Päätöspuut ja niiden käyttö terveydenhuollossa päätöksenteon tuen menetelmänä

Kristian Koivisto

Kandidaatintutkielma
HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

Helsinki, 13. marraskuuta 2015

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Matemaattis-luonnontieteellinen	Tietojenkäsittelytieteen laitos

Tekijä — Författare — Author
Kristian Koivisto

Työn nimi — Arbetets titel — Title

Päätöspuut ja niiden käyttö terveydenhuollossa päätöksenteon tuen menetelmänä

Oppiaine — Läroämne — Subject
Tietojenkäsittelytiede

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Kandidaatintutkielma	13. marraskuuta 2015	21

Tiivistelmä — Referat — Abstract

Tässä työssä perehdytään päätöspuiden perusteisiin, niiden muodostamisessa käytettyihin induktioalgoritmeihin sekä päätöspuiden sovellukseen terveydenhuollossa. Induktioalgoritmeista ID3:een perehdytään perusteellisemmin. Päätöspuut ovat tehokkaita luokittimia ja niitä voidaan käyttää päätöksenteon apukeinona tilanteissa, joissa päätöksenteko on luonteeltaan luokittelua - esimerkiksi lääketieteellisessä diagnostiikassa on usein kyse siitä, mihin luokkaan tai diagnoosiin potilastapaus kuuluu sen ominaisuuksien perusteella. Erilaisista luokittimista päätöspuut saattavat soveltua terveydenhuoltoon paremmin kuin muun tyyppiset luokittimet, koska niiden toiminta on visualisoitavissa ja lyhyen perehtymisen jälkeen on usein suhteellisen helppo kenen tahansa nähdä, mihin päätöspuun toiminta ja luokittelu perustuu. Tällä on merkitystä, koska terveydenhuollossa erilaisia päätöksentuen apukeinoja on ollut vaikea istuttaa sairaaloiden ja muiden toimintayksiköiden prosesseihin. Yksi syy tähän on, että terveydenhuollon ammattilaisten on vaikea luottaa sovelluksiin, joiden tarjoaman tuloksen syitä ja perusteita ei itse pysty verifioimaan.

Avainsanat — Nyckelord — Keywords
päätöspuu, päätöspuuinduktioalgoritmi, ID3, CART, terveydenhuolto, päätöksenteon tuki

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Sisältö

1 Johdanto

Terveydenhuollossa pyritään resurssit huomioiden takaamaan potilaille parasta mahdollista hoitoa. Tämä edellyttää laadukasta päätöksentekoa. Usein hyvän päätöksen tekeminen potilaan hoitoon liittyvässä päätöksentekotilanteessa on kuitenkin haastavaa ja valitettavan monesti tehty päätös osoittautuu jälkeenpäin epäedulliseksi. Tämän vuoksi terveydenhuollon ammattilaisille on pyritty kehittämään erilaisia päätöksenteon tuen menetelmiä. Luonnollisesti tietojenkäsittelytieteen keinoja on valjastettu terveydenhuollon päätöksenteon laadun parantamiseksi – myös älykkäiden järjestelmien soveltamista tähän on tutkittu aina tekoälyn alkutaipaleelta 1950-luvulta lähtien. Yksi käyttökelpoiseksi osoittautunut menetelmä on päätöspuu.

Tässä kirjoituksessa perehdytään päätöspuihin, niiden muodostamisessa käytettyihin induktioalgoritmeihin sekä muutamiin sovelluksiin terveydenhuollossa.

2 Päätöspuut ja niiden käyttö päätöksenteon tuen menetelmänä

Terveydenhuollon ammattilaiset kohtaavat työssään monenlaisia päätöksentekotilanteita. Mikäli päätöksenteossa on kyse siitä, mihin luokkaan jokin asia kuuluu siitä tehtävien havaintojen perusteella, voi päätöspuista olla apua kyseisen päätöksenteon tukimenetelmänä. Tämäntyyppisiä päätöksentekotilanteita on terveydenhuollossa runsaasti - esimerkiksi menetelmä, jolla voitaisiin luotettavasti kategorisoida päivystyspoliklinikalle saapuva rintakipupotilas pienen tai suuren riskin potilaaksi kuolemanvaaran suhteen voisi olla hyödyllinen. Kyseisen päätöksen tekeminen on harjaantuneellekin lääkärille vaativa tehtävä ja usein alustavan päätöksen joutuu tekemään aloittelija, jolloin virheellisen päätöksen mahdollisuus luonnollisesti kasvaa. Tällaisessa asetelmassa luotettavasta päätöksentuen apukeinosta voisi olla hengenpelastavaakin merkitystä.

Vastaavanlaista luokittelua voidaan aikaansaada monilla muillakin mene-

telmillä, kuten neuroverkoilla ja tukivektorikoneilla. Päätöspuu on kuitenkin terveydenhuollon kontekstissa erityisen mielenkiintoinen vaihtoehto, sillä sen toimintaperiaate on helposti kenen tahansa ymmärrettävissä, siitä voidaan piirtää visuaalinen esitys ja lyhyen perehtymisen jälkeen on usein suhteellisen helppoa nähdä, miksi päätöspuu päätyy tiettyyn tulokseen. Tämä on tärkeää, sillä vaikka monien tekoälysovellusten on osoitettu kykenevän parempaan päätöksentekoon kuin terveydenhuollon ammattilaiset ja näin niistä saattaisi hyvinkin olla hyötyä, niiden istuttaminen sairaaloiden ja muiden terveydenhuollon toimintayksiköiden prosesseihin ei ole onnistunut odotetunlaisesti. Tämän on arveltu johtuvan osittain siitä, että päätöksentekotilanteissa, joihin liittyy epävarmuutta ja potilaan kannalta mahdollisesti huonoja vaihtoehtoja, terveydenhuollon ammattilaisten on hankala luottaa laitteeseen, jonka toimintaa on vaikeaa tai mahdotonta ymmärtää [7]. Tässä mielessä päätöspuut saattavat olla helpommin hyväksyttävissä kuin muut luokittimet, kun päätöksentekijä voi itse verifioida, mihin päätöspuun tulos perustuu. Terveydenhuollossa on myös ennestään totuttu erilaisiin graafisiin päätöksenteon apuvälineisiin, jolloin päätöspuu ei välttämättä edes olisi niin erilainen kuin jo aiemmin tutuiksi tulleet menetelmät.

Päätöspuu ja sen käyttö päätöksenteon tukemisessa on helposti ymmärrettävissä. Ensinnäkin päätöspuu on rakenteeltaan nimensä mukaisesti puu ja on sellaisena visualisoitavissa. Sen käyttäminen päätöstä tehdessä liittyy polun kulkemiseen puun juuresta johonkin sen lehdistä. Puun lehdet edustavat päätöstä ja sen muut solmut ehtoja, joiden tuloksen perusteella valikoituu yksi solmun lapsipuiden juurista polun seuraavaksi askeleeksi. Näin kukin polku puun juuresta lehteen koostuu sarjasta ehtoja, joiden perusteella päädytään yhteen mahdollisista puun tuloksista eli päätöksistä.

Konkreettisena esimerkkinä päätöspuu voisi liittyä edellämainittuun kysymykseen siitä, onko päivystyspoliklinikalle saapuvalla rintakipupotilaalla suuri kuolemanriski vai ei. Tähän tarkoitukseen muodostetussa puussa juuressa oleva ehto voisi liittyä potilaan sukupuoleen: jos potilas on miespuolinen, valitaan juuren vasemmanpuoleinen lapsi, jos taas potilas on naispuolinen, valitaan juuren oikeanpuoleinen lapsi. Mikäli potilas on mies, saattaisi seuraava ehto liittyä siihen, onko potilaalla verenpainetauti ja seuraava ehto sen

jälkeen taas kolesterolitasoon. Näin potilaasta tehtävät havainnot johdattavat lopulta puun lehteen eli potilastapauksesta tehtävään luokitukseen ja sitä myötä päätökseen.

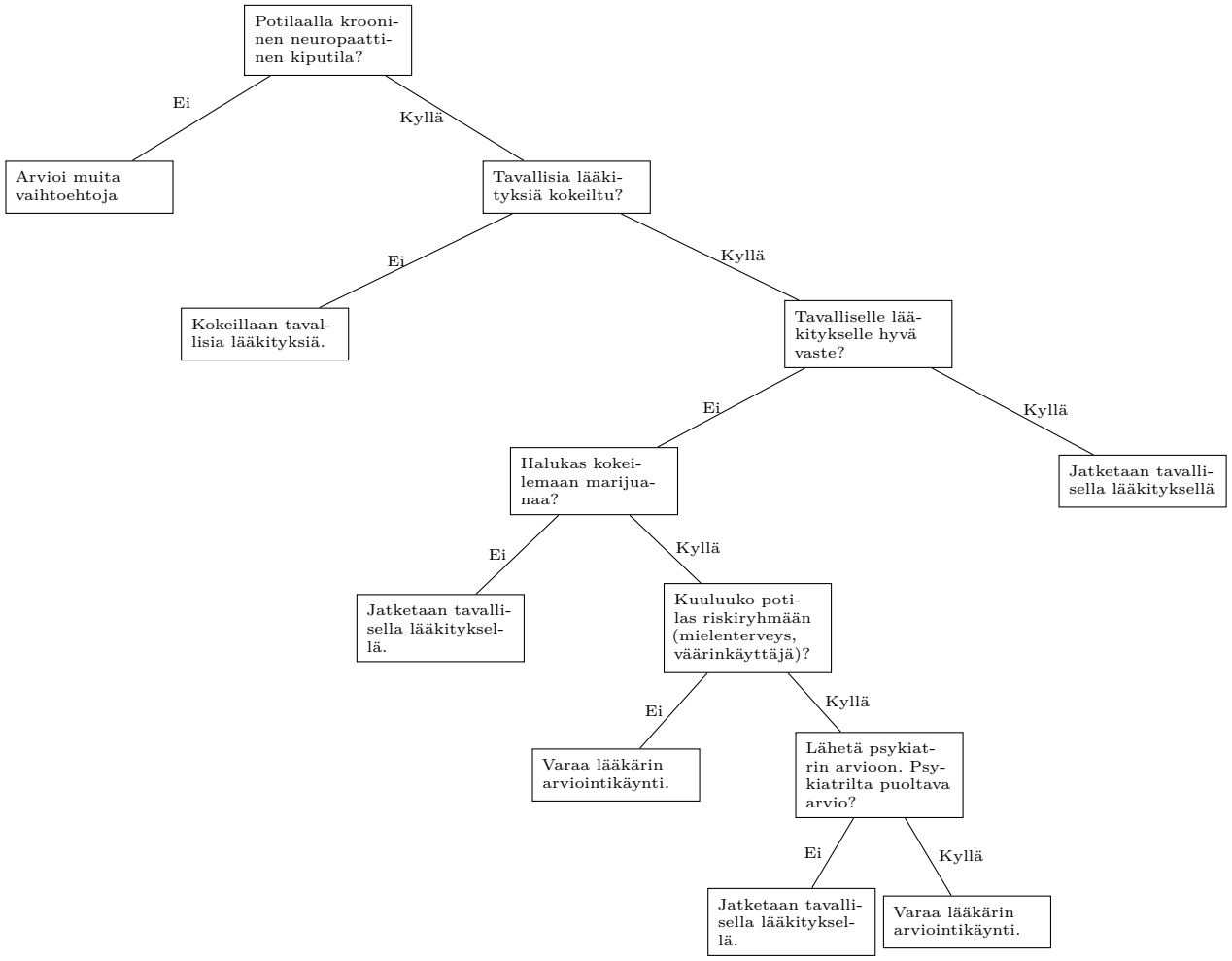
Periaatteessa päätöspuina voitaisiin pitää kaikkia sellaisia puita, joissa edellämainitulla tavalla päädytään puun lehdessä olevaan päätökseen tiettyjen ehtojen perusteella. Jäljempänä päätöspuilla tarkoitetaan kuitenkin vain sellaisia puita, jotka on muodostettu niin kutsutuilla induktioalgoritmeilla. Tällaisille puille on ominaista, että ne ovat luonteeltaan yleistämiseen kykeneviä luokittimia ja pohjautuvat induktioalgoritmille syötettyihin esimerkkitapauksiin. Induktioalgoritmeihin perehdytään tarkemmin kappaleessa 3. Yleistämisellä tarkoitetaan sitä, että päätöspuu kykenee luokittelemaan muitakin tapauksia kuin pelkästään niiden muodostamisessa käytettyjä esimerkkitapauksia.

Päätöspuiden ymmärrettävyys ja visuaalisuus ovat potentiaalisesti hyödyllisiä sivuominaisuuksia; induktioalgoritmien tavoitteena on kuitenkin vain puun mahdollisimman hyvä luokituskyky. Puun muilla ominaisuuksilla ei ole merkitystä. Päätöspuiden soveltamista terveydenhuollon päätöksentekotilanteisiin tarkastellaan kappaleessa 4. Siinä kontekstissa arvioidaan päätöspuiden meriittejä myös mainittujen sivuominaisuuksien näkökulmasta, vertailukohtana terveydenhuollossa tyypillinen, kotikutoinen, kuvassa 1 esitetty päätöspuu, jonka muodostamisessa ei ole käytetty mitään erityisempää formalismia. Tämän tyyppinen, vuokaaviomainen päätöspuu on ilmeisen käyttökelpoinen; induktioalgoritmipäätöspuiden olisi hyvä kyetä samantyyppiseen intuitiiviseen käytettävyyteen saavuttaakseen hyväksyttävyyttä terveydenhuollon ammattilaisten keskuudessa.

Päätöspuita on sovellettu tiedon louhinnassa ja lisäksi joillakin induktioalgoritmeilla voidaan muodostaa päätöspuiden lisäksi regressiopuita. Näitä aihepiirejä ei käsitellä tämän enempää.

3 Päätöspuiden induktioalgoritmit

Induktioalgoritmien tehtävä on muodostaa yleistämiseen kykenevä päätöspuu pelkästään niille syötettyjen, ennalta luokiteltujen esimerkkien perusteella.



Kuva 1: Päätöspuu liittyen kysymykseen, hyötyykö potilas lääkärin arviointikäynnistä.

Kaikki esimerkit oikein luokittelevan päätöspuun luominen on triviaali tehtävä, lisäksi samasta joukosta esimerkkejä voidaan muodostaa monenlaisia puita. Haaste onkin muodostaa sellainen puu, joka paitsi luokittelee esimerkit niin myös mitkä tahansa sille esitetyt tapaukset oikein. Kullakin induktioalgoritmilla on omanlaisensa strategia suoriutua tästä. Tunnetuin algoritmeista lienee Quinlanin ID3 (Iterative Dichotomiser 3) [8], joka käsitellään perusteellisesti jäljempänä. Muita tunnettuja induktioalgoritmeja ovat niin ikään Quinlanin C4.5 [9] ja C5.0 [10] sekä Breimanin ym. kehittämä joukko algoritmeja, joita kutsutaan yhteisellä nimellä CART (engl. classification and regression tree) [2].

3.1 Induktioalgoritmien yhteiset piirteet

Riippumatta käytetystä algoritmista päätöspuun muodostaminen on aina samanlainen tapahtuma sikäli, että algoritmin syötteenä on pelkästään joukko esimerkkitapauksia, joiden luokka on tiedossa, ja tuloksena yleistämiseen pyrkivä luokitin. Kaikkien algoritmien yhteisenä haasteena on esimerkkitapausten ominaisuuksien perusteella päätellä, mihin luokkaan muutkin mahdollisesti puulle myöhemmin esiteltävät tapaukset kuuluvat. Tämäntyyppistä päättelyä, jossa esimerkkitapausten perusteella tehdään oletuksia muistakin mahdollisista tapauksista, kutsutaan induktiiviseksi päättelyksi. Induktiivinen päättely on luonteeltaan epävarmaa, mistä seuraa, että myös päätöspuiden aikaansaama luokitus on epävarmaa. Induktioalgoritmien muodostamien päätöspuiden tulosta täytyykin pitää enemmänkin olettamuksena tai hypoteesina oikeasta luokasta kuin varmana luokituksena.

Matemaattisessa mielessä päätöspuu toimii kuvauksena eli funktiona kaikkien mahdollisten tapausten joukosta kaikkien mahdollisten luokkien joukkoon. Kutsuttakoon edellistä joukkoa joukoksi S ja jälkimmäistä joukoksi L. Oletetaan, että jokaisella joukon S alkiolla on luokka, joka on joukon L alkio. Tällöin jokin tuntematon funktio f kykenee liittämään jokaisen joukon S alkion oikeaan joukon L alkioon eli täydelliseen luokitteluun. Funktion f päättely ei yleensä ole mahdollista vaan induktioalgoritmin tuloksena on hypoteesi h, joka on approksimaatio funktiosta f. Verrattuna muihin keinoi-

hin approksimoida f päätöspuuinduktioalgoritmin erikoisuus on päätyminen luokittimeen, joka on rakenteeltaan puu.

Induktioalgoritmit ovat esimerkki ohjatusta oppimisesta; algoritmi pyrkii oppimaan edellisessä kappaleessa mainitun funktion f sille annettujen ennalta luokiteltujen esimerkkien perusteella. Kuten ohjatussa oppimisessa yleensäkin, myös induktioalgoritmien kohdalla käytettävissä olevat esimerkkitapaukset jaetaan kahteen osajoukoon, opetusjoukkoon ja testijoukkoon. Opetusjoukon tapauksia käytetään päätöspuun muodostamiseen ja testijoukon tapauksia kyseisen puun luokituskyvyn selvittämiseen: mitä suurempi osa testijoukon tapauksista luokitellaan oikein, sitä parempi päätöspuu on tehtävässään. Eri induktioalgoritmit päätyvät erilaisiin puihin vaikka hyödyntäisivät samaa opetusjoukkoa - vertaamalla puiden luokittelukykyä testijoukon alkioilla voidaan löytää puista se, joka kussakin päätöksentekotilanteessa toimii parhaiten.

Kukin opetusjoukon tapaus esitellään induktioalgoritmeille kokoelmana attribuutteja, joilla kullakin on arvo. Eri induktioalgorimit eroavat toisistaan siinä, minkälaisia arvoja attribuutit voivat saada: jotkut pystyvät käsittelemään reaaliarvoisia attribuutteja, toiset kokonaislukuarvoisia. Kuitenkin esimerkiksi ID3 pystyy käsittelemään vain sellaisia attribuutteja, joiden arvot tulevat ennalta määritellystä joukosta. Esimerkki tällaisesta attribuutista voisi olla rintakivun kovuus: ID3:n tapauksessa kovuus voisi saada vain ennalta määriteltyjä arvoja, kuten “kova”, “keskikova” ja “lievä”, kehittyneemmässä algoritmissa arvot voisivat olla jatkuvia esimerkiksi väliltä 0 - 10. Kukin tapaus kuvaillaan samalla joukolla attribuutteja, kullakin tapauksella on näille omat arvonsa. Tarvittavien attribuuttien määrä riippuu käsillä olevasta luokitteluongelmasta: joskus luokittelun onnistuminen saattaa vaatia tapausten kuvailua usealla kymmenellä attribuutilla, toisinaan saatetaan selvitä huomattavasti pienemmällä määrällä. Voi käydä niinkin, että luokittelu ei onnistu laisinkaan käytettävissä olevilla attribuuteilla.

Valmiissa päätöspuussa opetusjoukon tapausten attribuutteja vastaavat puun ehtosolmujen ehdot ja solmun lapsia attribuutin mahdolliset arvot. Mikäli rintakipuesimerkin päätöspuu olisi olemassa, siinä saattaisi olla - mutta ei tarvitse olla - ehtosolmu, jonka ehtona on nimenomaan rintakivun kovuus. Päätöspuuta käytettäessä rintakipupotilaan rintakivun kovuus pitäisi

olla tiedossa ja silloin päätöspuun ehtoon mahdollisuus vastata. Eteneminen ehtosolmusta eteenpäin määräytyisi juuri tämän rintakipupotilaan rintakivun kovuuden määrittelevän attribuutin arvon perusteella, jolloin päätöspuun seuraavaksi askeleeksi määräytyisi se ehtosolmun lapsi, jota vastaa tapauksen arvo.

3.2 ID3

Verrattuna sitä edeltäviin induktioalgoritmeihin ID3:ssa oivalluksena on ollut soveltaa informaatioteoriaa puun ehtosolmujen attribuuttien valinnassa. Algoritmin tiettyjen puutteiden vuoksi siitä on kehitty paranneltuja versioita (C4.5 ja C5.0). ID3 pystyy luokittelemaan vain kahteen erilliseen luokkaan (+ ja -), lisäksi attribuuttien arvot voivat olla peräisin vain ennalta määritellystä joukosta vaihtoehtoja. Se ei tue jatkuva- eikä kokonaislukuarvoisia attribuutteja lainkaan.

ID3 on rekursiivinen algoritmi, joka aloittaa puun muodostamisen sen juuresta. Juuren ollessa kyseessä se pyrkii valikoimaan koko opetusjoukon S alkioiden attribuuteista sen, jonka suhteen opetusjoukon jakaantumisesta osajoukkoihin seuraa paras mahdollinen informaatiolisä $IG(S)$ (engl. information gain). Algoritmi laskee informaatiolisän kaikille attribuuteille: juuren attribuutiksi valikoituu se, jonka informaatiolisä on suurin. Attribuutin valinnan seurauksena opetusjoukko jakaantuu osajoukkoihin attribuutin mahdollisten arvojen perusteella ja vastaavasti juuri saa niin monta lapsisolmua kuin attribuutilla on mahdollisia arvoja. Valittu attribuutti vastaa päätöspuun ehtosolmussa olevaa ehtoa. Jaon jälkeen algoritmi pureutuu jäljellä oleviin opetusjoukon osajoukkoihin samalla rekursiivisella periaatteella kuin koko opetusjoukon ollessa kyseessä, mutta jo valikoidut attribuutit jätetään huomioimatta suurinta informaatiolisää laskettaessa eikä niitä enää uudestaan valita ehtoattribuuteiksi.

Algoritmin rekursio päättyy seuraavissa tilanteissa (pysähtymisehdot):

1. Kaikki jäljellä oleva opetusjoukon osajoukon alkiot kuuluvat samaan luokkaan (+ tai -): muodostetaan puun lehti, jonka luokaksi asetetaan kyseisten alkioiden luokka.

2. Kaikki mahdolliset attribuutit on jo kulutettu loppuun mutta osajoukon alkiot eivät silti kuulu samaan luokkaan (+ tai -): muodostetaan puun lehti, jonka luokaksi valikoituu se luokka, johon suurempi osa osajoukon alkioista kuuluu

3. osajoukko on tyhjä: muodostetaan lehti, jonka luokaksi valitaan lehden vanhemman osajoukon alkioiden yleisin luokka.

ID3-algoritmia käytettäessä käytettävissä olevista esimerkkitapauksista erotetaan satunnaisesti tietty määrä tapauksia; muodostettua joukkoa kutsutaan ikkunaksi. Jäljelle jääneet tapaukset muodostavat testijoukon. ID3-algoritmi muodostaa ikkunan alkioiden perusteella päätöspuun, joka aina luokittelee ikkunan tapaukset oikein. Tämän jälkeen loput esimerkkitapaukset eli testijoukko ajetaan päätöspuun läpi ja mikäli ilmenee, että yksikin testijoukon alkio luokittuu väärin, valitaan satunnainen määrä näitä väärin luokitettuja tapauksia ja lisätään ne ikkunaan. Iteraatioita jatketaan, kunnes kaikki testijoukonkin alkiot luokittuvat oikein. Quinlanin artikkelin mukaan [8] tällä tavoin toimien on mahdollista nopeuttaa algoritmin toimintaa. Periaatteessa tällä strategialla toimien voisi käydä niin, ettei ID3 päätyisi lainkaan sopivaan päätöspuuhun; saman artikkelin mukaan silloisilla ID3:lle syötetyillä joukoilla ei olisi kertaakaan käynyt niin.

Pseudokoodiesitys ID3-algoritmista on esitettu kuvassa 2.

3.3 ID3 ja päätöspuun muodostaminen esimerkkidatan pohjalta

ID3-algoritmin toiminnan havainnollistamiseksi seuraavassa muodostetaan päätöspuu käyttäen taulukossa 1 olevaa esimerkkidataa. Kyseinen esimerkkidata on kuvitteellinen ja on terveydenhuollon kontekstiin mukailtu versio Quinlanin artikkelissa [8] olevasta samantyyppisestä esimerkistä.

Esimerkkidatassa on neljäntoista päivystyspoliklinikalla hoidetun nilkkavammapotilaan tietoja sekä tieto siitä, katsoiko potilaan päivystyspoliklinikalla hoitanut lääkäri tarpeelliseksi pyytää potilaan nilkasta röntgentutkimusta vai ei (viimeisen sarakkeen P tai N). Muihin sarakkeisiin on koottu potilaiden

```
function ID3(Esimerkit, Attribuutit, Luokat)
    Muodosta solmu t
    if Esimerkit-joukon kaikkien alkioiden luokka on + then
        Anna solmulle t leima +
        return t
    end if
    if Esimerkit-joukon kaikkien alkioiden luokka on - then
        Anna solmulle t leima -
        return t
    end if
    Leimaa solmu t Luokat-joukon yleisimmällä arvolla Esimerkit-joukossa
    if Attribuutit-joukko on tyhjä then
        return t
    end if
    Valitse A*:ksi parhaiten luokitteleva Attribuutti Esimerkit-joukossa
    Aseta A* t:n ehtoparametriksi
    for kaikille A*:n mahdollisille arvoille a do
        Muodosta t:lle uusi lapsisolmu jota vastaa ehto A* = a
        Muodosta joukko EsimerkkiA jossa A* = a
        if EsimerkkiA on tyhjä then
            Lisää lehtisolmu ja leimaa se Luokat-joukon yleisimmällä arvolla
joukossa Esimerkit
        else
            Lisää alipuu ID3(EsimerkkiA, Attribuutit \A*, Luokat)
        end if
    end for
    return t
end function
```

Kuva 2: ID3-algoritmin pseudokoodiesitys

tietoja: miltä nilkka on näyttänyt, mikä vammamekanismi on ollut, mikä kivun kovuus on ollut ja onko potilas kyennyt varaamaan (kävelemään kipeällä nilkalla) vai ei.

Taulukkoa 1 tarkastelemalla herää ajatus, pystyttäisiinkö siinä olevan datan, eli potilaiden ominaisuuksien ja luokituksen (P tai N) perusteella muodostamaan päätöspuu, jolla voitaisiin ennustaa hoitavan lääkärin päätös tilata potilaan nilkasta röntgenkuva? Tällainen päätöksenteon apukeino voisi päivystyspoliklinikalla olla hyödyllinen: päätös kuvaamisen tarpeesta voitaisiin tehdä jo potilaan saapuessa päivystyspoliklinikalle hoitajan tekemän lyhyen arvion perusteella, jolloin potilas voisi käydä nilkan kuvauksessa ennen lääkärin vastaanotolle saapumista. Tällöin lääkäri voisi hoitaa potilaan yhden kontaktin perusteella: vastaanotolla olisi potilastapauksen ratkaisemiseksi kaikki tarpeellinen tieto valmiina. Ilman kuvaamistarpeen ennakointia tarvittaisiin kaksi lääkärin ja potilaan kohtaamista: ensimmäisessä todetaan, ettei potilastapausta voida ratkaista ilman kuvaamista ja toinen potilastapauksen lopullista hoitoa varten. Yhden kontakin malli johtaisi potilaan nopeampaan hoitoon ja olisi rajallisten resurssien tehokasta käyttöä.

Taulukosta nähdään, että kullakin potilastapauksella on neljä attribuuttia ja taulukosta voidaan katsoa kunkin potilaan saamat attribuuttien arvot. Potilaan 2 attribuutit ja niiden arvot on esitetty taulukossa 2. Taulukosta 1 nähdään myös, että kyseisen potilastapauksen kohdalla ei päädytty kuvaukseen (luokka N).

Kyseinen data sopii ID3:n hyödynnettäviksi: luokitus on jompikumpi kahdesta luokasta (N tai P), luokat ovat erilliset (potilastapaus ei voi kuulua molempiin luokkiin N ja P samanaikaisesti) ja potilaiden attribuuttien arvot kuuluvat ennaltamääriteltyihin joukkoihin, jotka on esitetty taulukossa 3.

Joukon S entropia $H(S)$ on kyseiseen joukkoon sisältyvän epävarmuuden mitta. Se lasketaan kaavalla

$$H(S) = -\sum_{x \in X} p(x) \log_2 p(x),$$

jossa X on joukon S alkioiden mahdollisten luokkien joukko ja $p(x)$ niiden alkioiden osuus joukossa S joiden luokka on $x \in X$. Entropian yksikkönä on

tapana käyttää bittiä.

ID3-algoritmissa joukko S halutaan jakaa sen attribuutin A suhteen osajoukkoihin $T_1, T_2, ..., T_t \subset T = S$, joista seuraa suurin informaatiolisä $IG(A, S)$. Edellä joukot $T_1, T_2, ..., T_t$ ovat erillisiä ja $T_1 \cup T_2 \cup ... \cup T_t = \cup_{t=1}^{n} T_t = S$. Attribuutilla A on n mahdollista arvoa. Informaatiolisä on joukon S entropian ja osajoukkoihin jaosta seuraavan entropian erotus, joka lasketaan kaavalla

$$IG(A, S) = H(S) - \sum_{t=1}^{n} p(T_t)H(T_t),$$

jossa $p(T_t)$ on joukon T_t alkioiden osuus kaikista joukon S alkioista ja $H(T_t)$ joukon T_t entropia.

Lasketaan nilkkavammapotilasdatan entropia. Nilkkavammapotilaat ovat jakautuneet kahteen luokkaan N ja P. Luokkaan N kuuluu 5 tapausta ja luokkaan P kuuluu 9 tapausta. Yhteensä tapauksia on siis neljätoista. Näiden tietojen perusteella nilkkavammapotilasjoukon S entropia on

$$H(S) = -\left(\frac{5}{14}\log_2\left(\frac{5}{14}\right) + \frac{9}{14}\log_2\left(\frac{9}{14}\right)\right) \approx 0.940 \text{ bittiä.}$$

Lasketaan seuraavaksi, minkä attribuutin perusteella koko nilkkavammapotilaiden joukko kannattaa jakaa osajoukkoihin, eli minkä attribuutin suhteen jakamisesta seuraa suurin informaatiolisä. Aloitetaan attribuutista *Ulkonäkö*, joka taulukon 3 mukaisesti voi saada arvot *Normaali, Turvonnut* tai *Virheasento*. Informaatiolisän $IG(A_{Ulkonäkö}, S)$ laskemiseksi tarvitaan $p(T_{Normaali})$, $p(T_{Turvonnut})$ ja $p(T_{Virheasento})$, eli attribuutin *Ulkonäkö* saamien arvojen osuus kaikista joukon S alkioista. Taulukosta 1 voidaan laskea, että

$$\begin{aligned} p(T_{Normaali}) &= \frac{5}{14} \\ p(T_{Turvonnut}) &= \frac{5}{14} \\ p(T_{Virheasento}) &= \frac{4}{14}. \end{aligned}$$

Lisäksi tarvitaan joukkojen $T_{Normaali}$, $T_{Turvonnut}$ ja $T_{Virheasento}$ entropiat

$H(T_{Normaali})$, $H(T_{Turvonnut})$ ja $H(T_{Virheasento})$. Kaava on luonnollisesti sama kuin edellä, jossa laskettiin koko joukon S entropia $H(S)$, mutta osajoukkojen $T_{Normaali}$, $T_{Turvonnut}$ ja $T_{Virheasento}$ entropioita laskettaessa $p(x)$ tarkoittaa luokkien N ja P osuutta kussakin osajoukossa $T_{Normaali}$, $T_{Turvonnut}$ ja $T_{Virheasento}$. Entropiat ovat

$$\begin{aligned}
H(T_{Normaali}) &= -\left(\frac{3}{5}\log_2\left(\frac{3}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right) \approx 0.971 \text{ bittiä,} \\
H(T_{Turvonnut}) &= -\left(\frac{3}{5}\log_2\left(\frac{3}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right) \approx 0.971 \text{ bittiä ja} \\
H(T_{Virheasento}) &= -\left(\frac{4}{4}\log_2\left(\frac{4}{4}\right) + \frac{0}{4}\log_2\left(\frac{0}{4}\right)\right) = 0 \text{ bittiä.}
\end{aligned}$$

Viimeistä riviä laskettaessa törmätään lausekkeeseen $0 \log_2 0$, jolla ei ole määriteltyä arvoa. Informaatioteoriassa tämä saa yleisen käytännön mukaisesti arvon 0. Se, että tästä seuraa $H(T_{Virheasento}) = 0$, on helposti ymmärrettävissä, koska kyseisessä tapauksessa kaikki alkiot kuuluvat samaan luokkaan P ja näin entropia määritelmän mukaan on 0.

Nyt on kaikki tarvittava informaatiolisän $IG(A_{Ulkonäkö}, S)$ laskemiseksi:

$$\begin{aligned}
IG(A_{Ulkonäkö}, S) &= H(S) - \sum_{t=1}^{n} p(T_t)H(T_t) \\
&\approx 0.940 - \left(\frac{5}{14} \cdot 0.971 + \frac{5}{14} \cdot 0.971 + \frac{4}{14} \cdot 0\right) \\
&\approx 0.246.
\end{aligned}$$

Samalla tavalla on mahdollista laskea myös muidenkin attribuuttien suhteen jakamisesta seuraavat informaatiolisät:

$$\begin{aligned}
IG(A_{Vammamekanismi}, S) &\approx 0.029, \\
IG(A_{Kipu}, S) &\approx 0.151 \text{ ja} \\
IG(A_{Varaaminen}, S) &\approx 0.048.
\end{aligned}$$

Koska informaatiolisä $IG(A_{Ulkonäkö}, S)$ on suurin, valitsee ID3 ulkonäköattribuutin päätöspuun juuren ehtosolmun attribuutiksi. Tässä vaiheessa aikaansaatu päätöspuu on esitetty kuvassa 3. Juuren attribuutin valitsemisen jälkeen

Nro	Ulkonäkö	Vamma-mekanis-mi	Kipu	Varaa-minen	Kuvataan
1	Normaali	Nyrjähdys	Ei-Kova	Ei	N
2	Normaali	Nyrjähdys	Ei-kova	Kyllä	N
3	Virheasento	Nyrjähdys	Ei-kova	Ei	P
4	Turvonnut	Putoaminen	Ei-kova	Ei	P
5	Turvonnut	Muu	Kova	Ei	P
6	Turvonnut	Muu	Kova	Kyllä	N
7	Virheasento	Muu	Kova	Kyllä	P
8	Normaali	Putoaminen	Ei-kova	Ei	N
9	Normaali	Muu	Kova	Ei	P
10	Turvonnut	Putoaminen	Kova	Ei	P
11	Normaali	Putoaminen	Kova	Kyllä	P
12	Virheasento	Putoaminen	Ei-kova	Kyllä	P
13	Virheasento	Nyrjähdys	Kova	Ei	P
14	Turvonnut	Putoaminen	Ei-kova	Kyllä	N

Taulukko 1: Esimerkkidata, joka koostu neljäntoista nilkkavammapotilaan tiedoista.

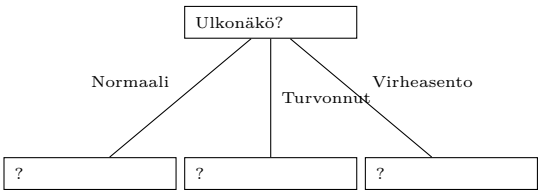
ID3 siirtyy rekursiivisesti rakentamaan juuren lapsipuita. Lapsipuita on kolme, koska juurisolmun ehtoattribuutilla on kolme mahdollista arvoa. Lapsipuiden juuria valittaessa ID3 ei enää ota huomioon koko opetusjoukkoa S vaan sen erillisiä osajoukkoja $T_{Normaali}$, $T_{Turvonnut}$ ja $T_{Virheasento}$, yksi kutakin lapsipuuta kohden. Tällä tavoin rekursiivisesti edeten ID3 rakentaa päätöspuun pysähtyen kunkin lapsipuun kohdalla kappaleessa 3.2 mainittuihin pysähtymisehtoihin. Joukot $T_{Normaali}$, $T_{Turvonnut}$ ja $T_{Virheasento}$ on esitetty taulukoissa 4, 5 ja 6. Huomionarvoista on, että kun ulkonäköattribuutti $A_{Ulkonäkö}$ on käsitelty ja kiinnitetty puun juureen, sitä ei enää uudestaan huomioida juuren lapsia rakennettaessa. Yleensäkin attribuutit voivat esiintyä puun poluissa vain kertaalleen - toisaalta sama attribuutti voi esiintyä useassa polussa eri puolilla päätöspuuta. ID3:n aikaansaama päätöspuu on esitetty kuvassa XXX (tekemättä).

Attribuutti	**Attribuutin arvo**
Ulkonäkö	Normaali
Vammamekanismi	Nyrjähdys
Kipu	Ei-kova
Varaaminen	Kyllä

Taulukko 2: Nilkkavammapotilaan numero 2 attribuuttien arvot

Attribuutti	**Attribuutin mahdolliset arvot**
Ulkonäkö	Normaali, Turvonnut, Virheasento
Vammamekanismi	Nyrjähdys, Putoaminen, Muu
Kipu	Kova, Ei-kova
Varaaminen	Kyllä, Ei

Taulukko 3: Nilkkavammapotilastapausten mahdolliset attribuutit ja niiden arvot.



Kuva 3: ID3 on valinnut juurisolmun attribuutin. Seuraavaksi algoritmi alkaa työstämään juuren lapsia.

Nro	**Vammamekanismi**	**Kipu**	**Varaaminen**	**Kuvataan**
1	Nyrjähdys	Ei-kova	Ei	N
2	Nyrjähdys	Ei-kova	Kyllä	N
3	Putoaminen	Ei-kova	Ei	N
4	Muu	Kova	Ei	P
5	Putoaminen	Kova	Kyllä	P

Taulukko 4: Opetusjoukon osajoukko $T_{Normaali}$.

Nro	**Vammamekanismi**	**Kipu**	**Varaaminen**	**Kuvataan**
1	Putoaminen	Ei-kova	Ei	P
2	Muu	Kova	Ei	P
3	Muu	Kova	Kyllä	N
4	Putoaminen	Kova	Ei	P
5	Putoaminen	Ei-kova	Kyllä	N

Taulukko 5: Opetusjoukon osajoukko $T_{Turvonnnut}$.

Nro	Vammamekanismi	Kipu	Varaaminen	Kuvataan
1	Nyrjähdys	Ei-kova	Ei	P
2	Muu	Kova	Kyllä	P
3	Putoaminen	Ei-kova	Kyllä	P
4	Nyrjähdys	Kova	Ei	P

Taulukko 6: Opetusjoukon osajoukko $T_{Virheasento}$.

3.4 CART

CART [2] on ID3:n tavoin keskeinen päätöspuuinduktioalgoritmi. Sen toiminta muistuttaa ID3-algoritmia sikäli, että samalla tavoin päätöspuun muodostaminen aloitetaan puun juuresta ja rekursiivisesti jatkuu kunnes puu on muodostettu. Keskeinen ongelma puuta muodostettaessa sekä CART:ssa että ID3:ssa on päättää puun ehtosolmujen valinnasta, jonka seurauksena puu saa muotonsa. Tämäntyyppisellä strategialla toimivia algoritmeja on kutsuttu yhteisellä nimikkeellä TDIDT, joka on kirjainlyhenne englanninkielisistä nimestä “top-down induction of decision trees” [8].

CART:n toiminta poikkeaa ID3:n toiminnasta monin tavoin ja on sitä huomattavasti monipuolisempi algoritmi. Se muodostaa aina binääripuita eikä ehtosolmujen attribuuttien valinnassa sovelleta yleensä informaatioteoriaa vaan kussakin solmussa pyritään aikaansaamaan sellainen joukon jako, jonka seurauksena muodostuvat kaksi osajoukkoa ovat alkuperäistä joukkoa “puhtaampia”. Käytettävä epäpuhtauden mitta riippuu tilanteesta: ID3:n tavoin luokittimena käytettäessä epäpuhtauden mittana voidaan käyttää ns. gini-indeksiä tai ns. twoing-indeksiä. Molemmissa periaate on sama: joukon epäpuhtaus vähenee kun sen homogeenisuus kasvaa. Epäpuhtaus saa minimiarvon, kun solmun kaikki tapaukset kuuluvat samaan luokkaan ja maksimiarvon, kun solmun tapaukset jakautuvat kaikkiin luokkiin tasaisesti.

CART on ID3:a joustavampi attribuuttien ominaisuuksien suhteen. Siinä missä ID3 sallii attribuuttien arvoiksi vain ennalta määriteltyjä, CART osaa ottaa huomioon myös näiden arvojen mahdollisen ordinaalisuuden ja toisaalta jatkuva-arvoisetkin attribuutit sallitaan. Koska CART muodostaa binääripuita, kussakin ehtosolmussa otetaan huomioon vain kaksi attribuuttien mahdollista arvoa. Toisaalta attribuutilla voi olla enemmän kuin kaksi

mahdollista arvoa, joten sama attribuutti saattaa esiintyä puun poluilla useaan kertaan - ID3:ssa attribuutin kaikki arvot käsitellään samassa solmussa ja solmu saa aina yhtä monta lasta kuin attribuutilla on mahdollisia arvoja.

Luokittelun lisäksi CART kykenee nimensä mukaisesti muodostamaan myös regressiopuita, joissa pyrkimyksenä ei ole luokittelu kategorisiin luokkiin vaan estimoida jatkuva-arvoisen funktion riippuvuus attribuuttien arvoista.

CART-algoritmin kokonaisvaltainen käsittely jää tämän työn ulkopuolelle. Esitellään kuitenkin gini-indeksi ja "epäpuhtauden" merkitys ehtosolmun valinnassa, jotta ero ID3:n informaatioteoriaan perustuvaan menetelmään olisi arvioitavissa tämän työn puitteissa.

Gini-indeksi joukolle T, jossa on n luokkaa, määritellään seuraavasti:

$$gini(T) = 1 - \sum_{j=1}^{n} p_j^2,$$

jossa $p_j(t) = \frac{n_j(t)}{n(t)}$ on luokan j osuus solmun t alkioista. Jaetaan joukko T kahteen osajoukkoon T_1 ja T_2, joiden alkioiden lukumäärät ovat m_1 ja m_2. Jaon gini-indeksi määritellään seuraavasti:

$$gini_{jako}(T) = \frac{m_1}{m_1 + m_2} gini(T_1) + \frac{m_2}{m_1 + m_2} gini(T_2).$$

Jos solmussa t tehtäisiin jako s seuraisi tästä gini-indexin koheneminen (epäpuhtuden vähentyminen), joka voidaan laskea koko joukon T gini-indexin ja jaosta seuraavan gini-indexin erotuksena:

$$\Delta(s, t) = gini(T) - gini_{jako}(T).$$

Jaolla s tarkoitetaan yhtä niistä mahdollisista jaoista, joita voidaan tehdä solmussa t. CART-algoritmi valitsee solmun t jaoksi sen jaon s, josta seuraa suurin $\Delta(s, t)$. Valinnan jälkeen ID3-algoritmin tavoin CART jatkaa tämän jälkeen rekursiivisesti solmun t lapsipuiden solmujen järjestyksen selvittämistä, josta seuraa lopulta kokonainen päätöspuu.

4 Päätöspuiden sovelluksia terveydenhuollossa

Esitellään kolme kirjallisuudessa kuvattua päätöspuiden sovellusta terveydenhuollon kontekstissa.

4.1 Aivoverenkiertopotilaiden ennustaminen

Duen-Yian ym. [3] vertasivat kolmea eri menetelmää aivoverisuonisairauksien ennustamiseksi potilaiden ominaisuuksien perusteella. Parhaimmaksi osoittautui päätöspuihin perustuva luokitin. Tutkimuksessa käytettiin C4.5-algoritmia, joka on Quinlanin ID3:n pohjalta kehittämä induktioalgoritmi.

Aivoverisuonisairaudet ovat joukko sairauksia, joihin liittyy paljon kuolleisuutta ja elämänlaadun heikentymistä muun muassa vammautumisen ja dementoitumisen vuoksi. Keskeinen aivoverisuonisairaus on aivoinfarkti, joka voi johtaa pahimmassa tapauksessa kuolemaan mutta lievänäkin erilaisiin neurologisiin vammoihin, kuten halvaantumiseen ja liikunta- ja puhekyvyn menettämiseen. Maailmanlaajuisesti aivoverenkierron häiriöt ovat toiseksi yleisin kuolinsyy [4]. Niiden osuus kaikista kuolemista on noin 10%. Vuoden 2002 tilastojen mukaan [6] Suomessa sairastuu vuosittain ensimmäiseen aivohalvaukseen

- 25–74-vuotiaista 252/100 000 miestä ja 137/100 000 naista,
- 75–84-vuotiaista 1 747/100 000 miestä ja 1 586/100 000 naista,
- yli 85-vuotiaista 3 013/100 000 miestä ja 3 029/100 000 naista.

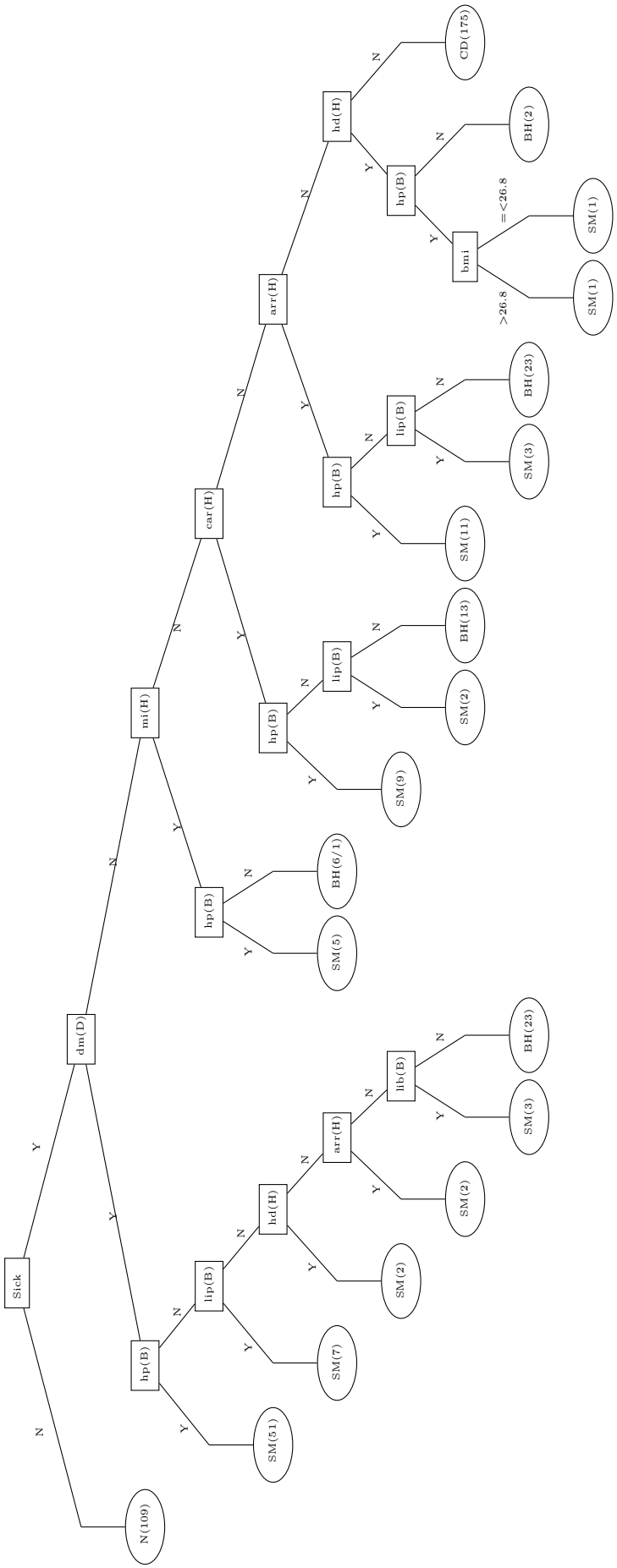
Aivoinfarktin kansantaloudellinen merkitys on valtava: elinikäisiksi terveydenhuollon kustannuksiksi Suomessa on arvioitu 80 000 euroa ja vuotuisiksi valtakunnallisiksi kustannuksiksi 1,1 miljardia euroa [5].

Aivoverenkierron häiriöiden kansantaloudellisten ja potilaille niistä koituvien haittojen vuoksi menetelmä, jolla voitaisiin ennustaa yksittäisen potilaan kohdalla riski saada kyseisiä häiriöitä, olisi luonnollisesti käyttökelpoinen. Riskipotilaisiin voitaisiin kohdistaa ennaltaehkäiseviä interventioita ja näin

Symboli	**Merkitys**
Sick	Onko sairautta?
dm(D)	Onko sokeritautia?
hp(B)	Onko verenpainetautia?
lip(B)	Onko hyperlipidemiaa?
hd(H)	Onko sepelvaltimotautia?
arr(H)	Onko rytmihäiriöitä?
mi(H)	Onko ollut sydäninfarktia?
car(H)	Onko ollut kardiogeenistä shokkia?
bmi	Painoindeksi?
N	Ei ole sairautta
SM	Aivoverenkiertohäiriö ja kaksi muuta sairautta tai sokeritauti ja yksi muu sairaus BH:n listasta
DM	Aivoverenkiertohäiriö ja sokeritauti
BH	Aivoverenkiertohäiriö ja yksi muu sairaus (verenpainetauti, reuma, hyperlipidemia, sepelvaltimotauti, rytmihäiriö, sydämen vajaatoiminta, sydäninfarkti)
CD	Vain aivoverenkierron häiriö, ei muita sairauksia

Taulukko 7: Aivoverenkiertopäätöspuun symbolien selitykset.

mahdollisesti estää näiden potilaiden kohdalla aivoverenkiertohäiriön toteutuminen. Edellä mainitussa artikkelissa [3] päädyttiin 493 potilaan tiedoista koostuvassa materiaalissa päätöspuuhun, jossa 29 attribuutin joukosta valikoituneiden yhdeksän attribuutin perusteella voitiin ennustaa potilaan kuuluminen aivoverenkiertohäiriöpotilaiden joukkoon. Päätöspuu on esitetty kuvassa 4.1. Puun lehdissä sulkeissa tapausten lukumäärät. Symbolien merkitykset on esitetty taulukossa 7.



Kuva 4: Päätöspuu aivoverenkiertohäiriön ennustamiseen.

4.2 Masennuksen ennustaminen

Batterham ym. [1]

5 Yhteenveto

Päätöspuut ovat tehokkaita luokittimia. Luokittelijoina päätöspuut ovat siitä erikoisia, että tapa, jolla ne aikaansaavat luokittelun, on intuitiivisesti suhteellisen helppo kenen tahansa ymmärtää. Puiden rakenteesta, sen solmuista, lehdistä ja poluista on parhaimmassa tapauksessa aikaansaatavissa sellaisia, että ne perusteet, jolla päätöspuu aikaansaa luokittelun, ovat myös kenen tahansa ymmärrettävissä. Tämän vuoksi on oletettavissa, että erilaisista luokittimista juuri päätöspuut saattaisivat soveltua sellaisiin tilanteisiin, jossa päätöspuuta käyttävän pitäisi kyetä varmistumaan, että sen tulos on järkeenkäypä. Etenkin terveydenhuollon päätöksentekotilaneissa, joissa päätöspuita voitaisiin soveltaa päätöksenteon tukemiseen, tällainen verifiointimahdollisuus saattaisi olla juuri se ratkaiseva tekijä, jolla terveydenhuollon ammattilaisten olisi helpompi hyväksyä koneellisia apukeinoja hankalien päätöksien tekemiseen. Esimerkkinä terveydenhuollon sovelluksista nostettiin päätöspuu, jota voidaan käyttää arvioitaessa potilaiden riskiä sairastua aivoverenkiertohäiriöihin. Puu osoittauti selkeäksi ja sen logiikka järkeenkäyväksi. Muutama lääkäri oli todennut puun alustavasti käyttökelpoiseksi, kuitenkaan puuta ei oltu kokeiltu terveydenhuollon päivittäisessä toiminnassa, mikä olisi luonnollisesti oleellista se selvittämiseksi, onko päätöspuihin perustuvilla luokittimilla ja päätöksenteon tukimenetelmillä käyttökelpoisuutta terveydenhuollon kontekstissa.

Lähteet

[1] Batterham, Philip, Christensen, Helen ja Mackinnon, Andrew: *Modifiable risk factors predicting major depressive disorder at four year follow-up: a decision tree approach.* BMC Psychiatry, 9(75), 2009. `http://www.biomedcentral.com/1471-244X/9/75`.

[2] Breiman, Leo, Friedman, Jerome, Stone, Charles ja Olshen, Richard: *Classification and regression trees.* Chapman and Hall/CRC, Monterey, CA, 1984.

[3] Duen-Yian, Yeh, Ching-Hsue, Cheng ja Yen-Wen, Chen: *A predictive model for cerebrovascular disease using data mining.* Expert systems with applications, 38:8970–8977, 2011.

[4] Lopez, Alan, Mathers, Colin, Ezzati, Majid, Jamison, Dean ja Murray, Christopher: *Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data.* The Lancet, 367:1747–1757, 2006.

[5] Meretoja, Atte, Kaste, Markku, Roine, Risto, Linna, Miika, Juntunen, Merja ja Häkkinen, Unto: *PERFECT Stroke - Aivohalvauksen hoidon aiheuttamat suorat terveydenhuollon kustannukset Suomessa 1999-2008.* Teoksessa Klavus, Jan (toimittaja): *Terveystaloustiede 2010.* THL, Terveyden ja hyvinvoinnin laitos, Helsinki, 2010.

[6] Pajunen, Pia, Pääkkönen, Rauni, Laatikainen, Tiina, Hämäläinen, Helena, Keskimäki, Ilmo, Niemi, Marja, Rintanen, Hannu ja Salomaa, Veikko: *Aivohalvausten ilmaantuvuuden ja kuolleisuuden muutokset Suomessa vuosina 1991-2002.* Suomen Lääkärilehti, 22:2437–2442, 2005.

[7] Patel, Vimla, Shortliffe, Edward, Stefanelli, Mario, Szolovits, Peter, Berthold, Michael, Bellazzi, Riccardo ja Abu-Hanna, Ameen: *The Coming of Age of Artificial Intelligence in Medicine.* Artificial intelligence in medicine, 46:5–17, 2008.

[8] Quinlan, Ross: *Induction of Decision Trees.* Machine Learning, 1:81–106, 1986.

[9] Quinlan, Ross: *C4.5: Programs for machine learning.* Morgan Kaufmann, San Mateo, 1993.

[10] Quinlan, Ross: *C5.0: An Informal Tutorial*, 2011. `http://www.rulequest.com/see5-unix.html`, vierailtu 2015-11-02 .