Elyas Addawe – Saku Ihalainen – Hanna Kaimo

Analyysiraportti (CRISP-DM)

Tehtävä n:o 1

Metropolia Ammattikorkeakoulu
Tieto- ja viestintätekniikka
Datan käsittely ja koneoppiminen
16.3.2022

1 Tavoitteet

Tämän harjoituksen tavoite on käsitellä tiettyä, Terveys_v3- dataa käyttäen prosessoinnissa CRISP- DM- mallin mukaisia vaiheita. Työkaluna käytämme Weka-ohjelmaa, jonka käyttöön tutustumme samalla. Ohessa käytämme ainakin Notepad++ ohjelmaa ja tarvittaessa muita ohjelmia. Etsimme virheelliset ja puuttuvat arvot ja käsittelemme ne sekä tarkistamme datan tietotyypit.
Eksaktina tehtävänä on saada aineistosta selville ne asiat, jotka ennustavat henkilön tupakan polttoa. Eli esimerkiksi ne relaatiot, jotka pätevät, muuttujien välillä, jos henkilö polttaa tupakkaa. Näin fiktiivinen tuleva valistuskampanja voidaan kohdistaa oikeille ihmisille.
Tehtävään kuuluu myös tutkia, mitkä muuttujat aineistossa ovat toisistaan riippuvaisia.

2 Data

Data on kerätty jonkinlaisella suullisella tai kirjallisella kyselyllä esim. puhelimitse tai sähköpostitse.

Attribuutit:

Henkilön paino: numeerinen muuttuja, vaihteluväli datassa 8-740 kg, 1 arvo puuttuu
Tupakointi: numeerinen muuttuja, attribuutin arvo on joko 0= ei tupakoi tai 1= tupakoi, 1 arvo puuttuu
Liikunta: numeerinen muuttuja, vaihteluväli datassa asteikolla 1-10, 8 arvoa puuttuu
Kolesteroli: String, vaihteluväli n. -7.5- 9.5
Kuukausitulo: numeerinen muuttuja, vaihteluväli 0-7690 €, 7 arvoa puuttuu
Koettu onnellisuus: numeerinen muuttuja, vaihteluväli 1-100, 3 arvoa puuttuu
Syntymävuosi: numeerinen muuttuja, vaihteluväli 1879-2020, 1 arvo puuttuu
Sukupuoli: nominaali- muuttuja, attribuutin arvo on joko M= mies tai N= nainen, poikkeavat arvot: Mies, Nainen

Yhdellä rivillä on yhdeksän arvoa, joten sen henkilön datan tulkinta on hyvin hankalaa(millä numerolla henkilö tarkoittaa mitäkin arvoa):

109,71,1,6,7,970,32,1986,M

Datan koko on 1000 henkilöä. 1000 henkilöä on mielestäni kohtuullisen pieni otanta, joten tulosten yleistäminen voi olla hieman hankalaa. Tuhat henkilöä on esimerkiksi suunnilleen Metropoliassa työskentelevien henkilöiden määrä. Tietysti kokoon vaikuttaa halutaanko tutkia tiettyä ihmis-ryhmää, jolloin otanta kyseisestä ryhmästä voi olla suppeampi kuin jos tutkinta- joukko on esimerkiksi: kaikki ihmiset.

Datan mahdollisia ongelmia ja virhelähteitä:

Moni ihminen kokee onnellisuuden eri tavalla. Ehkä onnellisuuden mittaamiseen voisi olla oma kyselynsä, jossa ihmisen onnellisuutta mitattaisiin samojen asioiden pohjalta. Koettu onnellisuus voi myös esim. vaihdella ihmisellä esimerkiksi eri päivinä.
Liikunnan määrän mittaamisessa vain yhdellä arvolla voi olla samankaltaisia tulkinta-ongelmia esim. määrässä ja mikä lasketaan liikunnaksi ja mikä on esim. "paljon liikuntaa".
Muut attribuutit ovat ihan konkreettisesti mitattavissa olevia arvoja tai joko tai vaihtoehtoja, joita ei tarvitse mitatessa tulkita.

3 Datan valmistelu

Korvasin attribuuttien arvojen pilkut pisteillä. Viimeisen sarakkeen Sukupuoli- arvot: Mies arvolla: M. korvattu sarakkeen Sukupuoli- arvot: Nainen arvolla: N. Sarakkeen 4, kolesteroli tyyppi on virheellisesti String, muutin sen numeeriseksi arvoksi. Muutin datan CSV- tiedostomuodon ARFF- tiedostomuodoksi, koska tätä muotoa on helpompi käsitellä Wekassa. Laskin mukaan puuttuvien arvojen prosentuaaliset osuudet kaikista arvoista.

Muutetut attribuutti- arvot:

Paino: 38- 120 kg
Kolesteroli: 2.4- 7.9
Kuukausitulo: 230- 7690 €
Syntymävuosi: 1920- 1995

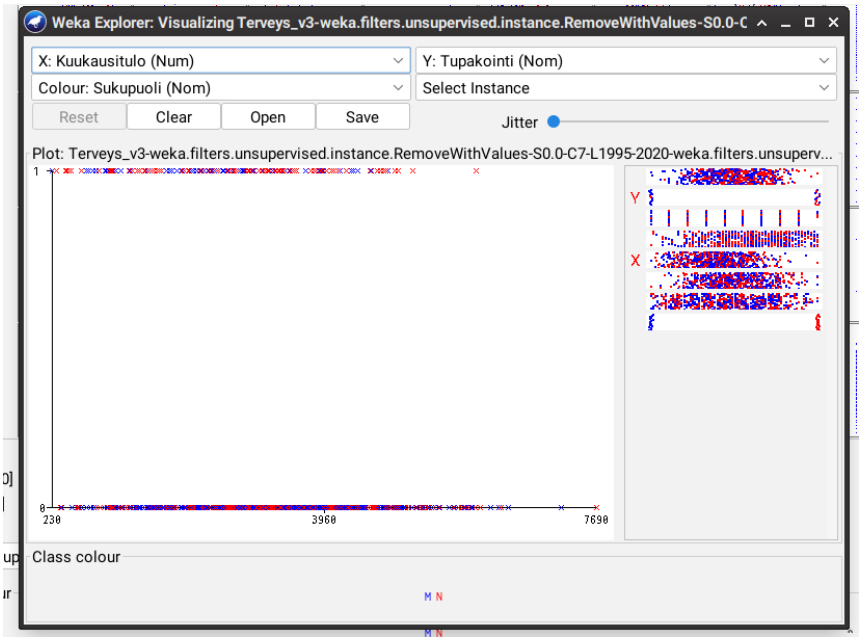
4 Mallinnus

Käytimme datan analysointiin päätöspuuta (decision tree), jonka mallinsimme Wekalla. Tämän tiedonlouhintamenetelmän yksi parhaista puolista on se, että puu on selkeä ja helposti ymmärrettävä tapa esittää asia. Ja sen teko Weka- ohjelmalla on melko yksinkertaista. Visuaalisia graafeja tutkimalla(Weka visualize) selvisi kiinnostavia yhteyksiä.

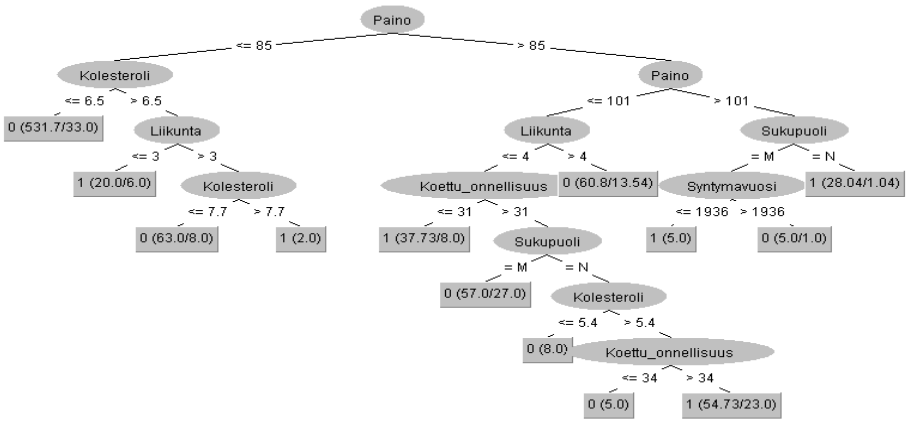
Tuloksista päätellen tupakointiin vaikuttaa kolme tekijää: kuukausitulojen vähyys, korkea kolesteroli ja korkea paino.

Tupakoimattomien ryhmällä oli tasaisesti erilaisia kolesteroliarvoja. Tupakoivien

ryhmässä oli hyvin vähän henkilöitä, joiden kolesteroli olisi alle 5.



Tupakoivien ja tupakoimattomien kuukausitulot olivat samankaltaiset muilta osin paitsi korkeissa kuukausituloissa. Tupakoivat eivät yllä korkeimpiin tuloihin, joihin tupakoimattomat yltävät.



5 Arviointi

Tupakoivilla koe- henkilöillä menee paljon rahaa tupakkaan, jonka takia kuukausitulot ovat pienemmät. Tupakkaa voi nauttia paljon, sillä nikotiini poistuu elimistöstä huomattavasti nopeammin kuin alkoholi. Tupakkaa voi tästä syystä nauttia huomaamattomasti todella paljon. Tupakan hintaa on nostettu tarkoituksella viime

vuosikymmeninä kansanterveydellisistä syistä.
Myös se voisi vaikuttaa, että hyvin korkea palkka usein tarkoittaa korkeaa koulutustasoa ja sitä kautta suurempaa määrää tietoa tupakan mahdollisista haittavaikutuksista. Myös jos tekee paljon töitä ei halua käyttää aikaansa tupakalla käymiseenkin, mikä vähentäisi työaikaa ja vähäistä vapaa- aikaa.

Tupakka vaikuttaa verenpaineeseen, sillä nikotiini tukkeuttaa verisuonia lisäten verenpainetta. Tulokset olivat loogisia kolesterolin osalta. Yllättävää oli kuitenkin se, että jotkut koehenkilöt kykenivät pitämään matalan verenpaineen tupakoinnista huolimatta. Tämän voi selittää ehkäpä siten että koehenkilöt nauttivat tupakkaa vain vähän. Datassa ei ollut tietoa kuinka paljon koehenkilöt nauttivat tupakkaa.

Tupakan polttaminen näyttäisi korreloivan korkeamman painon kanssa ja tupakka nostaa verenpainetta. Myös paljon liikuntaa harrastavien henkilöiden joukossa on vähemmän tupakoivia kuin vähän liikkuvien. Tupakka vaikuttaa negatiivisesti ihmisen fyysiseen kuntoon ja liikuntasuorituksiin. Joten tämä yhtälö: korkea paino + tupakointi + vähän liikuntaa toteutuu suhteellisen usein. Liikunta lisää ihmisen onnellisuuden tuntemuksia.

Myös esimerkiksi perintötekijät vaikuttavat siihen, miten tupakointi vaikuttaa terveyteen.

6 Täytäntöönpano

Muuttujat tietojen hankintaan oli valittu ihan hyvin, koska niillä oli jotain vaikutusta tutkittavaan asiaan, henkilön tupakan polttoon.
Toimenpidesuosituksena kannattaa valistaa ja kampanjoida henkilöitä pudottamaan painoaan, harrastamaan liikuntaa, opiskelemaan ja tekemään työtä. Voisi sanoa ympäripyöreästi, että ihmisen aktiivisuus kannattaa. Myös tupakan hinta kannattaa pitää kohtuullisen korkeana.
Näin voidaan yrittää minimoida ihmisten tupakan polttamisen aloittaminen, pyrkiä ihmisiä vähentämään tupakanpolttoa ja mieluiten lopettamaan se kokonaan.