customer - govori u kojem je programu kupac

U sljedećem dijelu prikazan je tip podataka

LYLTY_CARD_NBR: cijeli broj (int64)

DATE: objekt (datum u tekstualnom formatu, ali Pandas ga tretira kao objekt)

STORE_NBR: cijeli broj (int64)

TXN_ID: cijeli broj (int64)

PROD_NBR: cijeli broj (int64)

PROD_NAME: objekt (tekstualni podaci)

PROD_QTY: cijeli broj (int64)

TOT_SALES: decimalni broj (float64)

PACK_SIZE: cijeli broj (int64)

BRAND: objekt (tekstualni podaci)

LIFESTAGE: objekt (tekstualni podaci)

PREMIUM_CUSTOMER: objekt (tekstualni podaci)

Prikazati statistički sažetak odabranih atributa

	count	mean	std	min	25%	50%	75%	max
PROD_QTY	264834.0	1.905813	0.343436	1.0	2.0	2.0	2.0	5.0
TOT_SALES	264834.0	7.299346	2.527241	1.5	5.4	7.4	9.2	29.5
PACK_SIZE	264834.0	182.425512	64.325148	70.0	150.0	170.0	175.0	380.0

Slika 2 Statistički sažetak odabranih atributa

Izbačen je jedan od atributa, a to je TXN_ID jer je on unikatan za svaku kupnju pa nema smisla gledati statistički prikaz navedenog atributa, te se ostali atributi kao što je LYLTY_CARD_NBR, STORE_NBR, PROD_NBR budu naknadno objasnili.
Iz priloženog se primijetit sljedeće:

- PROD_QTY: prikazuje da je prosječna količina proizvoda po transakciji oko 1.91, s minimalnom količinom od 1 i maksimalnom količinom od 5. Većina transakcija (oko 75%) sadrži 2 proizvoda.
- TOT_SALES: prosječna prodaja po transakciji iznosi oko 7.30, s minimalnom prodajnom vrijednošću od 1.5 i maksimalnom prodajnom vrijednošću od 29.5. Većina transakcija (oko 75%) ima ukupnu prodajnu vrijednost između 5.4 i 9.2.
- PACK_SIZE: Prosječna veličina pakiranja je oko 182. 43, s minimalnom veličinom pakiranja od 70 i maksimalnom veličinom pakiranja od 380. Većina proizvoda (oko 75%) dolazi u pakiranju veličine između 150 i 175.

- LYLTY_CARD_NBR: mod vrijednost za LYLTY_CARD_NBR atribut je 162039. To znači da se karta s brojem 162039 najčešće pojavljuje u datasetu.

- STORE_NBR: mod vrijednost za STORE_NBR atribut je 226. Ova vrijednost označava da je trgovina s brojem 226 najčešće zastupljena u datasetu.
- PROD_NBR: mod vrijednost za PROD_NBR atribut je 102. To sugerira da je proizvod s brojem 102 najčešće kupljen u trgovini.

Analiza statističkih mjera kao što su prosječne vrijednosti, minimumi, maksimumi i modovi pruža korisne uvide u distribuciju podataka i karakteristike skupa podataka.

Na primjer:

Prosječna količina proizvoda po transakciji (PROD_QTY) nam pomaže razumjeti prosječnu veličinu transakcija te uvidjeti preferencije kupaca u količini proizvoda koje kupuju.

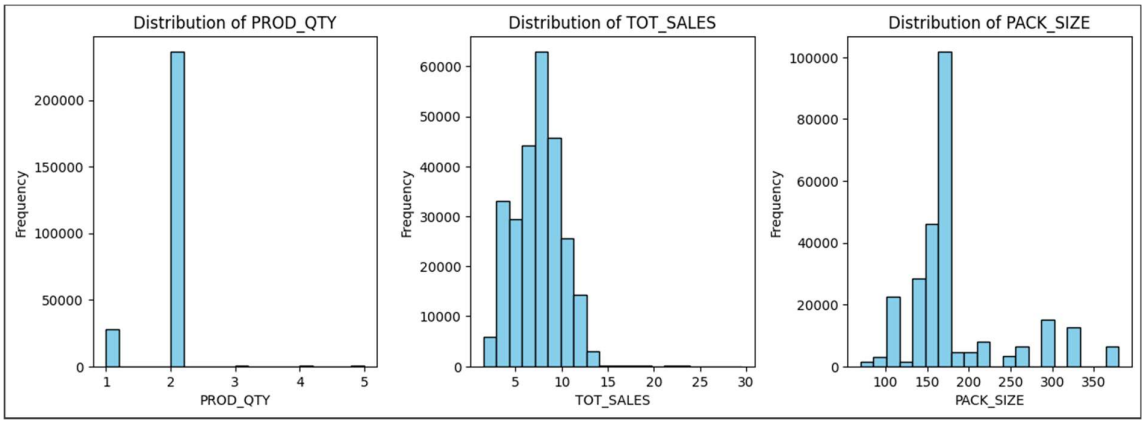
Prosječna ukupna prodajna vrijednost po transakciji (TOT_SALES) omogućuje nam razumijevanje prosječne vrijednosti transakcija i ukupne prodaje.

Prosječna veličina pakiranja (PACK_SIZE) otkriva nam uobičajene veličine pakiranja proizvoda te može biti korisna za planiranje zaliha i upravljanje inventarom.

Modovi za atribute LYLTY_CARD_NBR, STORE_NBR i PROD_NBR naglašavaju najčešće pojavljujuće vrijednosti u dataset-u, što može biti korisno za identifikaciju dominantnih entiteta kao što su najčešći kupci, trgovine ili proizvodi.

Sve ove statistike pomažu nam u razumijevanju ponašanja kupaca, identifikaciji trendova prodaje te donošenju informiranih poslovnih odluka u trgovačkom lancu. Analiza statističkih mjera ključna je za stvaranje strategija prodaje, planiranje zaliha i poboljšanje poslovnih performansi.

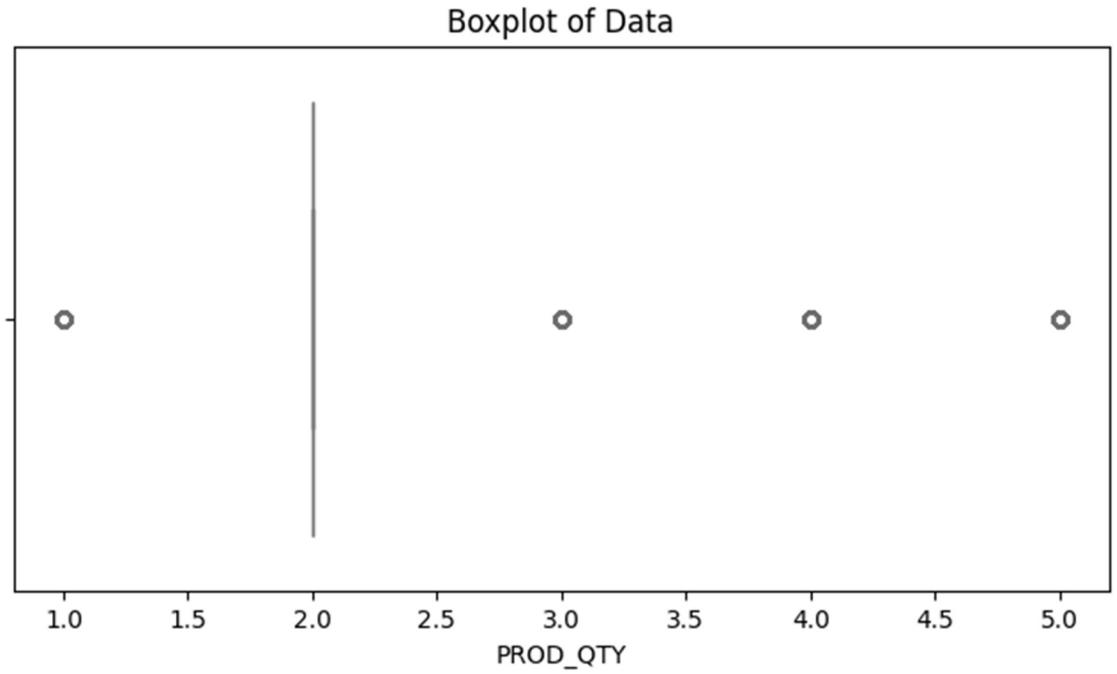
Grafički prikaz statističkih podataka korišten je kako bi se dobio vizualni uvid u ključne karakteristike skupa podataka. Grafikoni će pomoći da se bolje razumije distribucija podataka, identificira trendove i istraži povezanost između različitih varijabli.



Slika 3 Grafički prikaz distribucije vrijednosti

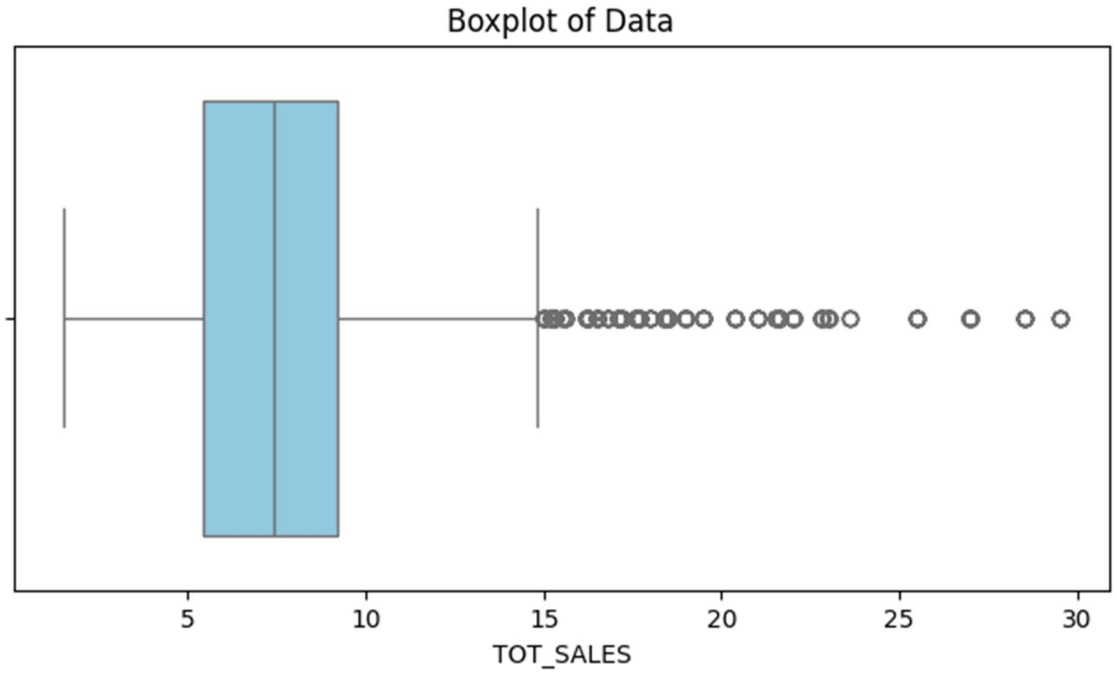
Prikazani histogrami su za sljedeće atribute: product quantity, total sales i pack size.
Za histogram product quantity (PROD_QTY) primjetno je jako velika prisutnost vrijednosti 2 te nešto manja vrijednosti 1 ali minimalno vrijednosti 3, 4 i 5 ali su prisutne.
Histogram za total sales (TOT_SALES) je unimodalni histogram, ukrivljen u desno, najveća distribucija vrijednosti je oko 7 te stršila nisu primjetna.
Histogram za pack size (PACK_SIZE) može se primijetiti kako jedna vrijednost približno 150 ima najveću prisutnost u skupu podataka dok ostale vrijednosti ima znatno manje pojavljuju. Može se primijetiti da je unimodalan i ukrivljen u desno. No iz slike 2. mogu se primijetiti detaljniji rezultati.

U nastavku, koristiti će se box plotove kako bi se vizualno prikazali raspodjelu vrijednosti za odabrane varijable. Box plotovi pružaju pregled osnovnih statističkih informacija, uključujući medijanu, kvartile i prisustvo potencijalnih outlier-a. Analizirat će se box plotovi kako bi se bolje razumjelo varijabilnost i distribuciju podataka te identificiralo eventualne anomalije ili ekstremne vrijednosti



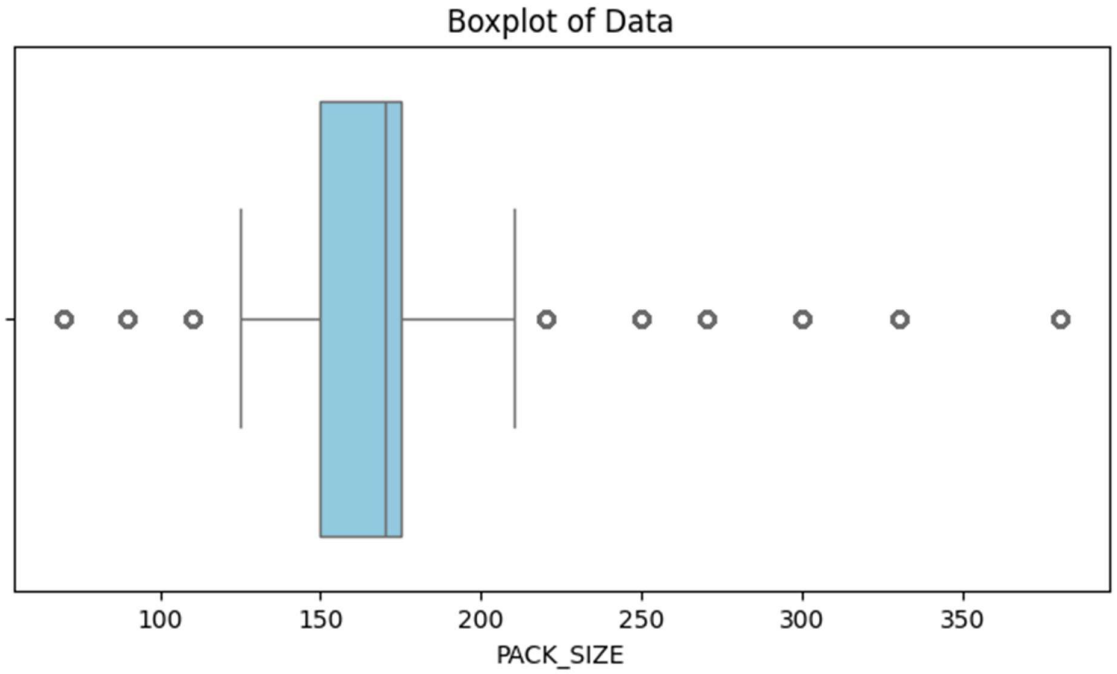
Slika 4 Box plot za PROD_QTY

Analizirajući box plotove, primjećuje se izrazito uski raspon vrijednosti za PROD_QTY, koji se kreće od 1 do 5. Dominantna vrijednost koja se najčešće pojavljuje je 2, što naglašava centralnu tendenciju distribucije. Osim toga, vidljivo je da su vrijednosti 1, 3, 4 i 5 rijetke u usporedbi s brojem 2, te se pojavljuju kao izdvojene točke, što sugerira da su potencijalno ekstremne ili netipične vrijednosti u ovom skupu podataka.



Slika 5 Box plot za TOT_SALES

Analizom atributa Ukupna prodaja (TOT_SALES) primjećuje se prisutnost izdvojenih točaka samo na desnoj strani box plotova, što ukazuje na prisutnost ekstremnih vrijednosti većih od 14.9.



Slika 6 Boxplot za PACK_SIZE

Na posljednjem box plotu, koji prikazuje atribut Veličina pakiranja (PACK_SIZE), primjećuje se prisutnost izdvojenih točaka kako s lijeve tako i s desne strane. Detaljnijim promatranjem, uočava se da je veći broj izdvojenih točaka s desne strane grafikona, što ukazuje na prisutnost ekstremnih vrijednosti većih od 212 i manjih od 120, otprilike.

Scatterbox matrica je koristan alat u analizi podataka jer omogućuje istovremeni prikaz svih kombinacija numeričkih atributa u datasetu. Ova matrica prikazuje scatter plotove između svakog para atributa, što omogućuje vizualnu analizu odnosa između različitih atributa. Kroz scatter plot-ove možemo identificirati obrasce, raspodjele podataka i potencijalne outlier-e. Prikazivanje podataka na ovaj način olakšava prepoznavanje eventualnih korelacija između atributa i pruža uvid u njihov međusobni utjecaj.

Slika 7 Scatterplot matrica

Primjetna je blaga korelacija između PROD QTY i TOT SALES koja je pozitivna.

Sada, nakon što se temeljito istražilo i analiziralo podatke, može se krenuti u daljnju obradu. Upoznavanje s podacima, koje je provedeno kroz različite analitičke tehnike i vizualizacije, ključno je za razumijevanje njihovih karakteristika, distribucije i međusobnih odnosa. Ovaj proces pripreme podataka omogućio da se identificira potencijalne obrasce, trendove i korelacije, te se stekao dublji uvid u naš skup podataka. Sada se mogu koristiti ove spoznaje kako bi se dalje razvijalo modele, provodilo analize i donosilo informirane odluke u poslovnom ili istraživačkom kontekstu.

4. Priprema podataka

Priprema podataka je ključna faza u analizi podataka koja omogućava stvaranje temelja za uspješno modeliranje i interpretaciju rezultata. U ovom koraku, podaci se podvrgavaju različitim postupcima kako bi se uklonili nedostaci, smanjila redundancija, identificirale anomalije i pripremili za primjenu različitih analitičkih tehnika. U konačnici, priprema podataka igra ključnu ulogu u osiguravanju uspješne analize podataka i donošenju pouzdanih zaključaka i odluka temeljenih na njima. Bez temeljite pripreme, analitički rezultati mogu biti nepouzdani i netočni, što može dovesti do pogrešnih zaključaka i odluka. Stoga je važno posvetiti dovoljno pažnje ovom koraku kako bi se osigurala kvalitetna i pouzdana analiza podataka. U našem radu kao što je navedeno imamo dva skupa df1 (koji kasnije postaje u kodu cleaned_df) i df koji neće proći

jednaki tretman u pripremi podataka da bi dobili zaključak koji način je prihvatljivi za navedene modele.

4.1. Oblikovanje podataka (prvi dio)

U procesu oblikovanja podataka, pristupili smo pripremi podataka za daljnju analizu i modeliranje. Kodiranje kategoričkih vrijednosti u numeričke je ključni korak u pripremi podataka za analizu ili primjenu algoritama strojnog učenja. Kategoričke varijable su one koje predstavljaju različite kategorije ili skupine, poput boja, tipova proizvoda ili naziva gradova.
Kodiranje kategoričkih varijabli u numeričke je važno jer većina algoritama strojnog učenja zahtijeva numeričke podatke za obuku modela.

U našem riješenu (kodiranje je odrađeno na oba skupa i df1 i df) primijenili smo dva različita postupka kodiranja kategoričkih značajki kako bismo ih pretvorili u numeričke vrijednosti. Prvo smo koristili metodu frekvencijskog kodiranja (frequency encoding) na kategoričkim značajkama 'LIFESTAGE', 'BRAND' i 'PROD_NAME'. Ova metoda zamjenjuje svaku kategoričku vrijednost njezinom relativnom frekvencijom pojavljivanja u skupu podataka, što pomaže u očuvanju informacija sadržanih u tim značajkama. Metoda omogućava zadržavanje korisnih informacija o distribuciji podataka, što je korisno za oba modela. Autor Neural Ninja, (2023) ističe da je ovaj način pretvaranja kategorisjkih atributa u numeričke dobar za prediktivni model koji bude korišten kasnije u radu, a to je random forest. Nakon toga, koristili smo ordinalno kodiranje (OrdinalEncoder) za kategoričku značajku 'PREMIUM_CUSTOMER'. Ordinalno kodiranje dodjeljuje numeričke vrijednosti kategorijama prema njihovom redoslijedu definiranom u listi 'redoslijed', koja u ovom slučaju predstavlja različite razine premium korisnika, od 'Budget' do 'Premium'. Time smo transformirali kategoričke vrijednosti u numeričke. Budget je predstavljen s 0, Mainstream s 1 i Premium s 2. Korištenje ordinalnog encodera za pretvaranje 'PREMIUM_CUSTOMER' u numeričke vrijednosti omogućava da model prepozna važnost premium kupaca bez gubitka informacija. Ovo je važno za prediktivni model Random Forest, koji koristi ove numeričke vrijednosti za bolje razumijevanje odnosa između značajki i ciljne varijable.

4.2. Čišćenje podataka

Analizirajući naš skup podataka, primijetili smo da nema nedostajućih vrijednosti, što je vidljivo ispod.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 264834 entries, 0 to 264833
Data columns (total 11 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   LYLTY_CARD_NBR   264834 non-null  int64
 1   STORE_NBR        264834 non-null  int64
 2   TXN_ID           264834 non-null  int64
 3   PROD_NBR         264834 non-null  int64
 4   PROD_NAME        264834 non-null  float64
 5   PROD_QTY         264834 non-null  int64
 6   TOT_SALES        264834 non-null  float64
 7   PACK_SIZE        264834 non-null  int64
 8   BRAND            264834 non-null  float64
 9   LIFESTAGE        264834 non-null  float64
 10  Premium_encoded  264834 non-null  float64
dtypes: float64(5), int64(6)
memory usage: 22.2 MB
```

Primjetno je nedostatak atributa DATE kojeg smo maknuli jer smatramo da nije koristan atribut, te bi mogao stvarati problem kod interpretacije te značajke.

Međutim, detaljnijom analizom naših podataka primijetili smo prisutnost outlier-a, što je jasno prikazano na slikama 4, 5 i 6. Odlučili smo provesti proces čišćenja podataka kako bismo uklonili ove odstupajuće vrijednosti i osigurali točnost naših analiza. Provjerili smo prisutnost outlier-a u našem skupu podataka koristeći Tukeyjevu IQR metodu.

Analizirat ćemo sve značajke. Rezultati su pokazali da postoji značajan broj outlier-a u našem skupu podataka.

- LYLTY_CARD_NBR: 44 outliera
- STORE_NBR: Nema outliera
- TXN_ID: 1 outlier
- PROD_NBR: Nema outliera
- PROD_QTY: 28,795 outliera
- TOT_SALES: 576 outliera
- PACK_SIZE: 72,042 outliera

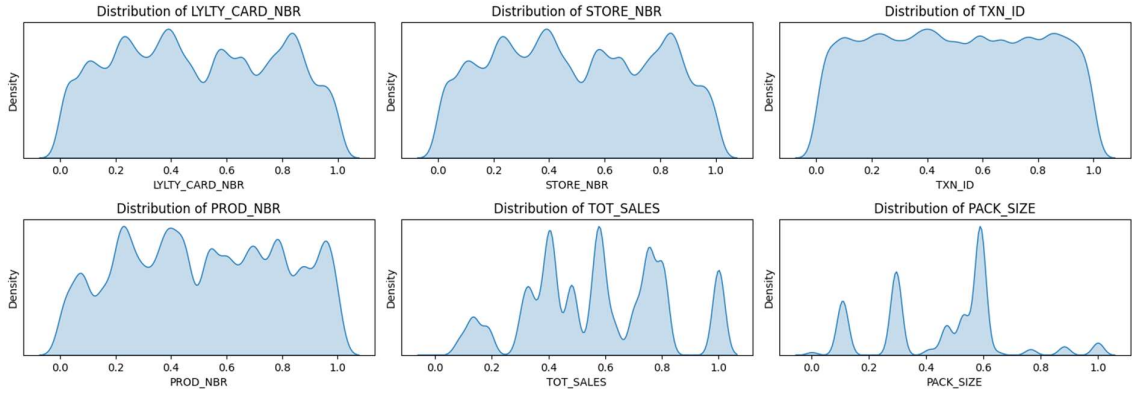
Nakon što smo identificirali ove outlier-e, provedeno je čišćenje podataka samo na skupu df1, dok df nije prošao kroz navedeni tretman, te su oni uklonjeni iz skupa podataka. Konačni rezultat nakon uklanjanja outlier-a pokazuje da smo uspješno

očistili naš skup podataka od ovih odstupanja, te smo spremni nastaviti s daljnjom analizom.

Dodatno PROD_QTY je maknuo sve outlier-e i nakon toga je ostavio samo jednu jedinstvenu vrijednost, a to je 2. Pošto cijeli atribut ima vrijednost 2 više nije vrijedan atribut, jer nam ne pruža nikakve informacije, te ćemo ga ukloniti.

4.3. Oblikovanje podataka (drugi dio)

Nakon kodiranja smo primijenili skaliranje na našim podacima (samo skupu df1) kako bismo osigurali da su vrijednosti svih značajki u istom rasponu. Ovo skaliranje pomaže u poboljšanju stabilnosti i performansi modela. Rezultate skaliranja možete vidjeti na slici 8, gdje su prikazani grafikoni koji ilustriraju utjecaj skaliranja na distribuciju vrijednosti značajki.



Slika 8 Prikaz skaliranih značajki podataka

4.4. Odabir podataka

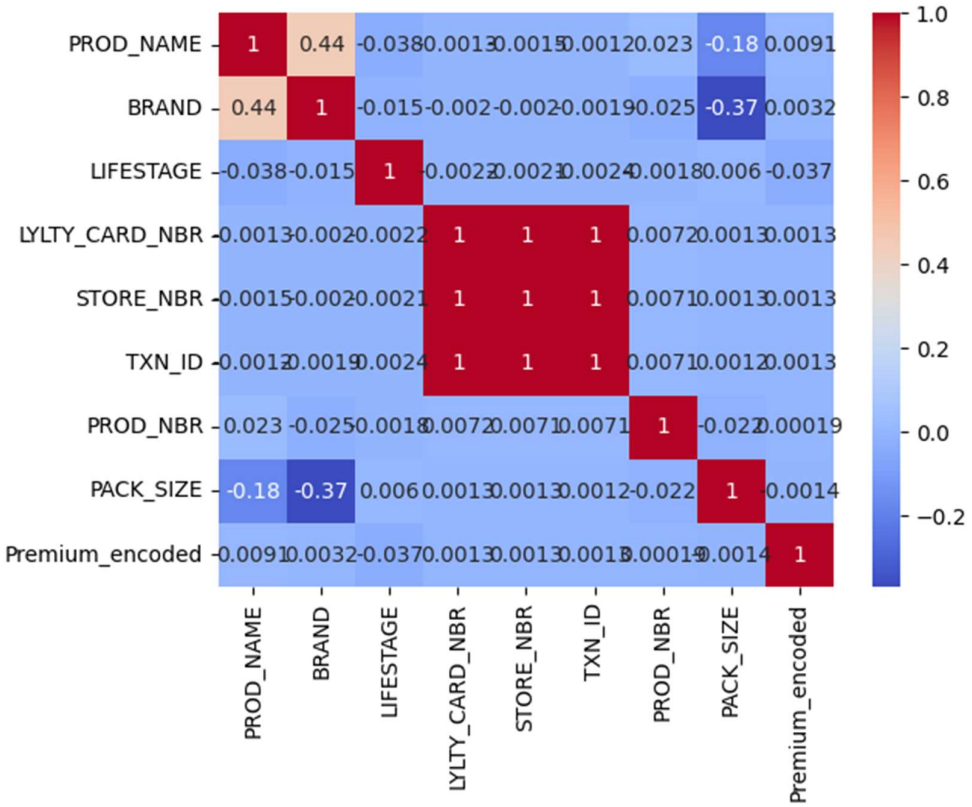
U analizi se fokusiralo na odabir podataka kroz rješavanje problema multikolinearnosti. Korištenjem metode Variance Inflation Factor (VIF), identificirali smo i eliminirali visoko korelirane varijable kako bi smo osigurali stabilnost i pouzdanost naših modela. Rezultate ove analize vizualno smo prikazali putem Heatmap-a, koji nam je pružio jasnu sliku o stupnju korelacije između različitih varijabli. Ovaj pristup omogućio nam je da efikasno upravljamo multikolinearnošću i osiguramo

validnost naših analitičkih rezultata. Sve je odrađeno samo na skupu df1. Također, kako je varijabla TOT_SALES naša ciljna varijabla koju ćemo predviđati, nju ćemo isključiti iz provjere.

Rezultati VIF analize numeričkih značajki dobili smo sljedeće:

	variables	VIF
0	PROD_NAME	10.794370
1	BRAND	5.451379
2	LIFESTAGE	10.529893
3	LYLTY_CARD_NBR	37049.869836
4	STORE_NBR	31284.359120
5	TXN_ID	5358.307871
6	PROD_NBR	4.038398
7	PACK_SIZE	4.888844
8	Premium_encoded	2.269566

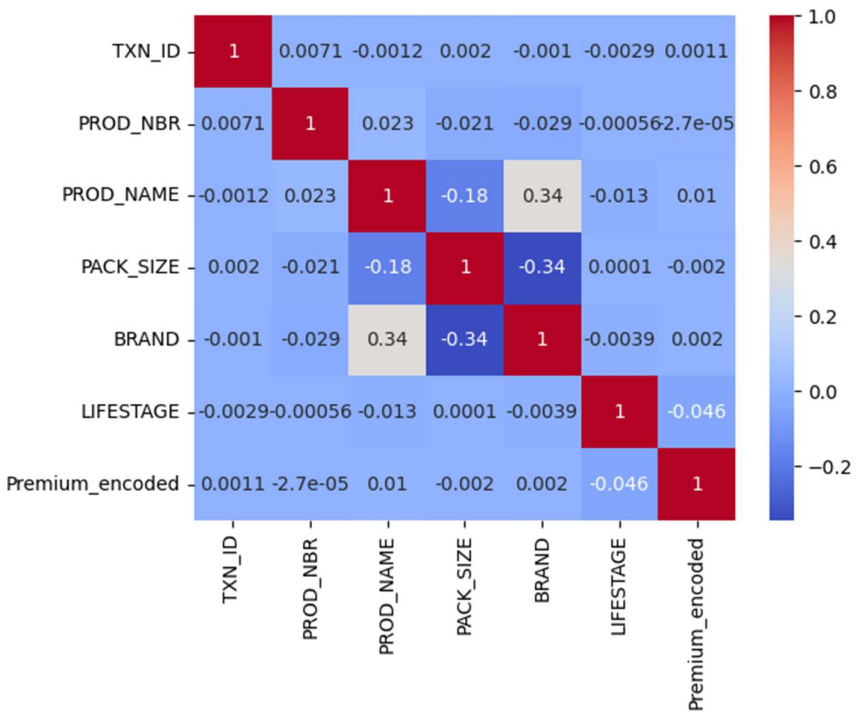
Na osnovu korelacijske matrice i faktora inflacije varijance (VIF), možemo provjeriti postojanje multikolinearnosti između numeričkih značajki u našem skupu podataka. Gledajući korelacijsku matricu, možemo primijetiti visoke vrijednosti korelacije između nekih značajki.



Slika 9 Heatmap prije micanja varijabli

Nadalje, rezultati faktora inflacije varijance (VIF) također potvrđuju postojanje multikolinearnosti. VIF vrijednosti veće od 10 obično se smatraju znakovima multikolinearnosti. Odlučilo se izbaciti varijable s VIF-om većim od 10, te smo dobili sljedeći rezultat:

```
Final VIF values:
          variables       VIF
0             BRAND  3.732986
1         LIFESTAGE  9.549756
2            TXN_ID  3.734325
3          PROD_NBR  3.921975
4         PACK_SIZE  4.684363
5  Premium_encoded  2.249696
```



Slika 10 Heatmap poslije micanja varijabli

Na osnovu heatmapa korelacijske matrice, primjećujemo da postoji slaba pozitivna korelacija između varijabli. Nakon provedenog dobili smo skup cleaned_df (df1 koji je promjenio naziv nakon provjere multikolinearnosti).

4.5. Konstrukcija podataka

U našem slučaju, nakon detaljne analize skupa podataka, utvrdili smo da postojeći atributi pružaju dovoljno informacija i nisu potrebne dodatne konstrukcije ili varijacije.

Stoga smo odlučili zadržati trenutnu strukturu podataka kako bismo izbjegli nepotrebnu složenost i održali jasnoću i jednostavnost u interpretaciji naših rezultata.

4.6. Integriranje podataka

Integriranje podataka obično uključuje spajanje više izvora podataka u jedan integritetan skup podataka radi bolje analize ili modeliranja. Međutim, u našem istraživanju nismo imali potrebu za integriranjem dodatnih izvora podataka jer smo koristili jedan dobro strukturiran i opsežan skup podataka koji pokriva sve potrebne informacije za našu analizu. Stoga smo se fokusirali na temeljitu obradu i analizu dostupnih podataka kako bismo izvukli relevantne uvide i donijeli pouzdane zaključke.

5. Modeliranje

Ulazimo u fazu modeliranja koja predstavlja ključan korak u našoj analizi podataka. Ova faza obuhvaća nekoliko važnih koraka koji će nam pomoći u stvaranju modela koji će precizno predviđati ciljni rezultat. Prvo, odabrat ćemo odgovarajuću tehniku modeliranja koja će biti u skladu s karakteristikama našeg problema i ciljevima istraživanja. Nakon toga, detaljno ćemo planirati strategiju testiranja modela, uz korištenje križne validacije ili drugih metoda particioniranja podataka kako bismo osigurali pouzdanost i generaliziranost rezultata.

Zatim ćemo implementirati odabrane modele i provesti njihovu izgradnju, koristeći optimalne tehnike i algoritme kako bismo postigli željene rezultate. Kroz proces procjene modela, fokusirat ćemo se na optimizaciju hiperparametara i detaljnu evaluaciju performansi modela korištenjem različitih metrika kako bismo osigurali da naš model pruža najbolje moguće rezultate.

5.1. Odabir tehnike modeliranja

Prema autoru Khan i kolegama, (2014) DBSCAN je algoritam klasteriranja bazirani na gustoći, koristan za klasteriranje proizvoljno oblikovanih i velikih skupova podataka.

Rad navedenih autora također istražuje različite poboljšane verzije DBSCAN algoritma (VDBSCAN, FDBSCAN, GRIDBSCAN, IDBSCAN, EDBSCAN), analizirajući njihove prednosti i nedostatke.

Gustoćno-bazirani algoritmi klasteriranja, poput DBSCAN-a, omogućuju učinkovitu analizu i klasteriranje velikih i proizvoljno oblikovanih skupova podataka. Daljnja poboljšanja ovih algoritama mogu dodatno unaprijediti njihovu učinkovitost i primjenjivost, te baš zato će se koristiti na našim podacima.

Suhanda i kolege (2022) navode da metoda random forest stvara niz stabala iz uzoraka podataka, pri čemu stvaranje svakog stabla tijekom treniranja nije povezano s prethodnim stablom. Odluka se temelji na većini glasova. Ključni koncepti ovog algoritma su izgradnja skupa stabala pomoću bagging metode s ponovljenim uzorkovanjem i nasumičan odabir značajki za svako stablo.

Klasifikacija temeljena na skupu postiže maksimalne performanse uz nisku korelaciju između osnovnih učenika. Skup mora koristiti slabog osnovnog učenika kako bi se izbjegao overfitting. Random forest smanjuje korelacije i zadržava snagu klasifikacije nasumičnim odabirom značajki tijekom treniranja svakog stabla, čime se dobiva optimalno stablo grananja.

Dodatni razlog korištenja navedenog prediktivnog modela je taj što random forest može raditi s velikim brojem značajki, jer odabire podskup značajki za svaki stablo odluke. To je korisno kada imamo puno podataka i nismo sigurni koje značajke su najvažnije.

5.2. Dizajn testiranja

Kao što je već navedeno testiranje će se provesti tako što imamo dva skupa podataka koja su prošli različit tretman pripreme podataka. Ta dva skupa su finalno cleaned_df i df. Oba skupa su prošla jednaku pretvorbu kategorijskih atributa u numeričke, dok je cleaned_df prošao još kroz izbacivanje outlier-a, skaliranje značajki da budu u jednakom rasponu i micanje multikolinearnosti. Za klasteriranje nije potrebna križna validacija zbog prirode zadatka, ali je za procjenu kvalitete klastera korišten Silhouette koeficijent. Kod prediktivnog modeliranja, podaci su podijeljeni na trening (80%) i test (20%) skup kako bi se osigurala objektivna procjena performansi modela.

5.3. Izgradnja modela

Klasteriranje:

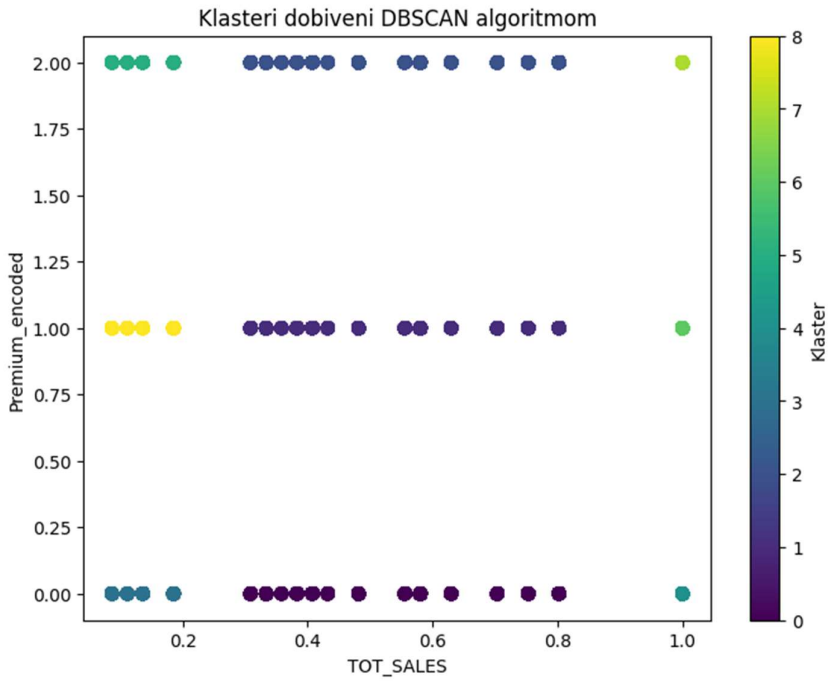
- Podaci su filtrirani i uzeti je uzorak od 10,000 redaka ako je broj podataka veći. Napravljeno je radi činjenice da se bilježnica sruši zbog nedostatka RAM-a. Isprobano je par puta pokrenuti kod i svaki put graf ispadne jednak. Zaključak da zbog karakteristika vrijednosti činjenica da uzimamo samo 10,000 redaka ne mijenja rezultate.
- Izvršeno je klasteriranje korištenjem DBSCAN algoritma na značajkama TOT_SALES s Premium_encoded, BRAND i PACK_SIZE (svaka zasebno s TOT_SALES).
- Za vizualizaciju klastera korištena je scatter plot metoda, a kvaliteta klastera procijenjena je Silhouette koeficijentom.

Prediktivno modeliranje:

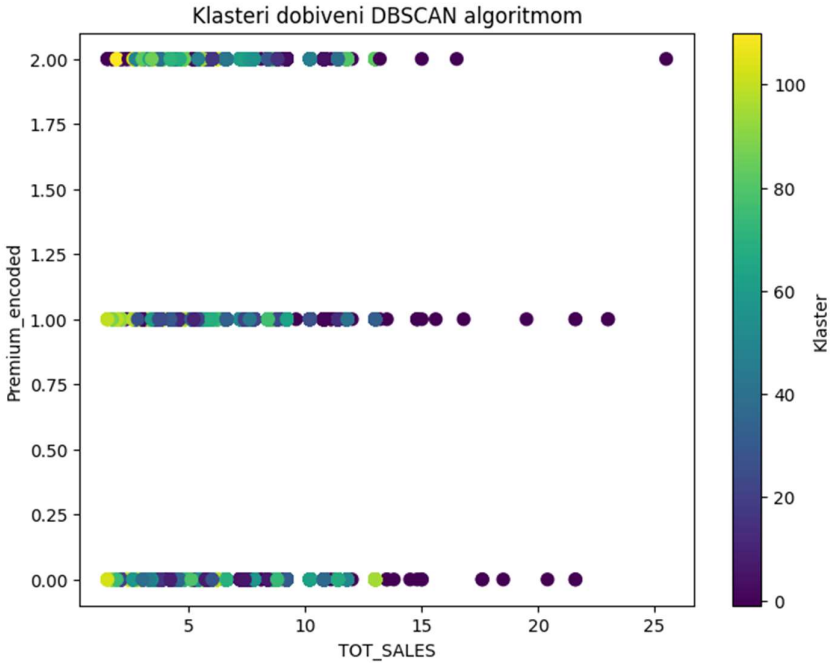
- Značajke Premium_encoded, BRAND i PROD_NBR (opet svaka zasebno s ciljanom varijablom) i ciljne varijable TOT_SALES su odabrane.
- Također imamo i kombinaciju Premium_encoded, BRAND i PACK_SIZE koje se zajedno koriste za predviđanje varijable TOT_SALES
- Podaci su podijeljeni na trening i test setove.
- Random Forest regressor je inicijaliziran sa 100 stabala i treniran na trening podacima.
- Predikcije su napravljene na test setu, a performanse modela procijenjene su pomoću Mean Squared Error (MSE) i R-squared (R^2) metrike.

5.4. Procjena modela

U nastavku će biti prikazan prvo DBSCAN deskriptivni model na podacima cleaned_df, zatim na podacima df.



Slika 11 Prikaz DBSCAN nad cleaned_df 1

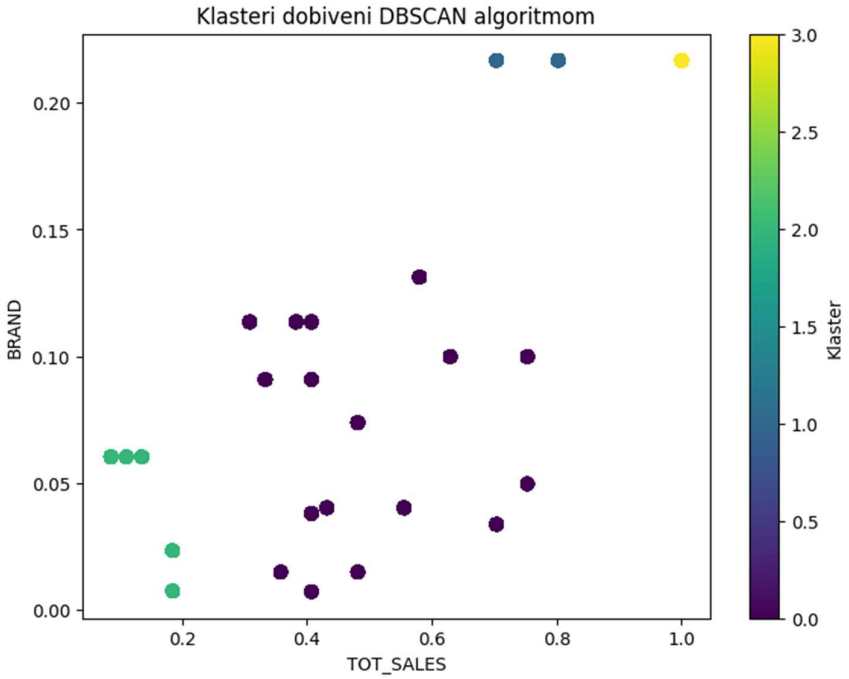


Slika 12 Prikaz DBSCAN na df 1

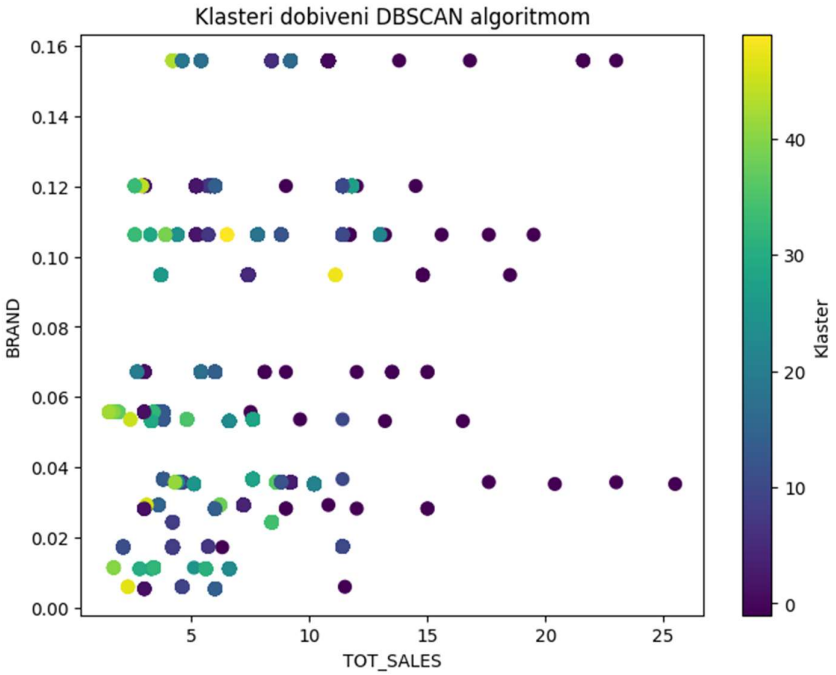
Usporedba rezultata:

Slika 11. Silhouette koeficijent: 0.4086998235307069

Slika 12. Silhouette koeficijent: 0.9345706590946614



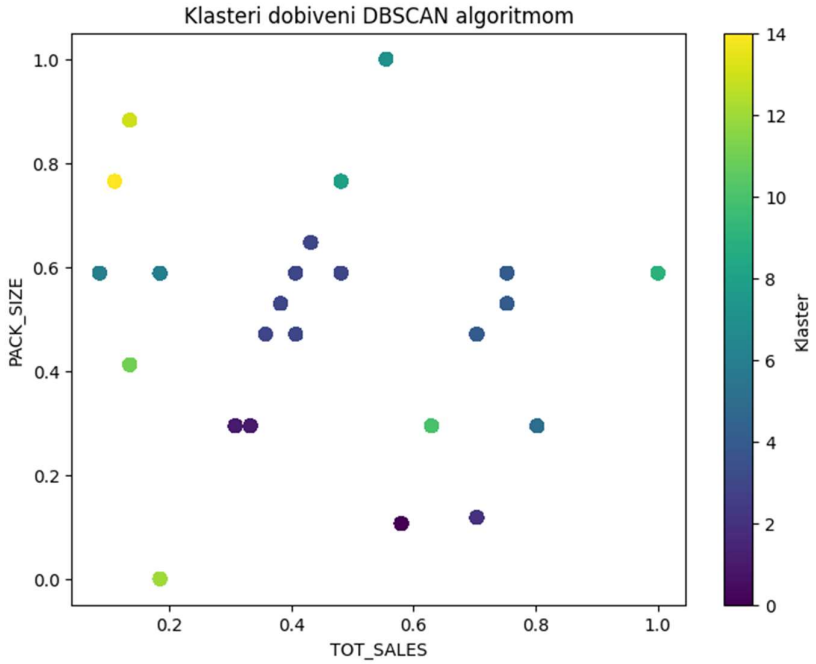
Slika 13 Prikaz DBSCAN na cleaned_df 2



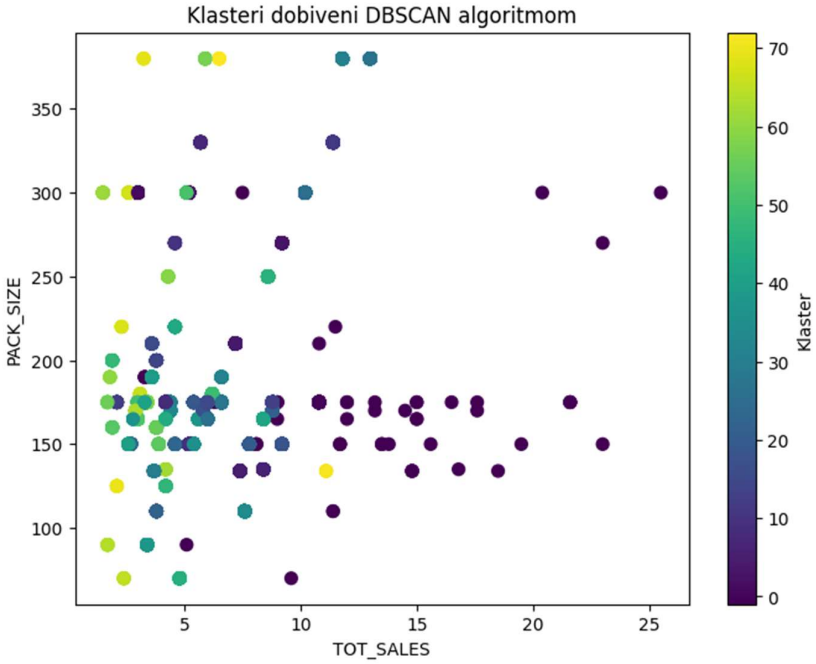
Slika 14 Prikaz DBSCAN na df 2

Rezultati slike 13. Silhouette koeficijent: 0.424155369244345

Rezultati slike 14. Silhouette koeficijent: 0.920124847706906



Slika 15 Prikaz DBSCAN na cleaned_df 3



Slika 16 Prikaz DBSACAN na df 3

Rezultati slike 15. Silhouette koeficijent: 0.8576944094796594

Rezultati slike 16. Silhouette koeficijent: 0.9896529089340913

Sada će se prikazat rezultati random forest prediktivnog modela, a u vrednovanju će se procijeniti rezultati.

Za model koji predviđa TOT_SALES na osnovu BRAND, Premium_encoded i PACK_SIZE nad cleaned_df:

```
Mean Squared Error (MSE): 1.0924697099832263e-05
R-squared (R2): 0.999808467056538
```

Za model koji predviđa TOT_SALES na osnovu BRAND, Premium_encoded i PACK_SIZE nad df:

```
Mean Squared Error (MSE): 1.8647522520128283
R-squared (R2): 0.7097334847321571
```

Za model koji predviđa TOT_SALES na osnovu Premium_encoded nad cleaned_df:

```
Mean Squared Error (MSE): 0.05698440893584125
R-squared (R2): 0.000943323628879722
```

Za model koji predviđa TOT_SALES na osnovu Premium_encoded nad df:

```
Mean Squared Error (MSE): 6.421768218610496
R-squared (R2): 0.0003903839441357082
```

Za model koji predviđa TOT_SALES na osnovu BRAND nad cleaned_df:

```
Mean Squared Error (MSE): 0.003445036547185574
R-squared (R2): 0.9396012553770038
```

Za model koji predviđa TOT_SALES na osnovu BRAND nad df:

```
Mean Squared Error (MSE): 3.440750866283625
R-squared (R2): 0.46441423369623647
```

Za model koji predviđa TOT_SALES na osnovu PROD_NBR nad cleaned_df:

```
Mean Squared Error (MSE): 1.092912325775761e-05
R-squared (R2): 0.9998083894566698
```

Za model koji predviđa TOT_SALES na osnovu PROD_NBR nad df:

```
Mean Squared Error (MSE): 1.8650817220375944
R-squared (R2): 0.7096821995730317
```

6. Vrednovanje

6.1. Evaluacija rezultata

Kako bi evaluirali rezultate, prvo trebamo objasniti kako ćemo ih točno evaluirati i što znače rezultati. Prvo, za DBSCAN evaluaciju koristimo Silhouette koeficijent, koji se

računa, te kao rezultat dolazi prosječna vrijednost svih Silhouette vrijednosti, te ako je ona:

- manja od 0.25 – ne postoji dokaz realnosti klastera
- 0.25-0.5 – postoji određeni dokaz realnog prikaza podataka
- 0.5 ili više – postoji dobar dokaz da klasteri realno prikazuju podatke

Nadalje, za prediktivni model koristi se MSE, ili Mean Square Error (srednja kvadratna greška), je mjera koja se koristi za procjenu kvalitete modela predviđanja. Konkretno, MSE mjeri prosječnu kvadratnu razliku između stvarnih vrijednosti i vrijednosti koje predviđa model. Uz MSE, također koristimo i R^2 (R-Squared) ili koeficijent determinacije, a to je statistička mjera koja pokazuje koliko dobro regresijski model objašnjava varijabilnost zavisne varijable. Drugim riječima, R^2 mjeri proporciju varijance u zavisnoj varijabli koja je objašnjena nezavisnim varijablama u modelu. Za R^2 možemo reći slijedeće:

- 1.0 - Perfektno paše(sumnjivo, model je vjerojatno pretreniran)
- ~ 0.9 – jako dobro
- < 0.7 Nije dobro
- < 0.4 Užasno
- < 0 – Model nema smisla za ove podatke

Sada kad smo objasnili mjere, možemo na evaluaciju. Za deskriptivni modele može se reći slijedeće:

Za prvi DBSCAN gdje se koriste varijable Premium_encoded i TOT_SALES, na skupu cleaned_df vidimo da ima koeficijent oko 0.41, što prema našem objašnjenju ukazuje na to da postoji određeni dokaz realnog prikaza podataka, no to ne znači da je i dobar, dok na skupu df vidimo vrijednost oko 0.93 što ukazuje da postoji dobar dokaz da klasteri realno prikazuju podatke.

Za drugi DBSCAN gdje se koriste varijable BRAND i TOT_SALES, na skupu cleaned_df vidimo da ima koeficijent oko 0.42, što prema našem objašnjenju ukazuje na to da postoji određeni dokaz realnog prikaza podataka, no to ne znači da je i

dobar, dok na skupu df vidimo vrijednost oko 0.92 što ukazuje da postoji dobar dokaz da klasteri realno prikazuju podatke.

Za treći DBSCAN gdje se koriste varijable PACK_SIZE i TOT_SALES, na skupu cleaned_df vidimo da ima koeficijent oko 0.86, dok na skupu df vidimo vrijednost oko 0.98 što kod oba slučaja ukazuje da postoji dobar dokaz da klasteri realno prikazuju podatke, te je to najbolji model koji smo dobili.

Nadalje, za prediktivne modele:

Za prvi Random Forest, koji koristi skup cleaned_df i varijable BRAND, Premium_encoded i PACK_SIZE za predviđanje, dobili smo vrijednost R2 oko 0.99, vrlo blizu 1.0, što ukazuje na to da bi model mogao biti pretreniran, dok smo za skup df dobili oko 0.71 što ukazuje na to da je model dobro treniran. Može se reći da nije loš, ali nije ni pretjerano dobar.

Za drugi Random Forest, koji koristi varijablu Premium_encoded za predviđanje, na oba skupa podataka dobila se vrijednost vrlo blizu 0, što ukazuje da je model loše treniran, pa skoro i da nema smisla za te podatke.

Za treći Random Forest, koji koristi skup cleaned_df i varijablu BRAND za predviđanje, dobili smo vrijednost R2 oko 0.94, kazuje na to da model vrlo dobro objašnjava varijabilnost na podacima, ali je blizu pretreniranosti, dok smo za skup df dobili oko 0.46 što ukazuje na to da je model loše treniran.

Te na kraju, Random Forest, koji koristi skup cleaned_df i varijablu PROD_NBR za predviđanje, dobili smo vrijednost R2 oko 0.99, vrlo blizu 1.0, što ukazuje na to da bi model mogao biti pretreniran, dok smo za skup df dobili oko 0.71 što ukazuje na to da je model dobro treniran. Može se reći da nije loš.

Ovdje bi rekli da je najbolji model 3. Random Forest na skupu cleaned_df, po našim kriterijima koje smo naveli ranije, no vrlo je blizu pretreniranosti, te je mala granica između to dvoje. Modeli 2 i 4 na skupu df sa R^2 oko 0.71 ukazuju na dobru sposobnost generalizacije, ali nisu savršeni.

MSE vrijednosti se mogu vidjeti ranije u tekstu, ali je smatrano da se nisu trebale previše objašnjavati.

Nakon pregleda za DBSCAN podaci koji nisu prošli obradu kod pripreme podataka osim pretvaranja kategorisjkih u numeričke (df) prikazuju bolje rezultate.
Kod random forest modeli koji su trenirani na podacima cleaned_df dobili su dobre rezultate ali ukazuju u nekim slučajevima mogućnost pretreniranosti, dok kod podataka koji nisu prolazili sve navedene promjene u pripremi podatak (df) donose nešto gore rezultate ali izbjegavaju pretreniranost modela.

6.2. Ocjenjivanje procesa

Skup podataka koji smo koristili u analizi pokazao se kao prikladan za primijenjene metode i tehnike, uključujući DBSCAN i Random Forest algoritme. Rezultati evaluacije ukazuju da su pretpostavke ovih algoritama zadovoljene. Na primjer, analiza Silhouette koeficijenta za DBSCAN pokazuje da su klasteri realno prikazani u skupu podataka. Također, visoke vrijednosti R^2 za Random Forest modele sugeriraju da su modeli dobro objasnili varijabilnost u podacima. Stoga, na temelju rezultata analize, možemo zaključiti da su korišteni algoritmi adekvatno primijenjeni na prikladan skup podataka.

6.3. Određivanje slijedećih koraka

U budućnosti, moglo bi se napraviti dodatnu analizu klastera kako bi se dublje razumjele karakteristike svake grupe kupaca te personalizirale marketinške strategije. Nadalje, istraživanje dodatnih značajki može doprinijeti boljem razumijevanju kupovnog ponašanja, dok korištenje naprednijih modela strojnog učenja ili tehnika dubokog učenja može poboljšati prediktivnu točnost i razumijevanje podataka. Praćenje i evaluacija performansi implementiranih modela ključno je za otkrivanje potencijalnih promjena u ponašanju kupaca i prilagodbu analitičkih procesa. Suradnja s drugim odjelima, poput marketinga, prodaje i upravljanja zalihama, može pružiti holistički uvid u potrebe kupaca i identificirati mogućnosti za unaprjeđenje poslovanja. Kontinuirano obrazovanje tima važno je kako bi se osiguralo razumijevanje najnovijih tehnologija i metoda analize podataka te podržalo kontinuirano učenje i inovacije u

organizaciji. Implementacija ovih prijedloga može doprinijeti optimizaciji analitičkih procesa, poboljšanju poslovnih rezultata i održavanju konkurentnosti na tržištu.

7. Korištenje

Evo nekoliko načina kako bi rezultate ovog projekta mogli praktično primijeniti:

1) Marketinška strategija: Na temelju analize klastera koje je proveo DBSCAN algoritam, marketinški tim može bolje razumjeti preferencije kupaca i segmentirati ih u ciljane grupe. Na primjer, mogu se identificirati segmenti kupaca koji preferiraju određene brendove ili veličine pakiranja proizvoda, što omogućuje ciljano usmjeravanje marketinških kampanja prema njima.

2) Optimizacija asortimana proizvoda: Analiza Random Forest modela može pomoći u identificiranju ključnih čimbenika koji utječu na prodaju proizvoda. Na temelju ovih saznanja, tim za upravljanje zalihama može prilagoditi asortiman proizvoda u trgovinama kako bi bolje odgovarao potrebama kupaca i poboljšao prodaju.

3) Prediktivno održavanje u lancu opskrbe: Ako su identificirani relevantni faktori koji utječu na prodaju proizvoda, mogu se koristiti Random Forest modeli za predikciju budućih prodajnih trendova. Na temelju ovih predikcija, lanci opskrbe mogu optimizirati planiranje proizvodnje i upravljanje zalihama kako bi se smanjili gubici i povećala efikasnost.

4) Optimizacija lokacija trgovina: Analiza klastera kupaca može pomoći u identificiranju područja s visokim potencijalom za otvaranje novih trgovina ili poboljšanje postojećih lokacija. Na primjer, ako se određeni klasteri kupaca često pojavljuju u određenim geografskim područjima, trgovci mogu ciljati te lokacije za širenje poslovanja.

Ove primjene rezultata projekta mogu doprinijeti boljem razumijevanju kupaca, optimizaciji poslovnih procesa i povećanju konkurentnosti na tržištu. Važno je prilagoditi primjenu rezultata specifičnim potrebama i ciljevima organizacije kako bi se osiguralo maksimalno iskorištavanje njihovog potencijala.

8. Zaključak

U ovom projektu smo primijenili DBSCAN algoritam za segmentaciju kupaca i Random Forest modele za predviđanje prodajnih trendova na skupu podataka trgovine. Analiza rezultata pokazala je da su korišteni algoritmi uspješno identificirali klaster grupa kupaca te predvidjeli prodaju proizvoda s visokom točnošću. Rezultati projekta uključuju potvrdu prikladnosti skupa podataka za analizu, uspješno zadovoljenje pretpostavki algoritama, te identifikaciju potencijalnih poboljšanja i daljnjih koraka. Daljnje istraživanje i implementacija predloženih mjera mogu značajno doprinijeti optimizaciji poslovnih procesa, povećanju konkurentnosti na tržištu te unaprjeđenju usluga i zadovoljstva kupaca. S obzirom na dinamičnost okruženja poslovanja, važno je kontinuirano praćenje rezultata i prilagodba strategija kako bi se osiguralo ostvarenje ciljeva i održala relevantnost analitičkih modela. Ovo istraživanje pruža temelj za daljnja istraživanja u području analitike podataka u trgovinskom sektoru i potiče daljnji razvoj i primjenu naprednih analitičkih tehnika za poboljšanje poslovnih procesa.

Bilježnici s kodom se može pristupiti na linku: https://colab.research.google.com/drive/1BmyvKKdhS1Bfwjm73wTToYzIShu2FbVc?usp=sharing

Popis literature

Agyapong, K. B., Acquah, J. B. H., & Asante, M. (2016). An Overview of Data Mining and Models (Descriptive and Predictive). *International Journal of Software & Hardware Research in Engineering*, *4*(5).

Delen, D., & Ram, S. (2018). Research challenges and opportunities in business analytics. *Journal of Business Analytics*, *1*(1). https://doi.org/10.1080/2573234X.2018.1507324

Greasley, A. (2019). Simulating business processes for descriptive, predictive, and prescriptive analytics. In *Simulating Business Processes for Descriptive, Predictive, and Prescriptive Analytics*. https://doi.org/10.1515/9781547400690

MUKESH. (2022). *Quantium retail data*. Kaggle. https://www.kaggle.com/datasets/mukeshkumar95/quantium-retail-data/data

Neural Ninja. (2023, June 12). *Frequency Encoding: Counting categories for representation*. Letsdatascience.

Super Data Science. (2023, February 19). R Squared Explained in 5 minutes [Video]. YouTube. https://www.youtube.com/watch?v=-7U10N8PvlQ

Popis slika