Elyas Addawe – Saku Ihalainen – Hanna Kaimo

Logistinen regressio (CRISP-DM)

Tehtävä n:o 9

Metropolia Ammattikorkeakoulu
Tieto- ja viestintätekniikka
Datan käsittely ja koneoppiminen
11.4.2022

1 Tavoitteet

Aineiston henkilöt ovat saaneet ensimmäisen sydänkohtauksen ja toipuneet siitä. Tavoitteena on logistista regressioanalyysiä käyttäen löytää aineistosta ne terveydentilaa kuvaavat tiedot ja henkilön elintavat, joiden takia henkilö saa todennäköisesti toisenkin sydänkohtauksen. Viimeinen muuttuja sisältää tiedon, onko näin käynyt. Näin tuloksen pohjalta saadaan toivottavasti tietoja, joiden avulla voidaan kohdentaa sydänkohtauksen saamisen ennaltaehkäisytoimet riskihenkilöille.

2 Data

Aineisto on saatu kuvitteellisilla haastatteluilla ja laboratoriomittauksilla eli se on itse generoitu. Attribuuttien määrä on 10.

Attribuutit:

ID: Numeerinen muuttuja, vaihteluväli 1 - 838
Paino: Numeerinen muuttuja, vaihteluväli 27-130 kg
Tupakointi: Numeerinen muuttuja, arvo joko 0 = ei tupakoi tai 1 = henkilö tupakoi
Liikunta: Numeerinen muuttuja, vaihteluväli 0-10
Kolesteroli: Numeerinen muuttuja, vaihteluväli 0.1-9.9
Kuukausitulo: Numeerinen muuttuja, vaihteluväli -790-5860 €
Koettu onnellisuus: Numeerinen muuttuja, vaihteluväli 0-100
Syntymävuosi: Numeerinen muuttuja, vaihteluväli v.1879-2020
Sukupuoli: Nominaali muuttuja, arvo joku M = Mies tai N = Nainen
Kohtaus: Nominaali muuttuja, arvo joko 0 = ei sydänkohtausta tai 1 = henkilö on saanut sydänkohtauksen

Datassa oli puuttuvia arvoja. Generoimme datan ReplaceMissingValues- filterillä, jolloin Weka poisti meille puuttuvat arvot (Kaikki missing values = 0%). Poikkeavia arvoja on painossa, kolesterolissa, kuukausitulossa ja syntymävuodessa.
Datan koko on 838 havaintoa.

3 Datan valmistelu

Avasimme Terveys_v3.csv:n Wekassa. Poistimme turhan ID- attribuutin. Aloitimme datan attribuuttien arvojen karsimisen, poistimme mahdottomat arvot, esim. henkilön painon vaihteluväli oli 8 kg - 740 kg, muutimme vaihteluvälin 47 kg - 124 kg. Teimme sen Wekan RemoveWihValues- filterillä. SplitPointit olivat 45 kg ja 250 kg, vaihdoimme vain invertSelection- kohdan arvoon true. Vaihdoimme henkilön kolesteroliarvon vaihteluvälin 4-7.9 ja syntymävuoden 1915-2012. Vaihdoimme datan tiedostomuodon csv- muotoon arff.
Mahdottomat arvot voivat vääristää tutkimustuloksia.

4 Mallinnus

Käytämme analysointiin opettajan pyynnön mukaan Wekan Simple Logistic- algoritmia. Datan validointiin käytämme Cross- Validation Test Option- vaihtoehtoa. Opetusjoukon tarkkuus on aika ylhäällä, n. 73.15 %.

```
Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         613               73.1504 %
Incorrectly Classified Instances       225               26.8496 %
Kappa statistic                          0.4399
Mean absolute error                      0.3394
Root mean squared error                  0.415
Relative absolute error                 69.6611 %
Root relative squared error             84.085  %
Total Number of Instances              838

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,811     0,378     0,748      0,811     0,778      0,442     0,815     0,858     0
               0,622     0,189     0,704      0,622     0,661      0,442     0,815     0,762     1
Weighted Avg.  0,732     0,299     0,729      0,732     0,729      0,442     0,815     0,818

=== Confusion Matrix ===

   a   b   <-- classified as
 394  92 |   a = 0
 133 219 |   b = 1
```

Tuloksen perusteella näemme, että kohtausta edeltää (Class 1) attribuuttien painon oleminen suuri, henkilön tupakointi ja kolesteroliarvojen oleminen koholla, liikuntamäärän vähyys ja se, että henkilö on mies. Henkilön kuukausitulo ja syntymävuosi eivät tämän tutkimuksen perusteella vaikuta sydänkohtauksen

saamiseen. Eniten henkilön sydänkohtauksen saamisen riskiä nostaa tupakointi(*0.15) ja vähentää liikunta (* -0.15).

```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

SimpleLogistic:

Class 0 :
7.82 +
[paino] * -0.04 +
[tupakointi] * -0.15 +
[liikunta] * 0.15 +
[kolesteroli] * -0.1 +
[kuukausitulo] * 0    +
[syntymavuosi] * -0 +
[sukupuoli=N] * 0.05

Class 1 :
-7.82 +
[paino] * 0.04 +
[tupakointi] * 0.15 +
[liikunta] * -0.15 +
[kolesteroli] * 0.1  +
[kuukausitulo] * -0 +
[syntymavuosi] * 0    +
[sukupuoli=N] * -0.05
```

Katsoimme, mitä vaikuttaa, kun käytimme Persantace Split- validointia ja katsoimme tuloksia 10 %:n opetusjoukosta. Tämä kuitenkin laski tarkkuuden n. 63.1 %:iin. Jos tekisimme oikeaa tutkimusta, olisi ainakin varteenotettava vaihtoehto käyttää Logistic -vaihtoehtoa SimpleLogistic -vaihtoehdon sijaan, mutta tunnilla pyysit tekemään tämän käyttämällä SimpleLogistic.

SMO- algoritmilla (20 %:n testijoukko) saamme myös kuukausitulon, onnellisuuden ja iän vaikutukset sydänkohtauksen saamiseen (tarkkuus n. 72 %):

```
Kernel used:
  Linear Kernel: K(x,y) = <x,y>

Classifier for classes: 0, 1

BinarySMO

Machine linear: showing attribute weights, not support ve

         3.516  * (normalized) paino
 +       0.4782 * (normalized) tupakointi
 +      -1.8202 * (normalized) liikunta
 +       0.897  * (normalized) kolesteroli
 +       0.0926 * (normalized) kuukausitulo
 +      -0.2787 * (normalized) koettuonnellisuus
 +       0.3306 * (normalized) syntymavuosi
 +      -0.1668 * (normalized) sukupuoli=N
 -       1.5422

Number of kernel evaluations: 61822 (69.626% cached)
```

Onnellisuus vähentää, kuukausitulo ja ikä vähän lisäävät riskiä.

5 Arviointi

Ehkä ainakin yllättävää, että ihmisen tuloilla oli suhteellisen vähän vaikutusta sydänkohtausriskiin. Ehkä se selittyy sillä, että jotkut paljon tienaavat eivät esim. ehdi harrastamaan liikuntaa ja tulee syötyä mitä sattuu, stressi lisää tupakanpolttoa. Ainakin näitä kaikkia asioita esiintyy laajalti alempituloisissa yleisesti ottaen, mutta nämä asiat eivät perustu tähän tutkimukseen vaan ovat omaa pohdintaa yllättävistä asioista tuloksissa.

6 Täytäntöönpano

Saimme selville henkilön sydänkohtauksen saamisen riskiä nostavia tekijöitä. Toimenpidesuositukset olisivat: ihmisiä kannattaa kannustaa lopettamaan tupakan poltto tai olemaan aloittamatta sitä ja lisätä ihmisten liikuntaa. Kannattaa välttää huonoja rasvoja, koska ne kohottavat kolesterolia ja riski sydänkohtaukseen kasvaa. Pudottaa painoa.