

Izveštaj o Analizi Podataka i Modeliranju Asocijacija

1. Pregled Skupa Podataka

Skup podataka se sastoji od transakcija u maloprodaji, gde su proizvodi kupljeni od strane članova tokom različitih datuma. Ključne karakteristike:

- **Karakteristike skupa podataka:** Broj clana, Datum, Opis proizvoda.
- **Broj redova:** Skup podataka sadrži veliki broj transakcija (vremenom kreiranih zapisa).
- **Broj jedinstvenih proizvoda:** Postoji velik broj različitih proizvoda koji se kupuju u transakcijama.
- **Nedostajući podaci:** Skup podataka je očišćen od duplikata na osnovu Member_number, Date, i itemDescription.

2. Istraživačka Analiza Podataka

- **Distribucija proizvoda:** Vizualizovane su distribucije kupljenih proizvoda pomoću histograma i scatter plotova.
- **Top proizvodi:** Kroz analizu najčešće kupovanih proizvoda sa Apriori i FPGrowth algoritmima, izdvojeni su proizvodi sa visokim stepenom učestalosti (kao što su "whole milk", "yogurt", "sausage").
- **Outlieri:** Identifikovani su proizvodi koji su se pojavili u vrlo malom broju transakcija, koji mogu ukazivati na specifične niše ili retke kupovine.

3. Modeliranje Asocijacija

Dva algoritma su primenjena za analizu asocijacija u skupu podataka: **Apriori** i **FPGrowth**.

- **Apriori Algoritam:**
 - Postavljen je minimalni prag podrške od 0.001, što omogućava pronalaženje čestih skupa proizvoda.
 - Najčešći skupovi proizvoda: "whole milk", "tropical fruit", "pip fruit".
 - Generisani su asocijativni pravci sa visokim **confidence** i **lift** vrednostima, kao što su:
 - (tropical fruit) -> (whole milk) sa lift-om od ~0.77, što pokazuje snažnu korelaciju između ova dva proizvoda.
- **FPGrowth Algoritam:**
 - Primenjena je slična strategija sa nižim minimalnim pragom podrške (0.0005) za pronalaženje redih asocijacija.

- Top proizvodi uključuju "yogurt", "sausage", "pastry".
- Slične asocijacije kao kod Apriori algoritma, sa većim lift vrednostima za proizvode kao što su:
 - (yogurt, sausage) -> (whole milk) sa lift-om od ~1.62, što znači da postoji snažna veza između ova tri proizvoda.

4. Vizualizacija Rezultata

- **Distribucija Lift-a:**
 - Vizualizovane su distribucije lift vrednosti za asocijacije generisane oba algoritma (Apriori i FPGrowth).
 - Lift za Apriori i FPGrowth pokazuje varijacije, s tim da FPGrowth generiše veći broj visoko koreliranih pravaca.
- **Scatter Plotovi za Confidence vs Lift:**
 - Scatter plotovi za oba algoritma prikazuju odnos između **confidence** i **lift** vrednosti, kao i njihovu povezanost sa **support**. Veći **lift** i **confidence** ukazuju na snažnije asocijacije između proizvoda.

5. Filtriranje Snažnih Pravaca

- **Snažni pravci za Apriori:** Pravci sa **confidence > 0.2** i **lift > 1.5** otkrivaju snažne asocijacije kao što su:
 - (yogurt, sausage) -> (whole milk) sa **lift** vrednošću od 1.62.
- **Snažni pravci za FPGrowth:** Takođe, pravci sa **confidence > 0.2** i **lift > 1.5** otkrivaju značajne povezanosti:
 - (yogurt, sausage) -> (whole milk) sa **lift** vrednošću od 2.48, što ukazuje na vrlo jaku korelaciju između ovih proizvoda.

6. Ključni Uvidi i Zapažanja

- **Popularni proizvodi:** Proizvodi kao što su "whole milk", "yogurt", i "sausage" imaju visok nivo korelacije, što ukazuje na to da se često kupuju zajedno.
- **Snažna povezanost između proizvoda:** Pravci sa visokim lift vrednostima (kao što su (yogurt, sausage) -> (whole milk)) pokazuju jake asocijacije koje mogu biti korisne za marketinške strategije.
- **FPGrowth vs Apriori:** FPGrowth je efikasniji u identifikaciji redjih, ali veoma koreliranih asocijacija zbog nižeg praga podrške.

7. Zaključak

- **Analiza asocijacija** uspešno identifikuje proizvode koji se često kupuju zajedno, što može pomoći u strategijama za promociju i plasman proizvoda u maloprodaji.
- **FPGrowth** algoritam je pružio detaljnije informacije o redjim, ali značajnim asocijacijama, dok je **Apriori** pružio širi pregled čestih asocijacija.