Tampereen ammattikorkeakoulu

HARJOITUSTYÖ

Marika Verger

Tomi Salminen

Hanna Rantanen

Jenna Räty

DATA-ANALYTIIKKA
Marraskuu 2021

Tietotekniikka

SISÄLLYS

1 OSIO 1

1.1 Ryhmän muodostaminen

Muodostimme neljän hengen ryhmän harjoitustyötä varten. Ryhmämme jäsenet ovat Marika Verger, Tomi Salminen, Hanna Rantanen ja Jenna Räty.

1.2 Datasetin valitseminen

Tutkittavaksi datasetiksi valitsimme cereal.mat-datasetin, jossa on tietoa 77 amerikkalaisesta murovalmisteesta. Kyseisessä datasetissä on 15 eri muuttujaa.

1.3 Dataan tutustuminen

a) Datakanavia on yhteensä 15, joista suurin osa liittyy murovalmisteen ravintoarvoihin. Muuttujien englanninkieliset nimet on suomennettu suluissa. Listassa esitetään myös, missä yksikössä lukuarvot on ilmaistu. Cereal-datasetin muuttujat ovat alla listattuna datasetin määräämässä järjestyksessä.

1. Name (murovalmisteen nimi)
2. Mfg (valmistaja)
    a) N = Nabisco
    b) Q = Quaker Oats
    c) K = Kellogg’s
    d) R = Ralston Purina
    e) G = General Mills
    f) P = Post
    g) A = American Home Food Prod
3. Type (tyyppi, 1 = kylmä, 0 = kuuma)
4. Calories (kalorit, per kuppi (2,4 dl))
5. Protein (proteiini, g)
6. Fat (rasva, g)

7. Sodium (natrium, mg)
8. Fiber (kuitu, g)
9. Carbohydrates (hiilihydraatit, g)
10. Sugars (sokerit, g)
11. Shelf (hylly, lattiasta ylöspäin laskettuna 1, 2 tai 3)
12. Potass (kalium, mg)
13. Vitamins (vitamiinit ja mineraalit, kolme luokitusta)
    a. 0 = ei mitään lisätty
    b. 25 = 25% FDA:n suosittelema
    c. 100 = 100% FDA:n suosittelema)
14. Weight (yhden tarjoiltavan annoksen paino, oz (28,35 g))
15. Cups (tarjoiltava annos, kuppi (2,4 dl))

b) Koska datasetin jotkin muuttujat oli ilmoitettu amerikkalaisilla yksiköillä (kuppi, unssi), oli datan ymmärtäminen ja dokumentointi hankalampaa. Tämän lisäksi esimerkiksi valmistaja-muuttujaan oli talletettu vain valmistajan alkukirjain, mistä oli vaikea päätellä valmistajan nimeä. Nämä tiedot löytyivät kuitenkin internetistä.

Datasettiä tutkiskellessa huomattiin, että joidenkin muuttujien kohdalla datassa oli negatiivisia arvoja. Esimerkiksi cups-, sugars- ja potass-muuttujien kohdalla, joissa pienimmän mahdollisen arvon pitäisi olla 0, esiintyi arvoa -1.

1.4 Alustava data-analytiikan tavoite

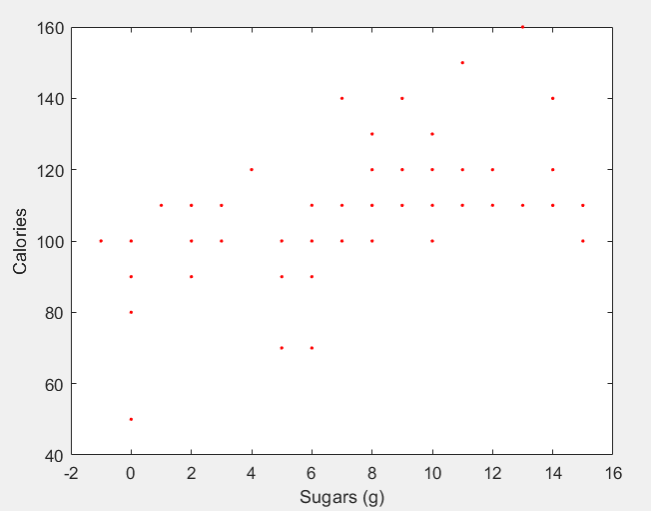
Datasetin muuttujia tarkastellessa mielekkäänä tavoitteena data-analyysille voisi olla eri muuttujien suhde toisiinsa sekä se, voiko niiden avulla ennustaa jonkin toisen muuttujan arvoa. Aihetta pohtiessa voidaan nopeasti muodostaa looginen yhteys esimerkiksi murojen sokerin ja kaloreiden määrän välillä. Data-analytiikan perustehtävä olisi tässä tapauksessa regressio.

2 OSIO 2

2.1 Datan visualisointi

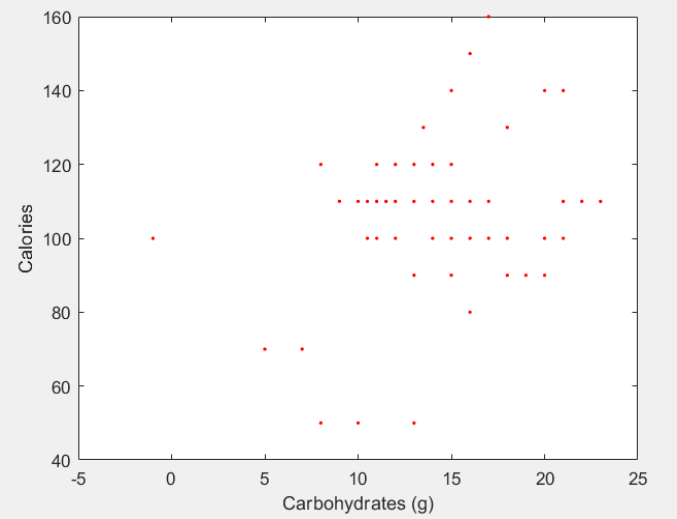
Pisteparvikuvaajissa on havainnollistettu kaloreiden sekä niihin vaikuttavien muuttujien suhdetta. Ensimmäisessä tapauksessa tarkasteltiin sokereiden ja kaloreiden yhteyttä. Kuvaajasta voidaan havaita, että trendi on nousujohdanteinen: sokereiden määrän lisääntyessä myös kalorit lisääntyvät.

```
>> plot(Sugars, Calories, 'r.')
```



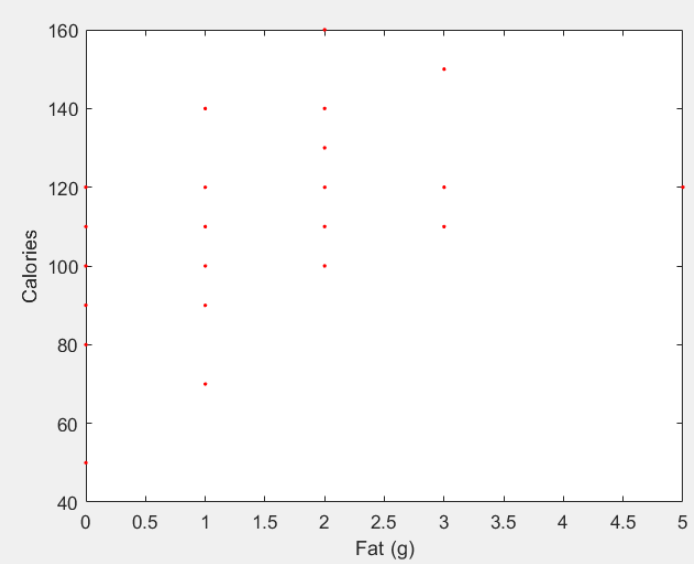
Toisessa tapauksessa tarkasteltiin hiilihydraattien ja kaloreiden yhteyttä. Myös tässä tapauksessa voidaan havaita, että trendi on nousujohdanteinen.

```
>> plot(Carbo, Calories, 'r.')
```



Kolmannessa tapauksessa tutkitaan rasvan yhteyttä kaloreihin. Koska rasvojen arvot ovat 0-5 välillä, mittapisteitä vaikuttaa olevan vähemmän aikaisempiin tarkastelutapauksiin verrattuna. Parvikuvaajasta voidaan silti huomata yhteys, että rasvojen määrän kasvaessa myös kalorien määrä kasvaa.

```
>> plot(Fat, Calories, 'r.')
```



2.2 Keskitunnusluvut

Sokereiden ja kaloreiden keskiarvot:

```
>> mean(Sugars)

ans =

    6.9221

>> mean(Calories)

ans =

  106.8831
```

Hyllyjen moodi on 3. Suurin osa datan muroista on siis ylimmällä hyllyllä. Tämä on loogista, sillä keskipituisen aikuisen ostopäätös kohdistuu helpoiten ylemmäs kuin lattiatasoon sijoitettuihin tuotteisiin.

```
>> mode(Shelf)

ans =

     3
```

Seuraavaksi tarkastelimme murojen sokeripitoisuuden ja niiden hyllypaikan välistä suhdetta. Laskimme Matlabin avulla keskiarvot murojen sokereista hyllyittäin.

```
>> shelf1 = find(Shelf==1);
>>
>> mean(Sugars(shelf1))

ans =

    4.8000

>> shelf2 = find(Shelf==2);
>> mean(Sugars(shelf2))

ans =

    9.6190

>> shelf3 = find(Shelf==3);
>> mean(Sugars(shelf3))

ans =

    6.5278
```

Saaduista keskiarvoista huomataan, ettei sokeripitoisuus muroissa jakaudu tasaisesti hyllyjen välillä. Toisella hyllyllä sijaitsevien murojen sokeripitoisuus on suurin ja se on noin kaksinkertainen ensimmäiseen eli alimmaiseen hyllyyn verrattuna. Jakauma tuskin on sattumaa, sillä toinen hylly yleensä sijaitsee lasten silmien korkeudella. Lapset suosivat yleensä sokeripitoisempia muroja, jolloin murojen sijainti voi vaikuttaa ostopäätökseen.

3 OSIO 3

3.1 Data-analytiikan tavoite

Data-analytiikan tavoite on selvittää, voidaanko loogisesti toisistaan liittyvillä muuttujilla ennustaa vastemuuttujan arvoja.

3.2 Tarpeellinen datan puhdistaminen ja esikäsittely

Potass-, sugars- ja cups-datakanavien negatiiviset arvot halutaan korvata arvolla 0.

Sokereiden arvoja ennen puhdistamista:

```
    10
    14
     3
     0
     0
     6
    -1
    12
     8
```

```
>> SugarsCleaned = max(Sugars, 0); % Puhdistetaan datasta miinusarvot korvaamalla ne nollalla
```

Samat arvot puhdistamisen jälkeen:

```
    10
    14
     3
     0
     0
     6
     0
    12
     8
```

3.3 Menetelmän valitseminen

Menetelmänä käytetään päätöspuita sekä satunnaismetsää regressiolle.

Aluksi luodaan all-matriisi, joka sisältää kaikki tarkasteltavat muuttujat.

```
>> all = [Calories Carbo Cups Fat Fiber Potass Protein Shelf Sodium SugarsCleaned Type Vitamins Weight];
>> % Luodaan matriisi joka sisältää tarkasteltavat muuttujat
```

Talletetaan y-muuttujaan eli vastemuuttujaan kalorit (all-matriisin ensimmäinen sarake). X-muuttujaan talletetaan selittävät muuttujat eli hiilihydraatit, rasvat ja sokerit (all-matriisin toinen, neljäs ja kymmenes sarake).

```
>> y = all(:,1); % Talletetaan vastemuuttujaksi kalorit
>> X = [all(:,2) all(:,4) all(:,10)]; % Talletetaan selittävät muuttujat (hiilihydraatit, rasvat, sokerit)
```

3.4 Valitun menetelmän soveltaminen

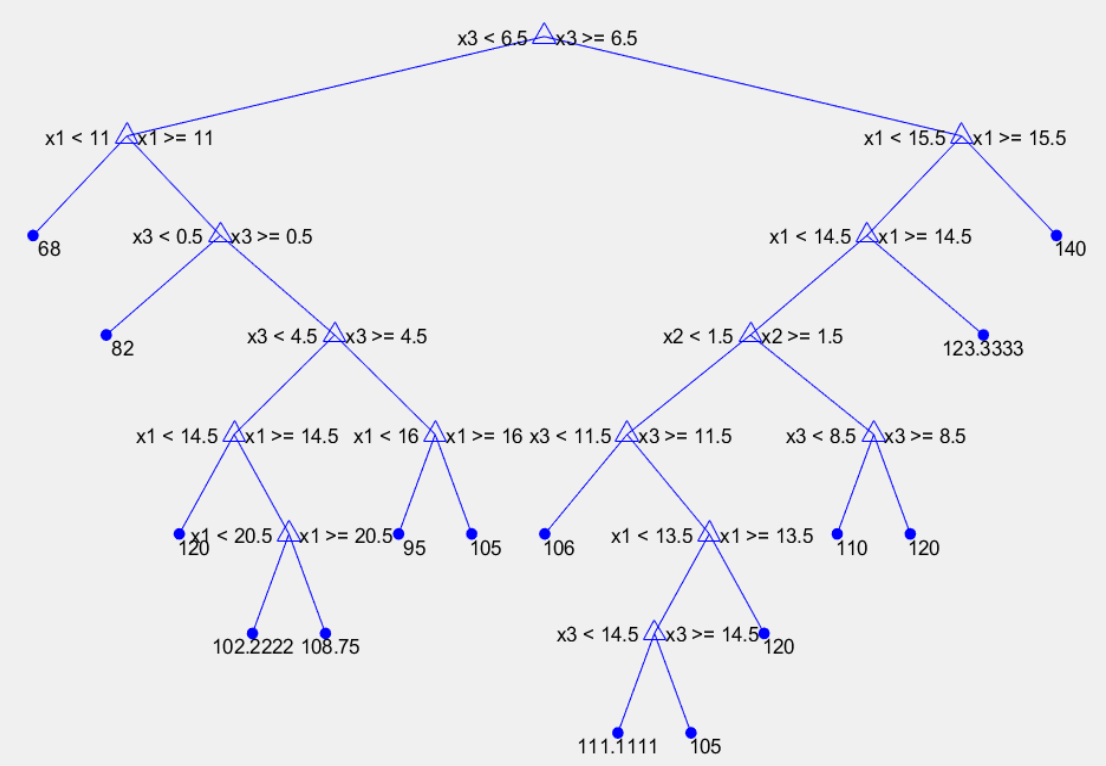
Kasvatetaan regressiopuu fitrtree-komennolla. Tämä on data-analyysin opetusvaihe.

```
>> cerealTree = fitrtree(X, y); % Regressiopuun kasvatus (opetusvaihe)
```

Opetusvaihe on visuaalinen esitystapa päätöspuiden ”oksien” jakautumisesta. Arvojen kautta voidaan lukea jakamisen syy ja mistä parametrista on kyse. Kuva auttaa havainnollistamaan, kuinka monta jakoa päätöspuu tekee.

Regressiopuun rakennetta voidaan visualisoida view-komennolla.

```
>> view(cerealTree, 'mode', 'graph')
>> % Visualisoidaan puurakenne
```



Ennustusvaiheessa pyritään ennustamaan kalorien määrä, kun hiilihydraateille, rasvoille ja sokereille määritetään ulkopuoliset arvot. Ennustaminen suoritetaan predict-komennolla.

```
>> predict(cerealTree, [12 3 14]) % Ennuste omalle havainnolle: hiilihydraatit 12g, rasvat 3g, sokerit 14g

ans =

  120.0000
```

Kasvatetaan satunnaismetsä fitrensemble-komennolla. Tämä on satunnaismetsän opetusvaihe.

```
>> cerealForest = fitrensemble(X, y); % Satunnaismetsän opetusvaihe
```

Ennustusvaihe tehdään samoin kuin päätöspuille predict-komennolla. Selittäville muuttujille asetetut arvot ovat samat. Tuloksesta nähdään, että satunnaismetsän avulla tehty ennustus on tarkempi arvo kuin päätöspuilla ennustettu.

```
>> predict(cerealForest, [12 3 14]) % Ennuste omalle havainnolle: hiilihydraarit 12g, rasvat 3g, sokerit 14g

ans =

  127.4743
```

Testidatasetin laajuudeksi päätettiin 20 %, joka 77 havainnosta on 16 arvoa. Dataa ei ole järjestetty kasvavaan tai vähenevään järjestykseen vaan listattu matriiseihin satunnaisesti, jolloin testidatasetiksi voidaan valita 16 viimeistä arvoa.

Tehdään päätöspuiden opetusvaihe siten, että datasta jätetään ulkopuolelle testidatasetiksi yllä mainittu 20 % eli 16 arvoa.

```
>> cerealTreeTest = fitrtree(X(1:end-16, :), y(1:end-16));
>> % Opetusvaihe ilman 20% arvoista (viimeiset 16 arvoa)
```

Tehdään ennustus 16 viimeiselle havainnolle kaloreiden määrästä.

```
>> predict(cerealTreeTest, X(62:77,:)) % Ennuste viimeiselle 16 havainnolle

ans =

  102.8571
  102.8571
   75.0000
   75.0000
   75.0000
  107.7778
   96.2500
   96.2500
  102.8571
  115.0000
   96.2500
  102.8571
  110.0000
  110.0000
  110.0000
  145.0000
```

Verrataan ennustettuja kaloriarvoja todellisiin arvoihin y-matriisissa.

```
>> y(62:77,:)

ans =

   110
   110
    80
    90
    90
   110
   110
    90
   110
   140
   100
   110
   110
   100
   100
   110
```

Tehdään satunnaismetsän opetusvaihe siten, että datasta jätetään ulkopuolelle testidatasetiksi yllä mainittu 20 % eli 16 arvoa.

```
>> cerealForestTest = fitrensemble(X(1:end-16, :), y(1:end-16));
>> % Opetusvaihe ilman 20% arvoista (viimeiset 16 arvoa)
```

Tehdään ennustus 16 viimeiselle havainnolle kaloreiden määrästä.

```
>> predict(cerealForestTest, X(62:77, :)) % Ennuste viimeiselle 16 havainnolle

ans =

  100.0949
  109.9854
   71.9168
   90.4250
   90.4250
  116.7822
   94.8864
   96.7958
  110.0045
  120.1764
  100.0251
  110.0045
  109.9078
  109.2387
  109.2387
  136.3889
```

Verrataan jälleen ennustettuja kaloriarvoja todellisiin arvoihin y-matriisissa.

```
>> y(62:77,:)

ans =

   110
   110
    80
    90
    90
   110
   110
    90
   110
   140
   100
   110
   110
   100
   100
   110
```

3.5 Data-analyysin tulokset ja havainnot

Analyysien menetelmäparametrit olivat kalorit, hiilihydraatit, rasva ja sokeri. Hiilihydraatit, rasva ja sokeri olivat selittävät muuttujat ja kalorit olivat selitettävä eli vastemuuttuja.

Saimme mallin ennustamiseen sekä päätöspuista että satunnaismetsästä.

Testidatasetillä mallin toimivuus oli kohtalaisen hyvä. Malli ennustaa kalorien arvot kymmenien tarkkuudella. 16 viimeiselle arvolle pienin heitto alkuperäisestä arvosta on 0 mittayksikköä ja isoin heitto 35 mittayksikköä. Tarkemmat ennusteet saatiin hyödyntämällä satunnaismetsää: tämä on loogista, sillä se laskee ennustukseen keskiarvon useiden päätöspuiden avulla.

Lopputulokset ovat järkeviä. Datasta valittiin selitettäväksi muuttujaksi kalorit ja selittäviksi muuttujiksi sokerit, rasvat ja hiilihydraatit, sillä näiden pääteltiin vaikuttavan kalorien määrään. Malliksi muodostettiin ensin päätöspuut ja niistä muodostettiin satunnaismetsä. Päätöspuilla tehtiin ennustus kalorien määrästä, jolla saatiin melko hyviä arvoja, mutta satunnaismetsällä tehdyllä ennustuksella saadut kaloreiden määrät olivat hyvin lähellä alkuperäisiä. Tehty malli voidaan todeta siis toimivaksi.