Ülesanne 5 raport

Marko Silla
250418TAF
07.03.2025

1. Palun too välja 5 huvitavamat sagedast hulka andmestikust input1.txt. Põhjenda! Mis parameetrite abil sa need said? Miks need?

```
apriori input1.txt output.txt -ts -m2 -n2 -q-1 -s11

Lihaproduktid Piim (13.3005)
Lihaproduktid Puuviljad (12.8079)
Lihaproduktid Juustud (11.8227)
Piim Puuviljad (11.8227)
Piim Saiad (11.8227)
```

Soovisin näha, milliseid tooteid ostetakse kõige sagedamini koos. Usun, et see info võiks kaupmeestele kasulik olla, aidates planeerida poodides kaupade paigutust nii, et kliendid saaksid oma ostud teha võimalikult kiiresti ja efektiivselt.

2. Palun too välja 5 huvitavamat assotsiatsioonireeglit andmestikust input1.txt. Põhjenda! Mis parameetrite abil sa need said? Miks need?

```
apriori input1.txt output.txt -tr -s5

saiad <- piim (6.40394, 92.3077)
6lled <- pant05 saiad (6.40394, 84.6154)
pant05 <- saiad 6lled (5.91133, 91.6667)
V6id <- margariinid Piim (5.41872, 81.8182)
margariinid <- V6id (7.88177, 100)
```

Soovisin näha, et millised kaubad esinevad sageli koos ostukorvis ning milliste kaupade vahel on tugev seos. Apriori algoritmi väljund näitas, et näiteks margariini ja või vahel on tugev seos (toetus 100%), kuid piima ja saia vahel on nõrk seos (usaldus 6.4%). See aitaks kaupmehel reaalselt mõista ostukäitumist ja kaubavalikut. Näiteks on olemas tugev seos "margariinid <- V6id", sest iga kord, kui ostetakse võid (V6id), ostetakse ka margariini.

3. Palun too välja 5 huvitavamat sagedast hulka andmestikust input2.txt. Põhjenda! Mis parameetrite abil sa need said? Miks need?

```
apriori input2.txt output.txt -ts -m2 -n2 -q-1 -s45

current_act_YES save_act_YES (53.1667)
current_act_YES married_YES (48.8333)
current_act_YES mortgage_NO (50.1667)
save_act_YES married_YES (46.1667)
save_act_YES mortgage_NO (45)
```

Soovisin näha, et millised tegurid on omavahel seotud ning kui sageli need koos esinevad. Apriori algoritmi väljund näitab, et näiteks kui inimesel on aktiivne konto (current_act_YES), siis tal on sageli ka säästukonto (save_act_YES) tõenäosusega 53.17%. Samuti on aktiivse konto omanikel sageli abielustaatus (married_YES) tõenäosusega 48.83%. Üldiselt aitab

see analüüs mõista, millised finantsnäitajad esinevad koos, mis võib olla kasulik pangateenuste pakkumise optimeerimisel.

4. Palun too välja 5 huvitavamat assotsiatsioonireeglit andmestikust input2.txt. Põhjenda! Mis parameetrite abil sa need said? Miks need?

```
apriori input2.txt output.txt -tr -s23

married_YES <- children_0 pep_NO (27.8333, 84.4311)
current_act_YES <- car_NO mortgage_NO (32.8333, 80.203)
married_YES <- pep_NO mortgage_NO save_act_YES (23.6667, 84.507)
married_YES <- pep_NO mortgage_NO current_act_YES (26.3333,
81.6456)
married_YES <- pep_NO mortgage_NO (34.8333, 81.8182)
```

Soovisin näha, millised omadused ja finantsnäitajad esinevad sageli koos. Apriori algoritmi tulemused näitavad, et abielustaatust, pangakonto omamist ja laenu puudumist seostatakse sageli teiste finants- ja elustiiliteguritega. See aitab tuvastada mustreid, mis võivad olla kasulikud finantsanalüüsis ja kliendiprofiilide koostamisel.

5. Mis oli selle teema praktiseerimisel kõige tüütum?

Kõige tüütum osa oli andmete eeltöötlus (data preprocessing) Apriori algoritmi jaoks, kuid see oli ootuspärane, kuna sellele oli loengutes eelnevalt tähelepanu juhitud (reaalelus võib see olla töömahust kuni 60-70%).

6. Mis sa uut enda jaoks õppisid?

Hakkas huvitama, kuidas see implementatsioon on tehtud. Nägin, et kõvasti on vaeva nähtud (kui koodist päist vaadata siis eeldatavalt aastast 1996!) ja see on teadus ja arendustöö oma parimas võtmes. Väga lahe oli teada apriori implementatsioonist. Dokumentatsioonist olulisim, mis ma enda jaoks “kaasa” võtan: “Assotsiatsioonireegli **usaldus** (confidence) näitab, kui suur osa eeltingimusele vastavatest juhtudest vastab ka järeldusele. **Toetus** (support) näitab, kui sageli see reegel esineb kogu andmestikus.”

7. Kes Amazoni ja Netflixi kõrval võiks sellega kõige rohkem raha teenida?

Küsitlusfirmad? Kõik, kes tegelevad andmekaevandamise/analüüsiga - valdkonniti pole piire. Dokumentatsioonist oli toodud näide valimiste usaldusväärsusest. Kui küsitlusfirma soovib prognoosida erakonna toetust, küsitleb ta näiteks 2000 inimest, kellest 857 ütlevad, et hääletaksid selle erakonna poolt. Sellest arvutatakse toetusprotsendiks 857 / 2000 = 43%. Kui aga küsitletaks vaid 100 inimest, kellest 43 vastaks jaatavalt, oleks tulemus 43%, kuid juhuslikud kõikumised oleksid palju suuremad. Suurem valim (2000 inimest) muudab tulemuse usaldusväärsemaks, sest see vähendab juhuslikke kõrvalekaldeid ja peegeldab paremini kogu inimeste eelistusi.