Tehtävä 2: Terveys ja elintavat

Tavoite

Tehtävän tavoitteena on oppia datankäsittelyprosessin vaiheet CRISP-DM-mallin mukaisesti. Työssä tutustutaan Weka-ohjelmiston käyttöön ja sovelletaan tarvittaessa muita työkaluja. Keskeiset tekniikat: virheellisten ja puuttuvien arvojen etsintä ja käsittely, muuttujien tyyppien tarkistus.

Aineisto

Työtiedosto on Oma-työtilan **Data**-kansiossa nimellä **Terveys_v3.csv**. (Versionumeron tilalla voi olla suurempi numero).

Aineisto[1] sisältää tutkimushenkilöiden elintapoja ja terveydentilaa kuvaavia tietoja, jotka on saatu haastattelemalla ja laboratoriomittauksin. Koska aineisto on käsin koottu, voi siinä olla runsaastikin virheitä ja puuttuvia tietoja.

Muuttujien tulkinta on seuraava:

Muuttuja	Selitys	Koodaus
ID	Henkilön tunnistenumero.	Juokseva kokonaisluku.
Paino	Paino kilogrammoina terveydenhoitajan mittamaana.	Kokonaisluku (tarvittaessa pyöristetty).
Tupakointi	Henkilön ilmoittama tupakointi. Tupakoivaksi määritellään henkilö, joka säännöllisesti polttaa vähintään kolme savuketta viikossa.	0 = ei tupakoi, 1 = tupakoi.
Liikunta	Liikuntakyselyn perusteella laskettu liikunta-aktiivisuusindeksi. Minimiarvo 1 (fyysisesti täysin passiivinen), maksimiarvo 10 (fyysisesti erittäin aktiivinen).	Kokonaisluku välillä 1-10.
Kolesteroli	Veren kokonaiskolesteroliarvo millimoolia litrassa laboratoriossa mitattuna.	Desimaaliluku yhden desimaalin tarkkuudella.
Kuukausitulo	Henkilön palkoista, pääomatuloista ja tulonsiirroista koostuva kokonaisbruttokuukausitulo itse ilmoitettuna.	Kokonaisluku.
Koettu onnellisuus	Kyselylomakkeeseen perustuva koetun onnellisuuden indeksi. Minimiarvo 0 (äärimmäisen onneton), maksimiarvo 100 (euforinen).	Kokonaisluku välillä 0-100.
Syntymävuosi	Väestötietojärjestelmästä haettu syntymävuosi.	Kokonaisluku.
Sukupuoli	Henkilön ilmoittama sukupuoli.	M=mies, N=nainen.

Kohteena on henkilöt, jotka ovat täyttäneet aineiston keruuhetkellä 18 vuotta. Aineisto kerättiin vuonna 2013.

Kysymyksenasettelu

Tavoitteena on saada kohdennettua valistuskampanjaa varten selville, mitkä tekijät voivat ennustaa henkilön tupakanpolttoa.

Tätä varten halutaan pieni ja selkeä (muutaman haarautumasolmun) päätöspuu henkilön tupakoimisen ennustamiseksi muiden muuttujien perusteella. Toissijaisesti halutaan saada selville, mitkä muuttujat aineistossa näyttävät olevan toisistaan riippuvaisia.

[1] Aineisto on tietokoneella generoitu, mutta se esitetään tässä kuin se olisi todellinen.

Voit tuottaa päätöspuun J48-päätöspuuluokittelualgoritmilla (*decision tree classification algorithm*). Päätöspuun koon säätö onnistuu muuttamalla (yleensä pienentämällä) *confidence factor* -parametria. Voit jättää muut parametrit oletusarvoihinsa; päätöspuihin, validointimenetelmiin ja parametreihin palataan opintojaksolla myöhemmin.

Muuttujien riippuvuuden tutkimiseen voit mm. käyttää hajontakuvia (Weka Explorerin Visualize -välilehti).

Jotta saat päätöspuun aikaiseksi, joudut tutustumaan aineistoon huolella ja tekemään sille esikäsittelyä. Tämän tehtävän painopiste on aineistoon tutustumisessa ja esikäsittelyssä sekä aineiston ja esikäsittelyn dokumentoinnissa (CRISP-mallin kohdat 2 ja 3): tee ja dokumentoi ne huolellisesti!

Tuotos

Tuotoksena palautetaan analyysiraportti, jossa selostaan datankäsittelyprosessin eteneminen CRISP-DM-mallin mukaisesti. Raporttipohja on Oma-työtilan Data-kansiossa nimellä **Raportin_mallipohja.docx**.

Arviointi ja palautus

Työ palautetaan Oman tehtäväpalautuslinkin kautta seuraavaan oppituntiin mennessä. Työstä saa korkeintaan kolme pistettä. Vaatimukset kolmen pisteen ratkaisulle:

- Kaikki datankäsittelyprosessin vaiheet on dokumentoitu selkeästi ja siten, että prosessi on toistettavissa.
- Aineiston virheet ja epävarmuudet on kuvattu, korjattu mahdollisuuksien mukaan ja otettu huomioon analyysissä ja tulosten tulkinnassa.
- Aineistolle tehdyt toimenpiteet ovat oikeita ja hyvin perusteltuja.
- Tulokset ovat virheettömiä ja niiden tulkinta sekä johtopäätökset ovat hyvin perusteltuja.

Jos työ arvioidaan nollaksi pisteeksi, on se korjattava vaatimusten mukaiseksi. Tämän jälkeen hyväksyttävästi korjatusta työstä saa yhden pisteen.