KNIME-ohjelmisto ja koneoppiminen

HAMK
HÄMEEN AMMATTIKORKEAKOULU
HÄME UNIVERSITY OF APPLIED SCIENCES

Ammattikorkeakoulututkinnon opinnäytetyö

Tietojenkäsittelyn koulutus, Hämeenlinnan korkeakoulukeskus
kevät, 2022

Juha Anttila

HAMK
HÄMEEN AMMATTIKORKEAKOULU
HÄME UNIVERSITY OF APPLIED SCIENCES

Tietojenkäsittelyn koulutus Tiivistelmä

Hämeenlinnan korkeakoulukeskus

Tekijä Juha Anttila Vuosi 2022

Työn nimi KNIME-ohjelmisto ja koneoppiminen

Ohjaajat Lasse Seppänen

TIIVISTELMÄ

Opinnäytetyön tarkoituksena oli tutustua KNIME-ohjelmistoon ja sen tarjoamiin koneoppimisen mahdollisuuksiin. KNIME-ohjelmistoon tutustumisessa hyödynnettiin ohjelmiston verkkosivuilla saatavilla olevia verkkokoulutuksia sekä muuta materiaalia, kuten valmiita työnkulkuja.

Opinnäytetyö on toiminnallinen työ, jossa teoria koostuu datatieteen ja koneoppimisen teoriasta sekä itse KNIME-ohjelmistoon liittyvästä teoriasta. Aineisto koostui ohjelmistolla tehdyistä työnkuluista sekä Excel-tiedostosta, johon oli kirjattu ohjelmiston koulutuksiin käytetty aika sekä muistiinpanoja koulutuksiin liittyen. KNIME-ohjelmiston koneoppimisen mahdollisuuksiin tutustuttiin numeraalisen ja kuvallisen datan kautta. Käytössä oli kaikkiaan kolme erilaista koneoppimisen työnkulkua, joihin tutustuttiin tarkemmin. Myös eri koneoppimisen mallien toimivuutta vertailtiin.

Opinnäytetyön tuloksena saatiin näkemys siitä, millaisen kurssikokonaisuuden KNIME-ohjelmiston ja koneoppimisen teemojen ympärille voi rakentaa. Opintokokonaisuus rakentuu ohjelmiston tarjoamien kahden kurssin kokonaisuudesta sekä koneoppimiseen painottuvaan oppimistehtävään. Opinnäytetyössä havaittiin, että KNIME-ohjelmiston mahdollistama kuvien muokkaus voi parantaa koneoppivan mallin toimintaa. Sen sijaan datan pyöristäminen heikentää koneoppivan mallin tarkkuutta. Myös käytettävän koneoppimisen menetelmän valinnalla on vaikutusta koneoppivan mallin tarkkuuteen.

Avainsanat Koneoppiminen, KNIME-ohjelmisto, data-analyysi

Sivut 53 sivua ja liitteitä 1 sivua

HAMK
HÄMEEN AMMATTIKORKEAKOULU
HÄME UNIVERSITY OF APPLIED SCIENCES

Degree Programme in Business Information Technology
Hämeenlinna University Centre

Abstract

Author Juha Anttila
Year 2022
Subject KNIME software and machine learning
Supervisors Lasse Seppänen

ABSTRACT

The aim of this thesis was to examine KNIME software and its machine learning opportunities. KNIME software's online learning platform and other material such as finalized workflows were used to study the software.

The thesis is a functional research. The theoretical background deals with data science, machine learning and the theory of the KNIME software. The material consists of workflows made with the software and an Excel file. The Excel file contains notes about online courses and times used to accomplish different online courses for KNIME software. KNIME software's machine learning opportunities were examined by using numerical and graphical data. In total, three different kinds of machine learning workflows were discussed in detail. Different machine learning models were also compared.

The project resulted in an outlook on what type of course could be built around KNIME software and machine learning. The course is based on two KNIME software's online courses and a learning assignment which focuses on machine learning. Other results prove that KNIME software's image editing nodes can improve the results of machine learning model. With numerical data, rounding the results for machine learning model would degrade the model's accuracy. The machine learning model used in a workflow does also affect the model's accuracy.

Keywords Machine learning, KNIME software, data-analysis
Pages 53 pages and appendices 1 page

Sanasto

Big data	Digitaalista, automaattisesti kerättyä ja kooltaan laajaa dataa.
Cohen's kappa	Luku kuvaa mallin tarkkuutta -1 ja 1 välillä, jos luku on negatiivinen, satunnainen arvonta on tarkempi
CRISP-DM	Cross Industry Standard Process for Data Mining, laaja teollisuuden tiedonlouhintaprosessi
IoT	Internet of Things, esineiden internet
KNIME-ohjelmisto	Avoimen lähdekoodin ilmainen data-analytiikan, datan käsittelyn ja - visualisoinnin monipuolinen työkalu
Koneoppiminen	Itsenäisesti mallin mukaisesti oppiva kone, jonka oppimisen malli luodaan tuloksen ja datan perusteella. Malli kehittyy jokaisen onnistumisen ja epäonnistumisen myötä
Label	Hahmotusta selkeyttävä leima esimerkiksi kuvassa tai osana tietokokonaisuutta
Noodi	Yksittäinen työnkulun osa, esimerkiksi CSV-Reader
R2-luku	0 ja 1 väliin sijoittuva luku, kertoo kuinka hyvin muut tekijät selittävät variaatiota
String	Merkkijono
Workflow	Työnkulku eli kaikkien noodien muodostama kokonaisuus

Sisällys

Kuvat ja taulukot

Liitteet

1 Johdanto

Opinnäytetyössäni käsiteltävä KNIME-ohjelmisto ja sen mahdollistamat koneoppimisen menetelmät sekä data-analytiikka ovat ajankohtaisia ja kiinnostuvia useilla eri liiketoiminnan aloilla. Toimeksianantajana työssä toimii Hämeen Ammattikorkeakoulu ja toimeksiantajalle annetaan opinnäytetyön myötä lisäksi esitys siitä, miten KNIME-ohjelmiston tarjoamia verkkokursseja ja muuta materiaalia voitaisiin hyödyntää tulevaisuudessa kolme opintopisteen kokonaisuudessa.

Tutkimuksen tavoitteena on tutustua KNIME-ohjelmistoon, joka oli aloitushetkellä vielä opinnäytetyön tekijälle vain pintapuolisesti tuttu. Lisäksi tavoitteena on arvioida mahdollisia vastaan tulevia teknisiä ongelmatilanteita sekä mitata ongelmatilanteiden ratkaisuun kulunutta aikaa. Myös KNIME-ohjelmiston verkossa tarjoamia koulutuksia arvioidaan sisällön ja ajankäytön näkökulmista.

Koneoppiminen ja koneellinen datan analysointi ovat nyt ja todennäköisesti myös tulevaisuudessa kasvavia tietojenkäsittelyn osa-alueita. Lisäksi arvioin itse niiden käytön lisääntyvän erityisesti yhä tehokkaampien tietokoneiden laskentatehon ja rutiininomaisten työtehtävien automatisoinnin ansiosta. Käytettävät ohjelmistot kehittyvät koko ajan ja myös uusia tulee markkinoille. KNIME-ohjelmisto on yksi merkittävimmistä data-analytiikan työkaluista. Muita vastaavia data-analyytikan ja koneoppimisen työkaluja tarjoavia ohjelmistoja ovat esimerkiksi Alteryx, DataBricks ja Orange.

Työni tutkimuskysymykset ovat:

- Mikä on KNIME-ohjelmisto ja millaisia datamanipuloinnin välineitä ohjelmistossa on?
- Miten KNIME-ohjelmistolla voi hyödyntää koneoppimista ja ennustaa tulevaisuutta?
- Miten KNIME-ohjelmistolla voi käyttää visualisointeja koneoppimisen työnkuluissa?
- Millaisia vaihtoehtoja on tehdä koneoppimiseen liittyvä kolmen opintopisteen opintojakso KNIME-ohjelmistolla hyödyntäen ohjelmiston verkkokursseja?

2 Työn tavoite ja tarkoitus

Tämän opinnäytetyön tavoitteena on tutustua KNIME-ohjelmistoon ja sen tarjoamiin data-analyysin ja erityisesti koneoppimisen mahdollisuuksiin. Ohjelmistoon tutustumisessa käytetään hyödyksi KNIME-ohjelmiston tarjoamia verkkokursseja sekä niihin liittyviä harjoitustehtäviä. Työssä on tarkoitus käsitellä koneoppimisen keinoin sekä taulukkomuotoista dataa että valokuvista koostuvaa dataa. Datan pohjalta voidaan luoda erilaisia koneoppivia työnkulkuja käyttäen erilaisia koneoppimisen keinoja. Työnkulut suoritetaan KNIME-ohjelmistossa.

Koneoppiminen on tällä hetkellä kasvava data-analyysin osa-alue ja siihen liittyvää osaamista arvostetaan myös työmarkkinoilla. KNIME-ohjelmisto taas tarjoaa koneoppimisen lisäksi paljon data-analytiikan työkaluja myös muuhun käyttöön, kuten esimerkiksi datan yhdistelyyn eri rajapinnoista.

Lisäksi työn pohjalta on tarkoitus luoda esitys siitä, millaisen kolmen opintopisteen kokonaisuuden KNIME-ohjelmistoon ja koneoppimiseen liittyen voisi luoda Hämeen Ammattikorkeakoululle (HAMK) hyödynnettäväksi. Tarkoitus on myös arvioida verkkokurssien kestoa, laatua ja soveltuvuutta opiskelijoiden käyttöön. HAMK voi hyödyntää työtä omassa kurssisuunnittelussaan suunnitellessaan koneoppimiseen keskittyviä kursseja.

Opinnäytetyö on toiminnallinen työ, joka jakautuu ohjelmistoon perehtymiseen, koulutusten suorittamiseen, koneoppivien työnkulkujen luomiseen, suorittamiseen ja raportointiin sekä omien kokemuksien pohjalta luotuun näkemykseen aiheeseen liittyvästä kurssikokonaisuudesta.

Tietoperusta jakautuu datatieteeseen, koneoppimiseen, dataan ja sen analysointiin sekä KNIME-ohjelmiston perusteisiin. Nämä tukevat itse käytännön osuutta, joka keskittyy KNIME-ohjelmiston koneoppiviin työnkulkuihin. Työnkuluissa esiintyy myös esimerkiksi koneoppivien mallien arviointia, joiden ymmärrys edellyttää teoriataustaa.

3 Datatiede ja koneoppiminen

Lyhyesti sanottuna datatiede on joukko periaatteita, ongelman määrittelyjä, algoritmeja ja sellaisia menetelmiä, jossa aineistomassasta tuodaan esille hyödyllistä, ei suoraan luettavissa olevaa, tietoa. Datatiede ja tiedonlouhinta tulkitaan usein synonyymeiksi koneoppimiselle. Näille kaikille yhteistä on se, että ne tukevat päätöksentekoa aineiston analysoinnilla. Koneoppimisen perusteiden hallitseminen on tulevaisuudessa tulosvastuullisten ihmisten perusosaamista. Vaikkei algoritmeja tai matematiikkaa niiden taustalla ymmärräkään, on perusasiat tärkeää osata. Datatieteen kokonaisuutta suppeampi koneoppiminen tähtää algoritmien suunnitteluun ja arviointiin aineistosta luotujen hahmojen pohjalta. Hahmoilla tarkoitetaan esimerkiksi sellaista tekniikkaa, jolla tunnistetaan samalla tavalla käyttäytyvä asiakas ja tämän perusteella voidaan asiakas segmentoida tiettyyn ryhmään. Toinen esimerkki hahmosta on tekniikka, joka tunnistaa usein yhdessä ostetut tuotteet verkkokaupassa. Verrattuna koneoppimiseen tiedonlouhinta tarkoittaa rakenteellista aineiston analyysia. Tiedonlouhinta on tyypillisemmin suosittua kaupallisissa sovelluksissa. (Kelleher et al., 2021 s.13-14; Merilehto, 2018 s.27)

Termiä datatiede on alettu käyttää 1990-luvulla, vaikkakin siihen liittyvillä aloilla onkin huomattavasti pidempi historia. Esimerkiksi erityyppisiä aineistoja on kerätty satoja – ellei tuhansia vuosia. Eri arvioiden mukaan kirjoittamisen keksimisen jälkeen 5000 vuoden mittaan kerätyn aineisto oli kooltaan 5 eksatavua (eksa on suuruudeltaan 10^{18}). Vuoden 2013 jälkeen ihmiskunta tuottaa 5 eksatavua dataa joka päivä. (Kelleher et al., 2021 s. 17,20)

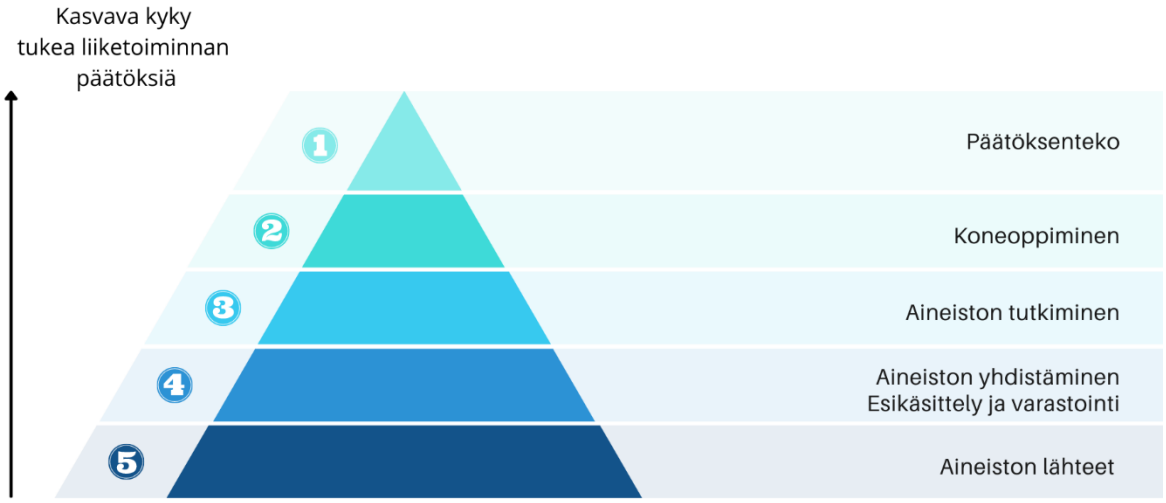
Vuonna 1956 Dartmouth Collegessa pidettiin työpaja, jossa käynnistyi tekoälyn tutkiminen. Jo tällöin tekoälyä kehitettäessä käytettiin terminä koneoppimista, kun tietokonetta käytettiin siten, että se oppi annetun aineiston perusteella. 1960-luvulla luotiin koneoppimisen päätöspuumalleja aineistosta. Samalla vuosikymmenellä tutkijat loivat myös ensimmäiset versiot k-means-algoritmista, joka on vielä nykyisinkin käytettävä väline asiakkaiden segmentoinnissa. Tästä eteenpäin 1970-luvulla luotu relaatiotietokantamalli johti tallennetun tietomäärän suureen kasvuun. Tämän jälkeen 1990-luvulla luotiin tietovarastot ja suuraineistot syntyivät. (Kelleher et al., 2021 s.24-25)

Koneoppimisen hyödyntäminen liiketoiminnassa ja sen kehityksessä tulee nykytiedon valossa kasvamaan tulevina vuosina. Koneoppimisella voidaan mahdollistaa se, että ajantasainen tieto saavuttaa oikeat ihmiset juuri silloin, kun tieto tarvitaan. Koneoppimisen menetelmillä on mahdollista korvata business intelligencen (BI) saavuttamat tulokset moninkertaisesti. (Lee, 2019 s. 3; Merilehto, 2018 s. 41)

3.1 CRISP-DM-prosessi

Datatieteen prosessin eri vaiheita voi esittää pyramidihierarkialla, jossa alhaalla ovat vähiten päätöksenteon kannalta informatiivisemmat toiminnot ja ylhäällä taas tärkeimmät. Tasojen leveys taas kertoo sen, kuinka paljon tasolla on aineistoa. Pyramidin (Kuva 1) alimmalla tasolla on lähdeaineisto. Neljännellä tasolla yhdistetään aineisto ja esikäsitellään se. Kolmas taso keskittyy aineiston tutkimiseen ja toinen taso koneoppimisen hahmojen hyödyntämiseen aineiston käsittelyssä. Ylimmällä tasolla aineistosta on voitu luoda liiketoiminnan päätöksentekoa tukeva kokonaisuus, jossa aineiston koko on selkeästi lähtötilannetta (taso 5) pienempi. (Kelleher et al., 2021 s. 62)

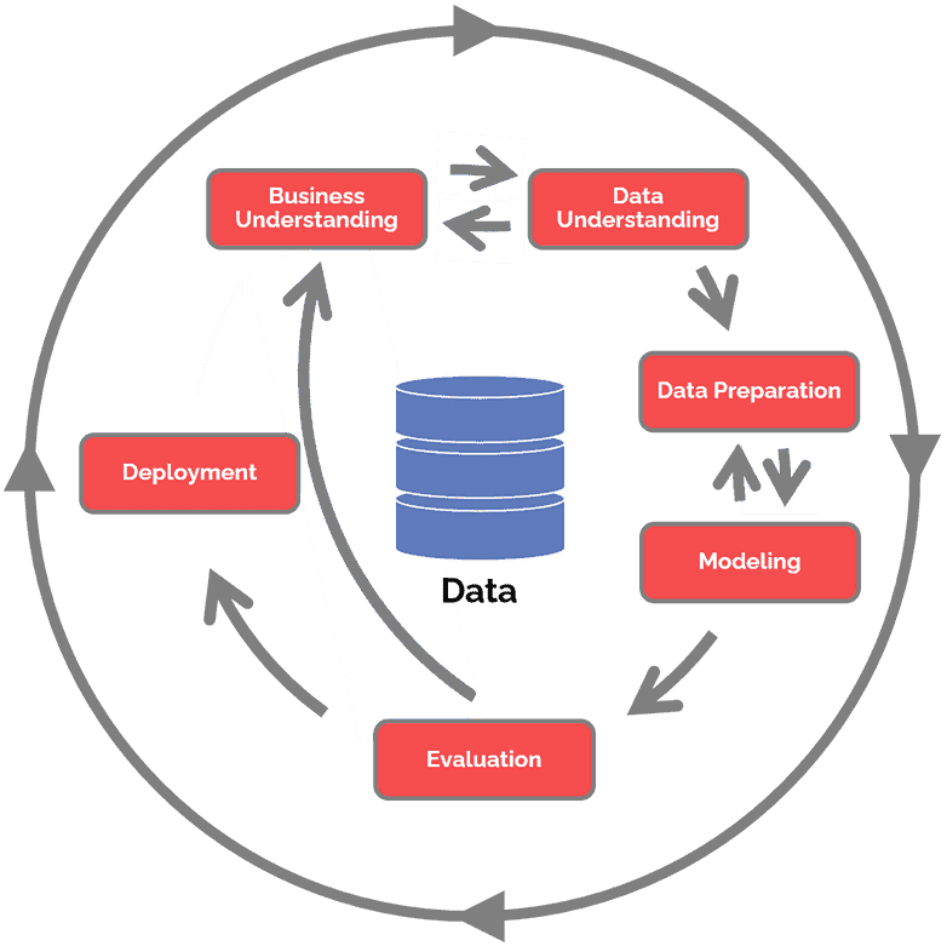
Kuva 1 Datatieteen pyramidi (Kelleher et al., 2021, s. 62)



Organisaatiot pyrkivät löytämään aina parhaan mahdollisen prosessin edetäkseen ylöspäin datatieteen pyramidilla. Suosituin käytettävä prosessi on CRISP-DM (Cross Industry Standard Process for Data Mining, laaja teollisuuden tiedonlouhintaprosessi). Prosessi on ollut jo pitkään käytössä laajalti teollisuudessa. Prosessi esiteltiin ensimmäisen kerran vuonna 1999, jonka jälkeen prosessia on yritetty kehittää ja parantaa useita kertoja. Parannusyrityksistä

huolimatta alkuperäinen prosessimalli on edelleen kaikkein suosituin. CRISP-DM-prosessia käytettäessä projektin toiminta voidaan jakaa kuuteen eri vaiheeseen: liiketoiminnan ymmärtäminen, aineiston ymmärtäminen, aineiston valmistelu, mallintaminen, arviointi sekä käyttöönotto. (Kelleher et al., 2021 s. 63-64)

Kuva 2 CRISP-DM elinkaari (DSPA, n.d.)



Kuva 2 esitellään CRISP-DM prosessin elinkaari. Prosessin keskiössä on datatieteessä tyypillisesti data. Prosessi kiertää myötäpäivään, mutta prosessissa datatieteilijän ei tarvitse kulkea vaiheittain lineaarisesta, vaan hän voi mennä myös taaksepäin nykyisessä vaiheessa saatujen tulosten perusteella. Ensimmäiset kaksi vaihetta ovat Business understanding (liiketoiminnan ymmärrys) ja Data understanding (datan ymmärrys). Niissä pyritään ymmärtämään liiketoimintaa ja aineistoa. Datatieteilijän tulee määritellä näiden tietojen perusteella projektille tavoitteet. Datatieteilijä voi esimerkiksi iteroida aineistoa, joka tarkoittaa liiketoiminnan ongelman tunnistamista ja selvitystä siitä, onko saatavilla olevasta aineiston perusteella mahdollista luoda tähän ongelmaan ratkaisu. Mikäli aineiston

perusteella ei voida luoda ratkaisua, tulee tunnistaa vaihtoehtoinen ongelma yhteistyössä organisaation eri liiketoimintayksiköiden kanssa. (Kelleher et al., 2021 s. 64-65)

Liiketoiminnan ongelman määrittelyn ja aineiston hyväksynnän jälkeen datatieteilijä voi siirtyä CRISP-DM:n seuraavaan vaiheeseen: aineiston valmisteluun (data preparation). Tyypillisesti aineiston valmistelussa yhdistellään monen tietokannan tietoja yhdeksi. Aineistoa valmistellessa on myös varmistettava, että se on laadultaan riittävää. Esimerkiksi puuttuvat arvot voivat vaikuttaa negatiivisesti laatuun ja algoritmien toimintaan, jotka aineistoa käsittelevät. Kun aineisto on valmisteltu, tehdään CRISP-DM:ssä mallintaminen (modeling). Tässä vaiheessa automaattisilla algoritmeilla luodaan aineiston pohjalta hahmoja ja luodaan myös malleja, jotka koodaavat hahmoja. Koneoppiminen keskittyy juuri näiden algoritmien suunnitteluun. (Kelleher et al., 2021 s. 65-66)

Mallintamisen vaiheessa tietoaineistoa käytetään mallin opettamiseen niin, että koneoppimisen algoritmeja ajetaan ja näin voidaan tunnistaa hyödyllisiä hahmoja. Nämä hahmot lisätään malliin. Yleensä tässä vaiheessa projektia datatieteilijä ei voi tietää, mitä hahmoja aineistosta on tarpeen etsiä, vaan hänen on tarpeellista kokeilla useita algoritmeja. Näiden useiden algoritmien tulosten perusteella voidaan valita ne algoritmit, jotka johtavat mahdollisimman täsmäävään malliin hankitusta aineistosta. Tässä vaiheessa datatieteilijä voi myös vielä tehdä havaintoja ongelmista aineiston kanssa. Joskus esimerkiksi malli ei toimi halutulla tavalla. Tässä kohtaa voidaankin siirtyä takaisin aineiston valmisteluvaiheeseen ja tehdä tarvittavia korjauksia tietoaineistoon. (Kelleher et al., 2021 s. 66-67)

CRISP-DM:n viimeisten vaiheiden arviointi (evaluation) ja käyttöönotto (deployment) aikana varmistetaan liiketoiminnan, sen prosessien ja luodun mallin yhteensopivuus. Arvioinnissa luotujen mallien sopivuutta ja vastaavuutta tavoitteisiin tarkastellaan. Samalla on myös tarkoitus varmistaa prosessien laatu: Onko kaikki liiketoiminnan kannalta oleelliset asiat huomioitu malleissa tai voisiko jotakin asiaa tarkastella vielä paremmin? Arvioinnin perusteella voidaan tehdä päätös siitä, otetaanko yksi tai useampi malleista käyttöön liiketoiminnassa vai pitääkö CRISP-DM-prosessia vielä iteroida. Jos yksi tai useampi malli hyväksytään käyttöön, voidaan aloittaa viimeinen vaihe eli käyttöönotto. Siinä on tarkoitus selvittää, kuinka mallit saadaan osaksi organisaation teknistä infrastruktuuria ja liiketoiminnan prosesseja. Käyttöönoton ohella on myös tärkeää luoda suunnitelma siitä,

kuinka mallin toimintaa seurataan ja arvioidaan käyttöönoton jälkeen. Esimerkiksi liiketoiminnan tarpeet voivat muuttua tai aineistovirtaan voi tulla muutoksia, joita käytettävät mallit eivät huomio. (Kelleher et al., 2021 s. 67-69)

3.2 Koneoppimisen teoria

Vuosien ajan perinteinen ohjelmistonkehitys ja web-ohjelmointi ovat olleet datan käsittelyn normeja. Lisäksi useita algoritmeja on luotu, jotta ohjelmat toimisivat yhä paremmin. Lähivuosina koneoppiminen on alkanut nousta perinteisen ohjelmoinnin rinnalle ja jopa ohi. Perinteisessä ohjelmoinnissa data ja ohjelmisto tuottavat lopputuloksen (output). Koneoppimisessa taas data ja lopputulos tuottavat ohjelman. Koneoppimisessa siis aiempien tapausten lopputuloksia sekä dataa käytetään ohjelman sääntöinä. Tätä kutsutaan malliksi. (Lee, 2019)

Organisaation aloittaessa koneoppimisen hyödyntämisen on tärkeää käydä koko liiketoimintaprosessi ja löytää päätökset ja päätöspisteet, jotka toistuvat prosessissa. Esimerkiksi pankeissa toistuva päätös on autolainan lainapäätös. Päätöstä varten tarvitaan kaikki mahdolliset datapisteet, eli ne tiedot mitkä puoltavat lainan antamista ja ne mitkä eivät puolla lainaa. Prosessin onnistumiseksi onkin tärkeää, että koneoppimisella ratkaistava ongelma on määritelty tarkasti ja dataa päätöksenteon avuksi on runsaasti. Kun organisaatiossa on otettu koneoppimisen menetelmiä käyttöön yksinkertaisempiin prosesseihin, voidaan kone ottaa avuksi myös monimutkaisiin ongelmiin. Monimutkaisemmissa ongelmissa ihminen tekee aina päätökset, mutta ihminen saa runsaasti hyötyä tehtäväänsä, kun kone tuottaa päätöksentekijälle kaiken saatavilla olevan ajantasaisen tiedon. (Merilehto, 2018 s. 42-43)

3.2.1 Algoritmit

Koneoppimisen tutkimusalan tehtävänä on luoda algoritmit, joilla tietokone luo hahmoja käsiteltävästä aineistosta. Koneoppimisen algoritmeja luodaan ja käytetään pääasiassa CRISP-DM:n mallintamisvaiheessa, josta kerrottiin tarkemmin aiemmassa luvussa. Koneoppimisen prosessi voidaan jakaa kahteen osaan: ensiksi algoritmien avulla eristetään

tietoaineistosta hyödyllisiä hahmoja ja luodaan mallit. Toisessa vaiheessa tehdään luoduista malleista analyysi. (Kelleher et al., 2021 s. 99-100)

Algoritmit, jotka luodaan koneoppimista silmällä pitäen, voidaan jakaa kahteen luokkaan: ohjattu ja ohjaamaton oppiminen. Ohjatussa oppimisessa tavoitteena on kehittää algoritmi, jolla tapausta kuvaavat arvot vastaavat mahdollisimman hyvin toisen, niin sanotun kohdepiirteen, arvoja. Esimerkki ohjatusta oppimisesta on sähköpostin roskapostisuodatin. Se pyrkii oppimaan ne funktiot, jotka korostavat tulevista sähköposteista piirteitä, jotka vastaavat joko normaalin sähköpostin tai roskapostin arvoja. Oppimista pidetään ohjattuna, koska tietoaineiston perusteella määritellään kaikille tapauksille syötearvo ja ulostuloarvo kokeilemalla eri funktioiden vastaavuutta olemassa olevaan tietoaineistoon. Ohjattu oppiminen edellyttää, että jokainen syötearvo saa myös ulostuloarvon. Tämän myötä on erityisen tärkeää, että tietoaineisto on mahdollisimman kattava, jotta mallin opetus onnistuu. (Kelleher et al., 2021 s. 101-102)

Ohjaamaton oppiminen eroaa ohjatusta oppimisesta siten, että siinä ei ole ulostuloarvoa. Tämän ansiosta ohjaamattomassa oppimisessa algoritmien käyttäminen ei edellytä ajankäyttöä tapausten yksilöimiseen ulostuloarvoilla (=kohdepiirre). Ohjaamattomassa oppimisessa algoritmin on tyypillisesti tarkoitus etsiä tapauksia, jotka muistuttavat enemmän toisiaan kuin aineiston tapauksia. Tätä kutsutaan klusteri- eli ryväsanalyysiksi. Ryväsanalyysin algoritmit toimivat esimerkiksi niin, että ne aloittavat arvaamalla ryväsjoukon ja päivittävät ryväsjoukkoa poistamalla tai siirtämällä tapauksia toiseen joukkoon. Tämän myötä syntyneet ryppäät ovat toisistaan poikkeavia, mutta rypäiden sisällä taas joukko on samankaltainen. Ryväsanalyysissa on haasteellista määritellä, kuinka samankaltaisuus mitataan. Jos tapaukset ovat mitattavissa numeroilla ja vaihteluvälit ovat tasaisia, on yksi keino laskea tapausten lyhin etäisyys (euklidinen etäisyys). Joissakin tapauksissa aineistossa voi olla tärkeämpiä ja ei-tärkeitä piirteitä numeerisina arvoina. Tärkeämpiä piirteitä on tarpeen painottaa etäisyyksien laskennassa. (Kelleher et al., 2021 s. 102-104)

3.2.2 Ennustavat mallit

Oppivia ennustavia malleja luodaan koneoppimisen yhteydessä. Ennustavien mallien käyttö on yksi suosituimmista koneoppimisella ratkaistavista ongelmista. Ennustuksessa on

tarkoituksena arvioida kohdepiirteelle arvosyötepiirteiden perusteella juuri tietyssä tapauksessa. Tässä voidaan hyödyntää koneoppivia algoritmeja, joiden lopputuotteena syntyy ennustavia malleja. Ennustavaa mallia käytetään esimerkiksi silloin, kun opetusaineistosta puuttuvista tapauksista tehdään arvio kohdepiirteen arvosta. Esimerkiksi sähköpostin roskapostisuodattimen aineistona ovat vanhojen sähköpostien aineisto. Tämän aineiston pohjalta luodun mallin perusteella määritellään, onko uusi sähköposti roskapostia vai ei. (Kelleher et al., 2021 s. 105-106)

3.2.3 Korrelaatio

Korrelaatio tarkoittaa kahden eri piirteen keskinäistä yhteyttä ja sen vahvuutta. Korrelaatiossa lineaarisen suhteen vahvuutta kuvataan arvoilla -1 ja +1 välillä. Pearsonin korrelaatiossa kerroin korrelaation arvoa kuvataan kirjaimella r. Jos r = 1, piirteet korreloivat täydellisesti. Jos r= 0, piirteet eivät korreloi. Jos taas r = -1, piirteet korreloivat täydellisesti negatiivisesti. Tällaisella negatiivisella korrelaatiolla yhden piirteen muuttuessa tapahtuu toiselle piirteelle päinvastainen muutos. Vahvan lineaarisen suhteen raja on r = 0,7 tai r = -0,7, kohtalaisen r = 0,5 tai r = -0,5 ja heikon r = 0,3 tai r = -0,3. Jos taas r = 0, ei kahden piirteen välillä ole keskinäistä suhdetta. Esimerkiksi ihmisellä kengännumeron ja pituuden välillä on keskinäinen korrelaatio, pituuden kasvaessa on tyypillisesti myös kengännumero isompi. (Kelleher et al., 2021 s. 106-107)

3.2.4 Lineaarinen regressio ja regressioanalyysi

Jos aineisto sisältää numeerisia piirteitä, käytetään tyypillisesti lineaariseen regressioon perustuvia malleja. Regressioanalyysin tarkoituksena on arvioida numeerisen kohdepiirteen odotusarvo. Regressioanalyysi voidaan jakaa kahteen vaiheeseen. Ensimmäisenä oletetaan syötepiirteiden ja kohteen suhteen rakenne. Toisessa vaiheessa määritellään parametrisoitu matemaattinen malli aiemmin tehdyn suhdeoletuksen pohjalta. Tätä parametrisoitua mallia kutsutaan regressiofunktioksi. Regressioanalyysia tehtäessä regressiofunktion parametrit ovat aluksi tuntemattomia. Parametrien asettaminen aloitetaan arvaamalla arvot ja niitä päivitetään siten, että funktion kokonaisvirhe suhteessa käytettävään aineistoon pienenee. Kokonaisvirheen laskeminen voidaan jakaa kolmeen vaiheeseen: Ensimmäisessä vaiheessa funktiota sovelletaan aineistoon ja arvioidaan kohdepiirteen arvo. Toisessa vaiheessa

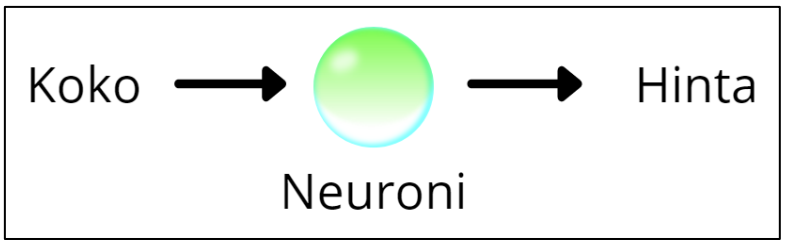
funktion virheen laskemiseksi vähennetään kohdepiirteen arvio sen todellisesta arvosta. Kolmannessa vaiheessa jokaisen eri tapauksen virhefunktio muutetaan toiseen potenssiin ja lasketaan näiden arvojen summa. Toiseen potenssiin korotus tehdään, jotta kohdearvosta tehdyt liian korkeat arviot eivät kumoaisi liian alhaisia arvioita. Tämä myös varmistaa sen, että kaikki virheet ovat positiivisia. Menetelmää, jossa tehdään virhefunktion toiseen potenssiin korotus, kutsutaan myös virheiden neliöinniksi (SSE, Sum of Squared Errors). (Kelleher et al., 2021 s. 113-116)

3.2.5 Neuroverkot

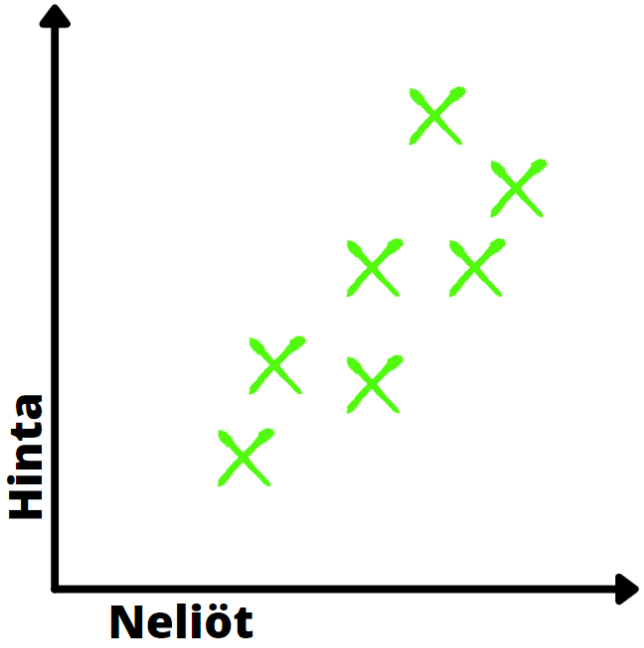
Koneoppimisen ja erityisesti syväoppimisen osana ovat neuroverkot. Neuroverkko tarkoittaa joukkoa toisiinsa kytkettyjä neuroneja. Neuroni taas tarkoittaa monesta syötearvosta syntyvää lineaarista regressiofunktiota. Näiden taustalla on iso määrä syötedataa. Neuroverkkojen voikin ajatella jäljittelevän ihmisaivojen tapaa toimia. Verkoissa on neuronien lisäksi synapseja. Neuronit ovat ikään kuin vain yhtä asiaa tekeviä solmukohtia ja synapsit kaaria seuraavaan solmukohtaan. Neuroverkossa jokaisen neuronin ominaisuuksiin kuuluu suunta sekä paino ja näitä käyttämällä saadaan matemaattista kaavaa käyttämällä neuroverkon ulostuloarvo. Neuroverkot vaativat optimaalisesti toimiakseen mahdollisimman paljon lähdeaineistoa. Mitä enemmän lähdeaineistoa on, sitä paremmin neuroverkot kykenevät oppimaan. Neuroverkkojen käyttöä on mahdollista soveltaa laajasti erilaisten ongelmien ratkaisuihin. Esimerkiksi kuvien tunnistaminen tai roskapostisuodatin hyödyntävät neuroverkkoja toiminnassaan. Kahden erilaisen toiminnon neuroverkot voivat olla samanlaisia, kun tarkastellaan neuronien määrää. Neuroverkkojen erona on koneoppimisen myötä tulleet painokertoimet. (Kelleher et al., 2021 s. 119-123; Merilehto, 2018 s. 47-48)

Yksinkertainen esimerkki neuroverkosta käytännössä on asuntojen hinnan ja neliöiden suhde. Nämä esitellään Kuva 3 ja Kuva 4. Tässä esimerkissä syötteenä on asunnon koko neliöinä ja vasteena hinta. (Merilehto, 2018 s. 49)

Kuva 3 Esimerkki: Koko kertoo asunnon hinnan (Merilehto, 2018 s.50)

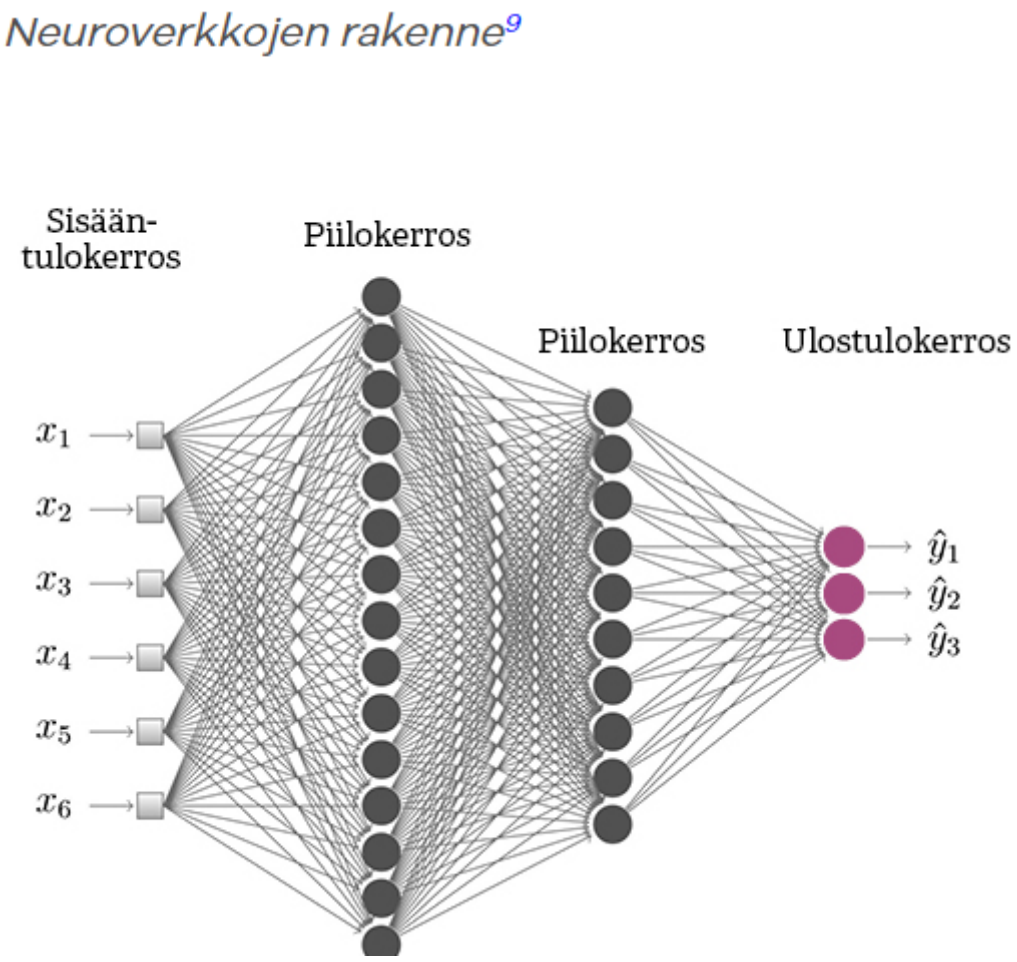


Kuva 4 Esimerkki: Hinnan ja neliöiden lähdeaineisto (Merilehto, 2018 s.49)



Neuroverkkoja käytettäessä datalla on suuri merkitys ja tyypillisesti riittävänä datamääränä pidetään kymmeniä- tai satoja tuhansia rivejä dataa. Monimutkaisempi neuroverkko on monikerroksinen perseptroniverkko (multilayer preceptron, MLP), jossa verkko koostuu yhteen tai useampaan kerrokseen tehdyistä keinotekoisista neuroneista (Kuva 5). Näitä kutsutaan perseptroneiksi. Kerrokset voidaankin luokitella kolmeen eri luokkaan: ulostuloskerroksiin, piilokerroksiin ja sisääntulokerrokseen. Sisääntulokerroksessa otetaan vastaan neuroverkolle tuleva tieto ja välitetään se eteenpäin ensimmäiselle piilokerrokselle. Sisääntulokerros poikkeaa muista kerroksista siten, ettei siinä tapahdu koneoppimista. (Merilehto, 2018 s. 51-52)

Kuva 5 Neuroverkkojen rakenne (Merilehto, 2018 s. 52)



3.2.6 Koneoppivat mallit ja arvioinnin tunnusluvut

Aiemmin esitelty lineaarinen regressiomalli toimii parhaiten numeerisia arvoja tarkasteltaessa. Jos tietoaineistossa on muuta kuin numeerista dataa, on tarkoituksenmukaisempaa käyttää esimerkiksi päätöspuuta. Päätöspuu toimii siten, että se luo joukon sääntöjä puunkaltaiseen malliin. Esimerkiksi sähköpostin roskapostisuodattimessa sääntö voi olla seuraava: jos sähköposti on saapunut ennalta tuntemattomalta lähettäjältä, niin se on roskapostia. Päätöspuussa päätökset tehdään puun latvasta lähtien kulkien alaspäin. Puun jokainen solmu testaa yhtä piirrettä ja puussa edetään tuloksen mukaisesti. Lopullinen tulos on haaran viimeisen päättävän solmun nimi. (Kelleher et al., 2021 s. 132-133)

Jokainen päätöspuun polku tekee määritelmän luokitussäännöstä. Tavoitteena on määritellä sellaisia luokitussääntöjä, jotka jakavat aineiston saman kohdepiirteen arvon ryhmiin. Kun

uusi arvo kulkee saman polun päätöspuussa, kuin aiempi luokiteltu arvojoukko, voi sen olettaa olevan kyseistä arvojoukkoa vastaava. Päätöspuumallin etuihin kuuluu helppo ymmärrettävyys. (Kelleher et al., 2021 s. 133-134, 137)

Satunnaisen metsän malli koostuu useista päätöspuista, joiden opetusaineisto on valittu satunnaisella otannalla koko opetusaineistosta. Ennustava malli toimii enemmistön periaatteella, eli yksittäistä arvoa arvioitaessa valitaan metsän kaikkien puiden ennustusten enemmistö. (Kelleher et al., 2021 s. 137-138)

Malleja tulee myös arvioida. Kaksi keskeistä arvioinnin lukua ovat Cohen's kappa ja R^2-luku. Cohen's kappa on luku -1 ja 1 väliltä. Lyhyesti sanottuna luku kertoo erotuksen mallin tarkkuuden ja satunnaisella arvonnalla saadun tarkkuuden välillä. Jos malli toimii hyvin, Cohen's kappa luku on lähellä 1. Satunnaisesti toimiva malli on lähellä arvoa 0. Satunnaista arvontaa heikompi malli voisi teoriassa olla myös miinusmerkkinen eli alle 0. (Widmann, n.d.)

R^2-luku kertoo kuinka paljon arvioitavan muuttujan vaihtelusta selittävät tekijät pystyvät selittämään. Tuloksena saadaan luku, joka on 0 ja 1 välillä. Jos saatu r^2 luku on esimerkiksi 0,44, tarkoittaa tämä sitä, että vaihtelua selittävät muuttujat kertovat 44 % vaihtelusta. (KvantiMOTV, n.d.)

3.3 Syväoppiminen

Syväoppimisen perustana ovat keinotekoiset piilokerroksia sisältävät neuroverkot. Järjestelmä, joka hyödyntää syväoppimista, pystyy opettamaan itsellensä tunnistamaan kuvista esimerkiksi auton, kunhan sille on näytetty riittävästi luokiteltuja esimerkkejä kuvista, joissa on auto. Ohjatun oppimisessa tärkeää on, että järjestelmän saama opetusaineisto koostuu sekä oikeista, että vääristä esimerkeistä. Tärkeimpiä hyötyjä syväoppimisessa on se, että sitä hyödyntävä järjestelmä voi oppia raakadatasta, jossa on mukana myös virheitä. Laskentatehon parantuminen on mahdollistanut monimutkaisten syväoppivien järjestelmien kehittämisen, joita hyödynnetään esimerkiksi lääketieteessä tautien havaitsemiseen otetuista kuvista. (Merilehto, 2018 s. 56-57)

Syväoppiminen on jo niin pitkällä, että luodut mallit ja järjestelmät voivat opettaa itse itseään. Tätä kutsutaan automatisoiduksi koneoppimiseksi. Syväoppivat automatisoidut järjestelmät kykenevät myös itsenäisesti ohjelmoimaan tekoälyä hyödyntäviä järjestelmiä. Automatisoitu koneoppiminen kykenee ohjelmoimaan vain perustason tekoälyohjelmia. Silti tämä mahdollistaa sen, että aikaa vievien ja toistuvien tehtävien sijaan tekoälyn parissa työskentelevät voivat keskittyä haastavampiin tehtäviin. Kehitys tekoälyn parissa on kuitenkin nopeaa ja automatisoidun koneoppimisen menettelyillä ohjelmistokehittäjien on mahdollista työskennellä tekoälyyn liittyvien teknologioiden parissa, vaikka ohjelmointitaidoissa olisikin puutteita. (Merilehto, 2018 s. 58-60)

3.4 Vahvistusoppiminen

Aiemmin mainittujen ohjatun ja ohjaamattoman oppimisen lisäksi konetta voidaan kouluttaa myös vahvistusoppimisen (reinforcement learning) keinoin. Vahvistusoppiminen koostuu agentista ja ympäristöstä. Agentti on algoritmi, joka liikkuu ympäristössä. Tämä algoritmi on suunniteltu niin, että sen tavoitteena on löytää mahdollisimman paljon positiivisia pisteitä ympäristöstä. Algoritmin toimii ikään kuin koira: Se saa palkinnon oikeasta suorituksesta ja negatiivista palautetta väärästä suorituksesta. Vahvistusoppimisen perustana onkin ongelmanratkaisutekniikka, eli algoritmi tutkii ympäristöä ja toimii ympäristön mukaisesti. Vahvistusoppimista on ollut haastavaa hyödyntää liike-elämässä, mutta sitä on hyödynnetty mm. itseohjautuvissa autoissa ja dynaamisessa hinnoittelussa. (Kananen, 2019 s. 158-159)

4 Data ja sen analysointi

Data voidaan määritellä koostuvan faktatiedosta, havainnoista ja raa’asta tiedosta. Datalla itsessään ei kuitenkaan ole suurta tarkoitusta, jos se on prosessoimatonta. Datan analysoinnilla tarkoitetaan datan hyödyntämisen käyttötapoja, joilla voidaan käsitellä suuria datamääriä. Kun suurta määrää ns. raakadataa käsitellään, voidaan tuoda esille mielenkiintoisia näkökulmia datan perusteella. Jotkut datan analysointitavat auttavat hahmottamaan paremmin eri dataryhmien välisiä yhteyksiä, toiset taas auttavat ryhmittelemään dataa, jotta siitä voidaan tehdä havaintoja ja ymmärtää kokonaisuutta paremmin. (Sedkaoui, 2018 s. 25)

Puhuttaessa datasta puhutaan yhä useammin myös niin kutsutusta Big datasta. Big datan taustalla on ajatus siitä, että kaikki tekemämme asiat jättävät yhä suuremman digitaalisen jäljen (tai palan dataa), jota voidaan käyttää ja analysoida toiminnan tehostamiseksi. Big data on melko uusi termi, sillä datan määrä on moninkertaistunut lyhyessä ajassa. Jos tarkastellaan kaikkea maailmassa syntynyttä dataa vuoteen 2010 asti, arvioiden mukaan pian luodaan sama määrä dataa joka minuutti. Telekommunikaation alalla toimivat yritykset käyttävät Big dataa esimerkiksi asiakastyytyväisyyden määrittämiseen seuraamalla puhelimen käyttöä ja asiakkaan toimintaa sosiaalisen median alustoilla. Big dataa hyödyntämällä jokainen asiakas voitiin segmentoida tarkkaan kategoriaan esimerkiksi sen suhteen, olivatko he todennäköisesti lopettamassa sopimustaan ja siirtämässä sopimustaan kilpailevalle operaattorille. (Marr, 2015 s. 9-11, 57)

Big data syntyy digitaalisesti ja sitä kerätään automaattisesti. Big dataa voidaan tarkastella ja arvioida kolmen V:n periaatteella. Volume (määrä) viittaa datan kokoon. Big data viittaa jo lähtökohtaisesti suureen datamäärään. Datan määrä vaihtelee organisaatiosta ja sektorista riippuen. Esimerkiksi IoT-laitteet (Internet of Things), joissa on lukuisia sensoreita tuottavat suuria määriä dataa joka sekunti. Velocity (nopeus) viittaa datan keräyksen ja tallennuksen nopeuteen sekä aikaan, jossa dataan tulisi reagoida. Erilaiset datalähteet ja dataformaatit ovat yksi teknologinen haaste datan käsittelyn nopeudessa. Variety (moninaisuus) viittaa taas siihen, kuinka heterogeenistä data on ja miten sitä voidaan tulkita. Jos esimerkiksi dataa tulee tekstinä, kuvina ja videoina niin dataa voidaan pitää moninaisena. Lisäksi joissakin tapauksissa Big datan arvioinnissa voidaan käyttää myös neljättä V:tä. Tämä on todenmukaisuus (veracity), joka kertoo luodun datan siisteydestä. Esimerkiksi sosiaalisen median viesti, jossa on aihetunnisteita, kirjoitusvirheitä, lyhenteitä ja puhekielen viittauksia on haasteellinen analysoinnissa. Big dataa käsiteltäessä on tärkeää tunnistaa eri datatyypit, niiden merkittävyys ja datalähteet.
(Marr, 2015 s. 79-80; Sedkaoui, 2018 s. 41-45)

Data voidaan jakaa myös strukturoituun ja strukturoimattomaan dataan. Strukturoitu data tarkoittaa dataa, joka on jaoteltu etukäteen määritellysti omiin kenttiinsä ja tiettyyn tiedostoon. Data voi olla myös tietokannassa tai laskentataulukossa. Strukturoitu data tarjoaa sellaisenaan eniten tukea analysointiin ja analysointi on helppoa. Strukturoidussa datassa jokainen kenttä on nimetty ja siinä määritellään eri kenttien väliset suhteet.

Esimerkiksi myyntitilasto, talouden data tai asiakastiedot ovat strukturoitua dataa. Strukturoimaton ja osittain strukturoitu data koostuu datasta, jota on vaikeaa asettaa omiin riveihin, kenttiin tai sarakkeisiin. Yleensä strukturoimaton data on tekstipainotteista, mutta voi sisältää myös päivämääriä, numeroita tai muuta dataa, kuten kuvia. Koska strukturoimaton data on sisällöltään epätasapainoista, sen analysointi on vaikeaa perinteisillä tietokoneohjelmilla. Esimerkiksi kuvat, videot, verkkosivut, tekstitiedostot tai Powerpoint-esitykset ovat strukturoimatonta dataa. Osittain strukturoitu data on sekoitus strukturoitua ja strukturoimatonta dataa. Tällaisessa datassa on osittain tunnistettava rakenne, jota voidaan käyttää analyysiin, mutta osa datasta on strukturoimatta. Esimerkiksi Facebookiin tehty kirjoitus on osittain strukturoitu. Siitä on saatavilla kirjoittajan nimi, pituus, kirjoitusaika ja tunnetila (sentiment), mutta päivityksen sisältö on tyypillisesti strukturoimatonta. (Marr, 2015 s. 59-62)

Datamallilla tarkoitetaan mallia, jolla organisaatio tallentaa, prosessoi ja käyttää organisaation toimintaan liittyvää dataa. Datamallissa jokainen eri datakentän sisältö määritellään. Esimerkiksi asiakastietokannassa datakenttiä ovat nimi, osoite, puhelinnumero ja sähköposti. Näiden datakenttien määritellyssä voidaan ottaa huomioon esimerkiksi se, että puhelinnumeroa pyytävään kenttään hyväksytään vain numeroita tai sähköpostiosoite-kenttä edellyttää @-merkkiä. (Marr, 2015 s. 60)

Erilaisista sensoreista tulevan datan määrä on kasvanut ja jatkaa kasvuaan tulevaisuudessa. Esimerkiksi juuri sensoreista saatava data tuottaa Big dataa. Sensoreita on tällä hetkellä monissa laitteissa. Esimerkiksi älypuhelimissa on seuraavia sensoreita: GPS-sensori, kiihtyvyysmittari, gyroskooppi, etäisyyssensori, valoisuutta seuraava sensori sekä NFC (Near Field Communications) sensori. GPS kertoo laitteen sijainnin käyttäen GPS satelliittien dataa. Kiihtyvyysmittari taas mittaa puhelimeen liikettä ja se auttaa esimerkiksi ottamaan parempia kuvia, vaikka puhelin liikkuisikin kuvaa ottaessa. Gyroskooppi seuraa puhelimen näytön orientaatiota ja auttaa pitämään ruudun käyttäjän kannalta oikein päin, vaikka puhelinta kääntäisikin. Etäisyyssensori taas nimensä mukaisesti mittaa laitteen läheisyyttä muihin esineisiin tai asioihin. Valoisuutta seuraava sensori (ambient sensor) seuraa taas puhelimen ympäristön valoisuutta ja muuttaa näytön kirkkautta sen mukaisesti. NFC-sensori taas mahdollistaa esimerkiksi maksamisen toiseen puhelimeen tai maksupäätteeseen. Analysoitavaa dataa saadaan sensoreista siis runsaasti. (Marr, 2015 s. 73-74)

4.1 Datan analysointi

Big datan analysointi on tullut tärkeäksi osaksi nykyaikaista liiketoimintaa. Big datan analysoinnilla on yritystoimintaan positiivisia vaikutuksia, kuten päätöksen tukeminen, kulujen pienentäminen, asiakaskäyttäytymisen ymmärtäminen ja avoimen datan hyödyntäminen esimerkiksi uusien tuotteiden ja palveluiden rakentamiseen. Haasteena datan käytössä ei ole vain datan keräys, vaan datan käsittely ja analysointi paremmin siten, että se tukee yrityksen päätöksentekoa ja auttaa yritystä toimimaan tehokkaammin. Koska dataa syntyy koko ajan, myös datan analysointi on tehtävä jopa reaaliaikaisesti. Perinteinen datan analysointi tehdään niin, että data kerätään ensin, sitten se muutetaan ja lopuksi analysoidaan. Nämä kolme askelta on tehty yksi kerrallaan. Nykyisin dataa voidaan analysoida jo reaaliaikaisesti, esimerkiksi terveyspalveluiden tuottaja voi seurata potilaitaan, joilla on vakava riski sairastua. Tämä tapahtuu yhdistämällä reaaliaikaista dataa useista laitteista, joilla seurataan terveyden tilan muutoksia ja erilaisia oireita. Kun tämä data käsitellään ja yhdistetään potilaan terveystietoihin, analysointityökalut voivat ilmoittaa terveydenhuollon ammattilaisille, mikäli ennakoiville toimille on tarvetta. (Sedkaoui, 2018 s. 91, 93-94)

4.2 Datan käsittelyn tulevaisuus

Ympäröivää maailmaa aistivien ja näiden havaintojen perusteella reagoivien antureiden ja näitä hyödyntävien laitteiden määrä kasvaa jatkuvasti. Esimerkiksi autoissa ja kodeissa on yhä enemmän älykkäitä laitteita, jotka keräävät havainnoistaan dataa. Tämä kehitys haastaa ihmisten yksityisyyttä ja lisää analysoitavan Big datan määrää, mutta antaa myös mahdollisuuksia, kuten henkilökohtainen lääketiede ja älykkäät kaupungit. (Kelleher et al., 2021 s. 209)

Lääkärit ovat joutuneet sairauden diagnosoinnissa luottamaan omiin kokemuksiinsa ja vaistoihinsa. Lääketieteessä on myös näkemys, jonka mukaan lääkärin päätösten tulee perustua aineistoon, joka parhaassa tapauksessa voidaan yhdistää yksittäisen potilaan tilasta ja elämäntavoista saatuun aineistoon. Lisäksi data-analyysia käyttämällä voidaan määritellä se, kenelle jokin lääke tai antibiootti tulisi antaa ja kuinka tehokkaasti se on toiminut. Kehitteillä on myös potilaan kantamia, nieltyjä tai istutettuja lääketieteellisiä antureita, joilla

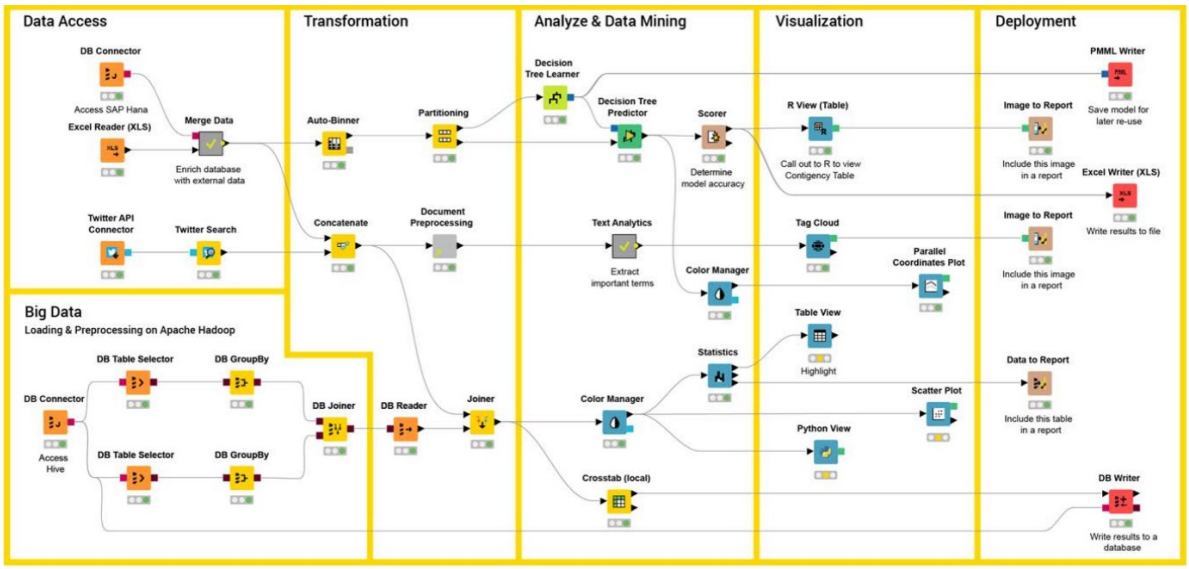
on mahdollista seurata potilaan elintoimintoja. Näistä antureista tietoa kerätään vuorokauden ympäri ja saatu tieto lähetetään keskitettyyn palvelimeen, josta hoitava henkilökunta voi nähdä hoidon vaikutukset ja arvioida tarpeet tuleville toimenpiteille. (Kelleher et al., 2021 s. 209-210)

Maailmanlaajuisesti yhä useammat kaupungit ovat ottaneet käyttöönsä uutta teknologiaa, jolla voidaan kerätä asukkaiden tuottamaa dataa ja hyödyntää sitä kaupungin eri toiminnoissa. Tämän mahdollistaa kolme asiaa: datatiede, Big data ja esineiden internet (IoT). Älykkäät kaupungit yhdistävät useista lähteistä saatua reaaliaikaista dataa yhdeksi tietokoosteeksi, jossa sitä voidaan hyödyntää analyysin jälkeen kaupungin päätöksenteossa. Tyypillisesti älykaupunki-hankkeiden perustana on jo olemassa oleva kaupunki, johon asennetaan antureita ja tietojenkäsittelykeskuksia. Esimerkiksi Espanjassa SmartSantander-projektin myötä yhteen kaupunkiin on asennettu yli 12 000 anturia, jotka mittaavat lämpötilaa, melua, valaistusta, häkätasoa sekä pysäköintiä. Myös liikennettä voidaan hallita data-analyysin keinoin. Jotkin kaupungit ovat asentaneet liikenteenseurantajärjestelmiä, jotka seuraavat antureilla reaaliajassa liikennettä. Järjestelmä voi tarvittaessa muuttaa esimerkiksi liikennevalojen toimintaa niin, että se tukee julkisen liikenteen sujuvuutta. (Kelleher et al., 2021 s. 211-213)

5 KNIME-ohjelmisto

KNIME-ohjelmisto (viralliselta nimeltään KNIME Analytics Platform) on avoimen lähdekoodin ilmainen ohjelmisto, jota hyödynnetään esimerkiksi datatieteessä. KNIME-ohjelmiston avulla voidaan käsitellä kaikki data-analytiikan vaiheet datan hankinnasta koostettuihin loppuraportteihin asti, luoda kustomoituja työnkulkuja sekä käyttää lukuisia ohjelmiston työkaluja ja komponentteja apuna. Ohjelmistoa käytettäessä luodaan visuaalinen työnkulku (Kuva 6), jossa dataa käsiteltäessä haetaan ensin lähdedata yhdestä tai useammasta lähteestä. Sen jälkeen dataa muokataan, yhdistellään tai käsitellään erilaisilla työkaluilla sopivaksi. Lisäksi erilaisia koneoppimisen menetelmiä on käytettävissä, kuten esimerkiksi syväoppiminen. Kun datasta on saatu luotua uusia kokonaisuuksia, ne voidaan tuoda visualisointeina tai raportteina hyödynnettäväksi loppukäyttäjälle. (Knime, 2020)

Kuva 6 Esimerkki KNIME-ohjelmiston työnkulun vaiheista.(Knime, 2020)

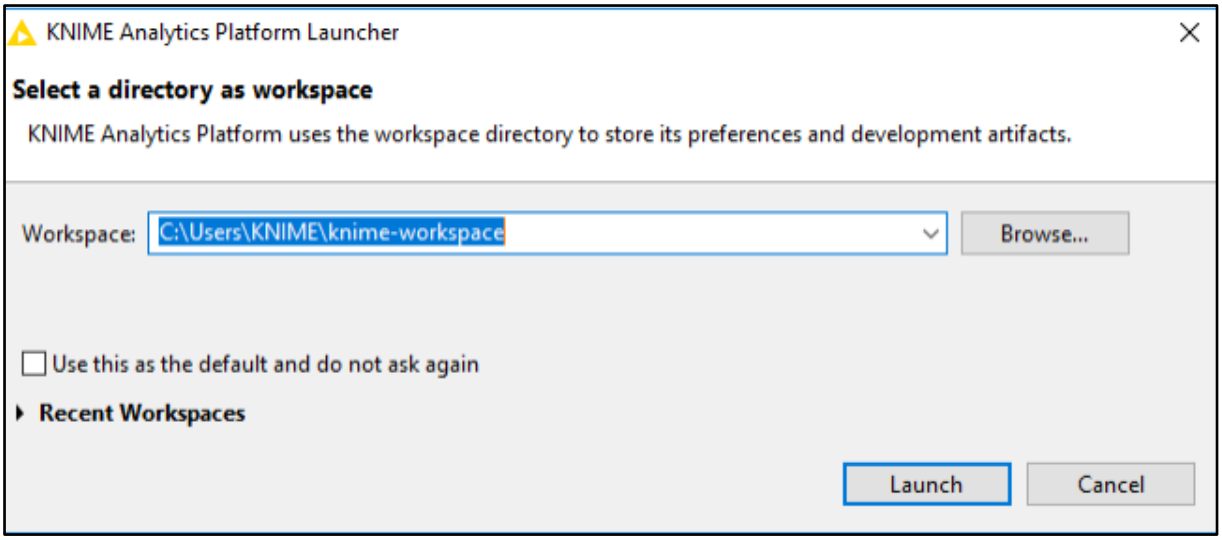


KNIME Analytics Platform covers all stages of the data science life cycle.

5.1 KNIME-ohjelmiston käyttö

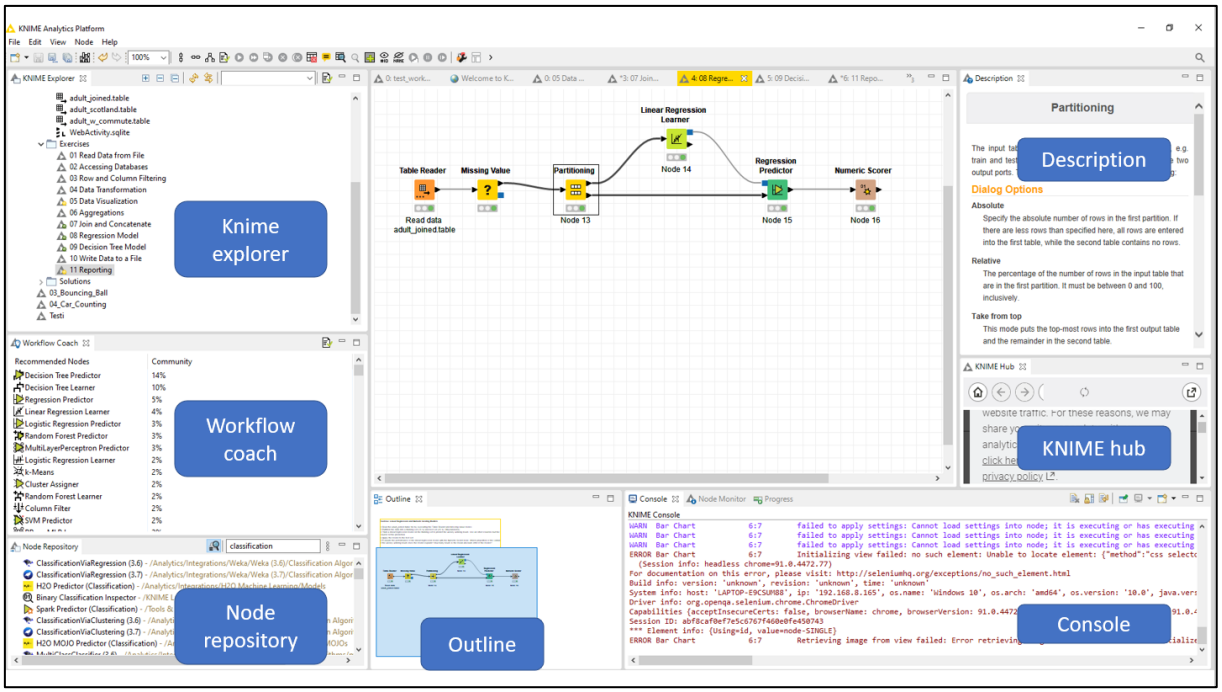
Ohjelmisto on ilmainen ja ladattavissa verkosta osoitteesta https://www.knime.com. Ohjelmiston asentamisen jälkeen ohjelmisto on valmis käytettäväksi. Ensimmäistä kertaa käynnistettäessä ohjelmisto kysyy työtilan osoitetta. Oletuksena on valittuna työtilaksi C:\Users\käyttäjänimi\knime-workspace. Työtila on kansio, jossa kaikki KNIME-ohjelmistolla tehdyt työnkulut (workflow), asetukset sekä luotu data tallennetaan. Työtilan valintaikkuna on Kuva 7. Työtilan sisältö on saatavilla ohjelman ollessa käynnissä vasemmalla ylhäällä sijaitsevasta KNIME Explorer -ikkunasta. (Anon., 2021 s. 2)

Kuva 7 KNIME-ohjelmiston työtilan valinta (Anon., 2021 s. 2)



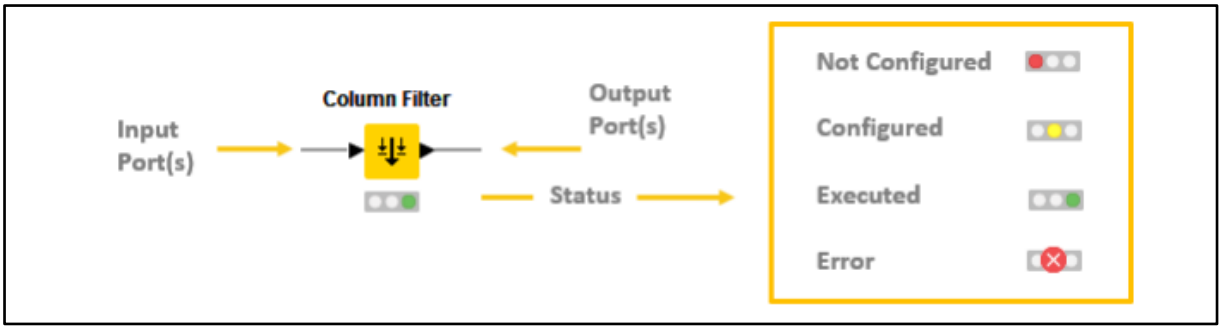
KNIME-ohjelmiston näkymä jakautuu seitsemään eri komponenttiin. Komponenttien sijainti ohjelman yleisnäkymässä on esitelty Kuva 8. KNIME explorer näyttää saatavilla olevat työnkulut ja työnkulun ryhmät omassa työtilassa, KNIME serverillä ja henkilökohtaisen KNIME hub tilan. Workflow coach listaa suositeltavat noodit (node) sen perusteella, millainen työnkulku tällä hetkellä on ohjelmassa käytössä. Suositukset perustuvat KNIME-yhteisön käyttäjien rakentamiin työnkulkuihin. Node repository listaa kaikki noodit, jotka ovat saatavilla ohjelmistossa. Saatavilla olevien noodien määrä riippuu asennetuista lisäosista. Noodeja on mahdollista myös etsiä ryhmittäin tai hakusanalla hakukentästä. Workflow editor on ohjelmiston keskiössä ja se sisältää visuaalisessa muodossa tällä hetkellä avoinna olevan työnkulun ja siihen kuuluvat noodit. Description kertoo nykyisen aktiivisen työnkulun tai valitun noodin kuvauksen. Outline näyttää koko työnkulun kuvan ja on hyödyllinen varsinkin laajempien työnkulkujen hahmottamisessa. Console on nimensä mukaisesti konsoli, josta näkee mitä tapahtuu suoritettaessa noodeja tai kokonaisia työnkulkuja. (Anon., 2021 s. 3)

Kuva 8 KNIME-ohjelmiston komponentit.



KNIME-ohjelmistossa yksittäisiä tehtäviä hallitaan noodeilla. Jokaisen noodi näytetään värillisenä laatikkona, jossa on tulo- ja lähtöportti (input ja output) tai joissakin tapauksessa useampia portteja. Lisäksi noodin yhteydessä on liikennevalojen kaltainen statusta ilmaiseva kuva. Tuloporttiin tulee se data, jonka noodin tulee käsitellä ja lähtöportissa on käsitelty data, joka esimerkiksi siirtyy seuraavaan noodiin. Jokaisella noodilla on omat asetukset, joita voidaan käsitellä konfigurointi-ikkunan kautta (configuration). Konfigurointi-ikkuna avataan esimerkiksi painamalla noodia oikealla hiiren painikkeella ja valitsemalla ”Configuration”. Kun noodin asetukset ovat kunnossa, status muuttuu punaisesta keltaiseksi. Kun taas noodi on ajettu onnistuneesti, status on vihreä. Mikäli status on punainen valo, jonka päällä on risti keskellä statusta ilmaisevaa kuvaketta, tarkoittaa se noodin virhetilaa. Noodien portit ja tilat on esitelty myös Kuva 9. (Anon., 2021 s. 4)

Kuva 9 Noodi sekä sen portit ja statusta ilmaiseva kuvake. (Anon., 2021)



5.2 Ohikulkevien autojen laskeminen hyödyntäen KNIME-ohjelmaa

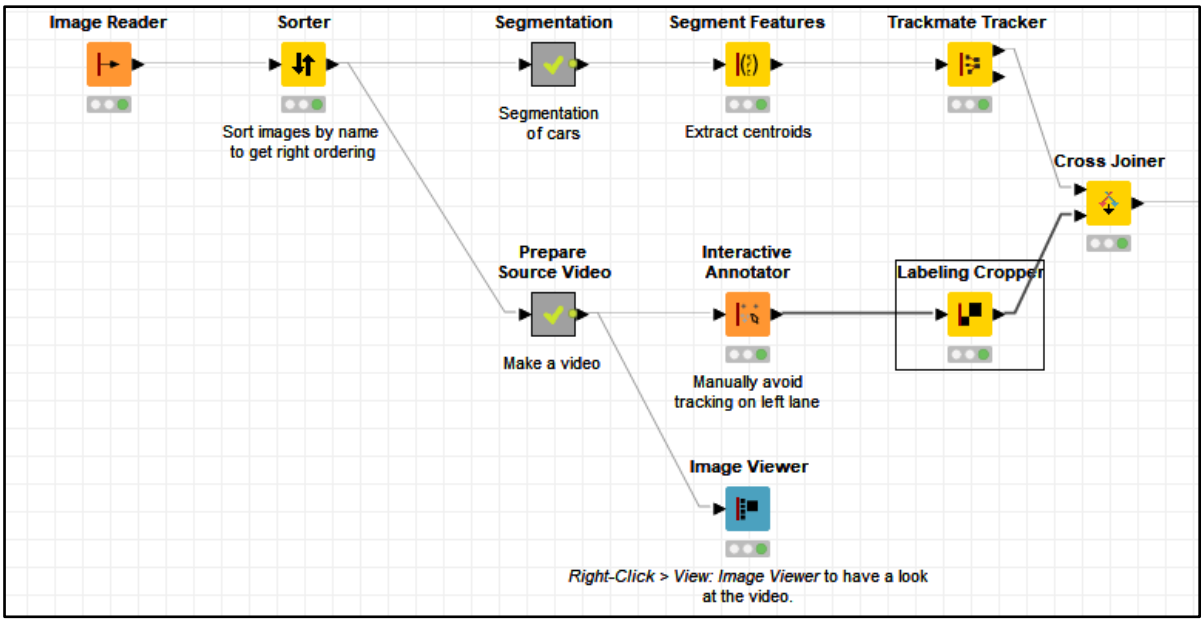
Ohikulkevien ajoneuvojen laskeminen videolta on ihmiselle kohtuullisen helppo tehtävä, mutta jos analysoitava aineisto on tuntien mittainen, voi työ olla puuduttavaa ja virheitä voi syntyä helposti. KNIME-ohjelmistolla tämä autojen laskeminen on mahdollista automatisoida. Tässä esimerkissä käytetään KNIME Hubissa saatavilla olevaa työnkulkua ja siihen kuuluvaa 300 kuvan aineistoa, joka on saatavilla osoitteesta https://hub.knime.com/knime/spaces/Examples/latest/99_Community/01_Image_Processing/03_Applications/04_Car_Counting~6W_4CLjiCAykbVe0. Prosessi alkaa sillä, että video jaetaan yksittäisiksi kuviksi. Tässä esimerkissä video on kuvattu 30fps kameralla ja kuvia on 300. Tämä tarkoittaa siis kymmenen sekunnin videota, joka on sopivan lyhyt ja pitää prosessin ja käsittelyn riittävän lyhyenä. Kun työnkulku on todettu toimivaksi, voidaan käyttöä laajentaa myös reilusti pidempiin, esimerkiksi tunnin mittaisiin videoihin. (Knime, n.d.-a)

Kun kuvien luku on valmis, ne voidaan järjestää nimen mukaan. Tämä varmistaa, että kuvat ovat kronologisesti oikeassa järjestyksessä ja prosessi toimii oikein. Epäkronologisessa järjestyksessä olevat kuvat johtaisivat vääriin tuloksiin, koska samasta autosta tehtäisiin useita havaintoja. Tämän Sorter-noodin jälkeen työnkulku jakautuu kahteen eri osaan ajoneuvojen segmentointiin ja lähdevideon luontiin yksittäisistä kuvista. Segmentoinnissa yksittäisiä kuvia käsitellään siten, että siitä voidaan erottaa liikkuva objekti eli tässä tapauksessa auto. Samalla myös liian pienet kuvassa liikkuvat asiat poistetaan väärien havaintojen välttämiseksi. Videotiedosto muuttuukin tämän myötä liikkuviksi erimuotoisiksi ja erivärisiksi elementeiksi. Nämä elementit lasketaan yhteen vasta lopuksi. Lähdevideon luonnin haarassa taas luodaan alkuperäisistä kuvista yksittäinen koostettu kuvatiedosto ja

tähän kuvatiedostoon määritellään alue, jota tarkastellaan. Tarkasteltava alue on videossa oikeanpuoleiset kolme kaistaa ilman kiihdytyskaistaa. Tätä aluetta kuvaa vaaleansininen väritys. (Knime, n.d.-a)

Kuva 10 Auton havainto on merkitty violetilla ja vaaleansinisellä havaintoalue.

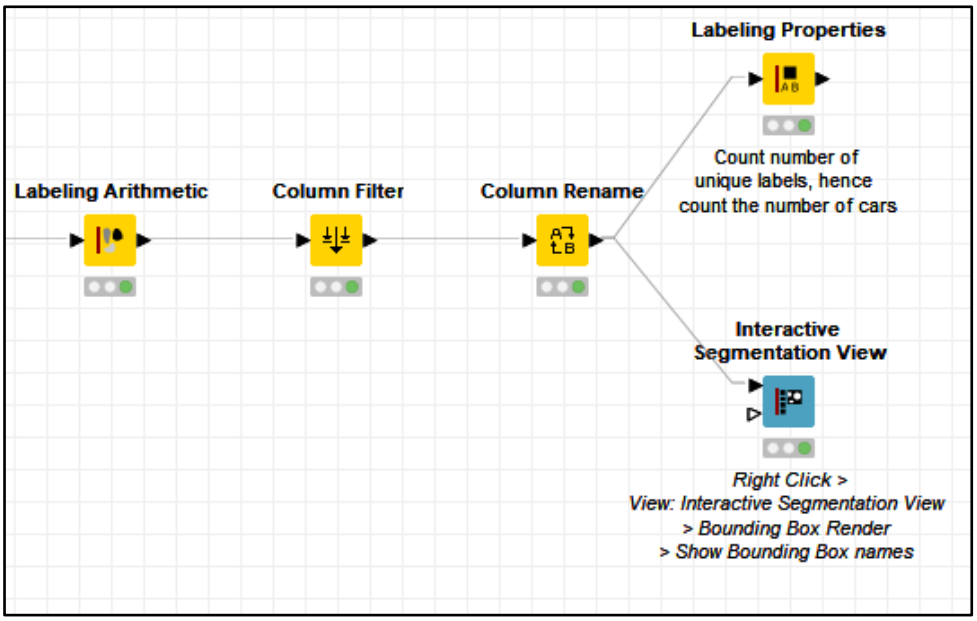
Kuva 11 Autojen havaitsemisen työnkulun ensimmäinen osa



Segmentaation ja videonvalmistelun haara yhdistyvät Cross Joiner -noodin avulla, jossa kaikki kolme eri kuvasarjaa ovat yhdessä taulussa. Autoja havaitsevat labelit ovat ”Tracking” taulussa, kevyesti käsitellyt kuvat ”Merged Img” taulussa ja havaintoaluetta rajaava label-kuva ”Merged Img_roi” taulussa. Nämä ovat Kuva 10. Yhdistämisen jälkeen tauluihin tehdään vielä eri labeleiden ja alkuperäisvideon yhdistämisiä. Näin saadaan aikaan video, jossa autoja havaitsevat labelit näkyvät videolla, jossa itse autotkin liikkuvat. Kun interaktiivinen video on valmis Labeling Properties -noodilla lasketaan yhteen havaitut autoja vastaavat elementit ja tulos tuodaan yksinkertaiseen tauluun, jossa ensimmäisessä

sarakkeessa on kuva-aineisto ja toisessa sarakkeessa uniikkien autojen lukumäärä. Yleiskuva työnkulusta KNIME-ohjelmistossa on esitelty kuvissa Kuva 11 jaKuva 12.(Knime, n.d.-a)

Kuva 12 Autojen havaitsemisen työnkulun toinen osa



Kuva 13 Video ja siihen yhdistetyt havaintoja kuvaavat labelit

Kuva 13 on itse video ja aiemmin luodut violetit ja keltaiset labelit on yhdistetty yhdeksi kuvakokonaisuudeksi. Tämä havainnollistaa hyvin lopullista työnkulun visuaalista tulosta. Autojen laskemisen voisi viedä vielä tätä työnkulkua pidemmälle. Esimerkiksi yhdistelmäajoneuvot ja muut suuret ajoneuvot voisi eritellä laskentaan omalle rivilleen. Analyysia voisi syventää myös siten, että tarkastelisi autojen määriä eri kaistoilla rajaamalla havaintoalueen kolmeen osaan joka kaistalle, yhden suuren havaintoalueen sijaan. Tässä

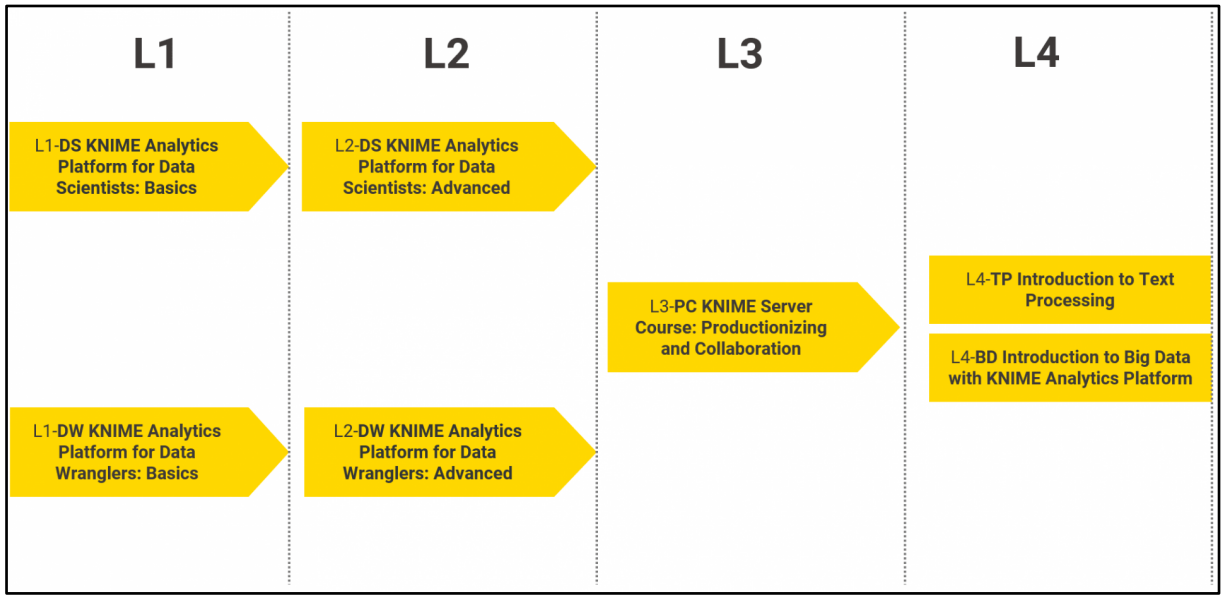
tulisi ottaa huomioon mahdolliset kaistavaihdot, jotka tapahtuvat tarkastelualueella. (Knime, n.d.-a)

Videon analysointi KNIME-ohjelmistossa on toimivaa ja tätä työnkulkua esimerkkinä hyödyntäen on mahdollista tehdä muita vastaavia laskutoimituksia videomateriaalin perusteella. Materiaalia on helpointa käsitellä silloin, kun kamera on paikallaan ja kuvausolosuhteet koko ajan identtiset. (Knime, n.d.-a)

5.3 KNIME-ohjelmiston koulutukset

KNIME-ohjelmistolle tarjotaan koulutuksia kaikkiaan neljässä luokassa (Kuva 14). Koulutukset toimivat verkossa KNIME-ohjelmiston sivustolla ja ovat englanninkielisiä. L1-tasolla koulutus on perustasoa, L2 syvennetään osaamista, L3 käyttöönottoon liittyvää koulutusta ja L4 tarkkaan joihinkin osa-alueisiin keskittyvää koulutusta. L1-tason koulutukset jaetaan kahteen eri alaluokkaan datatieteilijöille (data scientists) ja datakäsittelijöille (data wranglers). Datatieteilijöiden koulutus koostuu datan hankinnasta, siivoamisesta, visualisoinneista, koneoppimisesta ja käsitellyn datan tallentamisesta. Datankäsittelijöiden koulutus taas keskittyy enemmän datan siivoamiseen ja datan yhdistämiseen. (Knime, n.d.-e)

Kuva 14 KNIME-ohjelmiston koulutukset. (Knime, n.d.-e)



L2-kurssit syventävät L1 kursseilla hankittua osaamista. Datatieteilijöiden L2-kurssi koostuu työnkulun muuttujista, toistoista (loops) ja virhekäsittelystä. Lisäksi mm. päiväykseen ja aikaan liittyviä integraatioita käsitellään. Datakäsittelijöiden kurssilla on paljon samaa kuin datatietelijöiden kurssilla, mutta lisäksi mukana on datan käsittelyä ja luokitusmallin luontia. L3-tason kursseja on ainoastaan yksi. Koulutuksessa opitaan käsittelemään KNIME-palvelinta (KNIME server), verkkoportaalia (KNIME web portal), automaatioita ja ajastuksia. (Knime, n.d.-e)

L4-tason koulutukset keskittyvät tarkasti kahteen data-analytiikan osa-alueeseen. Ensimmäinen kursseista keskittyy Big dataan, tietokantojen käsittelyyn, Apache Hadoop -järjestelmään ja eri pilvipalveluiden integraatioihin. Toinen kursseista tarkastelee tekstin prosessointia ja tekstinlouhintaa (text mining). Kurssilla käsitellään mm. sentimenttianalyysia ja aihetunnistusta. (Knime, n.d.-e)

5.3.1 KNIME-ohjelmiston suoritettavissa olevat sertifikaatit

KNIME-ohjelmistoon liittyvä verkko-opiskelualusta tarjoaa koulutuksien lisäksi mahdollisuuden suorittaa maksullisia sertifikaatteja. Sertifikaatti on hinnaltaan 100 dollaria ja pysyy voimassa kaksi vuotta suorituksesta. Sertifikaatteja on tarjolla neljä. L1 Basic Proficiency in KNIME Analytics Platform -sertifikaatti on alimman tason sertifikaatti, joka perustuu kahteen L1-tason koulutukseen. Sertifikaattiin liittyy tentti ja se sisältää 15 monivalintakysymystä, joista on vastattava vähintään 11 kysymykseen oikein 30 minuutin sisällä. Tentti sisältää datan tuontia ja vientiä, yleisiä KNIME-ohjelmiston käyttötapauksia, datamanipulaatiota ja datan koostamista. Jos suoritus epäonnistuu, sertifikaattia voi yrittää uudelleen yhden kerran. Sertifikaatin suorituksessa ei ole sallittua käyttää apuvälineitä tai KNIME-ohjelmistoa. Kun sertifikaatti on suoritettu, saa siitä digitaalisen tunnuksen sekä PDF-muotoisen todistuksen. (Knime, n.d.-c)

L2-tason sertifikaatti L2: Advanced Proficiency in KNIME Analytics Platform perustuu vastaaviin L2-tason koulutuksiin. Se sisältää muuttujia, konfigurointinoodeja, aika- ja päivämäärätietojen käsittelyä, työnkulun hallintaa ja datan visualisointia. Myös tämän sertifikaatin tentti koostuu 15 kysymyksestä ja hyväksytty suoritus edellyttää 11 kysymykseen oikein vastaamista. L3-tason sertifikaatti L3: Proficency in KNIME Software for

Collaboration and Productionizing of Data Science perustuu L3-tason koulutuksiin. Se sisältää työnkulkujen käyttämistä sovelluksissa ja palveluissa, työnkulkujen etäkäyttö, raporttien luontia, tietokantojen käyttöä ja KNIME verkkoportaalin käyttöä. Myös tämän sertifikaatin tentti on laajuudeltaan 15 kysymystä. (Knime, n.d.-c)

Neljäs sertifikaatti on nimeltään ”KNIME Server Administration”. Tämä on erikoisempi KNIME serveriin pääasiassa keskittyvä kokonaisuus ja opiskelua varten koulutusportaalissa on tarjolla vain oppaita verkkokoulutuksen sijaan. Sertifikaatin tentti koostuu 50 tenttikysymyksestä ja aikaraja tentissä on 90 minuuttia. Se koostuu yleisistä KNIME-ohjelmiston asioista sekä mm. KNIME serverin autentikointeihin, lokeihin, turvallisuuteen ja arkkitehtuuriin. Tentin läpäisemiseksi on 70 prosenttiin kysymyksistä vastattava oikein. (Knime, n.d.-c)

5.3.2 Ulkopuoliset koulutukset

KNIME-ohjelmiston verkkosivuilla on saatavilla myös ohjelmiston ulkopuolisten tahojen järjestämiä kursseja. Pääosin kurssit keskittyvät yhteen tai muutamaan aiheeseen ohjelmistoon liittyen. Tarjolla on datan valmisteluun keskittyvä kurssi, data analyysia ja koneoppimista. Osa kursseista on vasta-alkajille ja osa jo pidempään dataa analysoineille henkilöille. Kurssit ovat logiikaltaan hyvin samankaltaisia kuin KNIME-ohjelmiston omat kurssit, eli kurssit koostuvat videoista ja harjoituksista. Kurssit ovat osittain maksullisia ja vuonna 2022 esimerkiksi viiden tunnin mittainen data-analyysin ja koneoppimisen kurssi maksoi 15 euroa. (Knime, n.d.-b)

Verkkokurssien lisäksi tarjolla on myös yliopistojen järjestämiä data-analyysiin, ennustavaan analytiikkaan ja liiketoiminnan analytiikkaan liittyviä kursseja. Kursseja tarjoaa mm. Readingin ja Oklahoman yliopistot. Nämä kurssit hyödyntävät osana oppimista myös KNIME-ohjelmistoa. (Knime, n.d.-b)

6 KNIME-ohjelmisto ja koneoppiminen käytännössä

KNIME-ohjelmisto on hyödynnettävissä monella tavalla koneoppimisessa. Ohjelmistolla voidaan laskea elementtejä kuvasta tai videosta, tehdä ratkaisuja aineiston perusteella, hankkia dataa tietokannasta tai esimerkiksi Twitterin rajapinnasta sekä organisoida, muuntaa ja hallita dokumentteja. Tämän lisäksi ohjelmistossa on paljon muitakin mahdollisuuksia. Ohjelmiston käyttö ei vaadi koodausta, vaan perustuu visuaaliseen käyttöliittymään. Siten käyttäjän taustasta riippumatta ohjelmistoon on mahdollista tutustua matalalla kynnyksellä.

KNIME Hubissa on saatavilla lukuisia esimerkkejä ja valmiita ratkaisuja erilaisten datatyyppien kanssa työskentelevien henkilöiden tarpeisiin. Esimerkiksi työnkulku, joka pystyy päättelemään eläimestä otetusta kuvasta, onko siinä kissa vai koira, löytyy KNIME Hubista tausta-aineistoineen. Tausta-aineisto tässä esimerkissä koostuu yli 25 000 eläinkuvasta. Koneoppimisen soveltaminen ohjelmistolla edellyttääkin aina riittävää tausta-aineistoa, jotta tulos on riittävän hyvä.

6.1 RMS Titanicin matkustajien selviytymisen analysointi

Selviytymisen analyysia alettiin tekemään 887 rivisellä CSV-muotoisella lähdeaineistolla, jossa on luetteloitu matkustajia nimellä. Aineiston tarjoaa Stanfordin yliopisto ja se on saatavilla osoitteesta https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/problem12.html. Data vietiin sellaisenaan KNIME-ohjelmistoon ilman ennakkokäsittelyä. Aineiston sarakkeissa on nimien lisäksi tieto selviytymisestä (0 ei selviytynyt tai 1 selviytyi), matkustajaluokka (1,2 tai 3), sukupuoli, ikä, matkalla olleiden sisarusten määrä, matkalla olleiden omien vanhempien tai lapsien lukumäärä sekä matkalipun hinta.

Käytin analysoinnissa kolmea erilaista koneoppimisen mallia. Satunnaisen metsän mallia, päätöspuumallia sekä lineaarista regressiomallia. Malleissa on paljon yhteneväisyyksiä, mutta koneoppimisen menetelmät ja KNIME-ohjelmistossa käytettävät noodit vaihtelevat jokaisessa mallissa. Jokaisessa mallissa kaikki alkaa CSV Reader -noodilla, joka lukee työnkulkuun lähdeaineiston sisällön ja luo siitä taulun. Tämä taulu on esitelty Kuva 15.

Kuva 15 CSV Reader -noodin tuottama taulu CSV-tiedostosta

Table "default" - Rows: 887 | Spec - Columns: 8 | Properties | Flow Variables

Row ID	Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
Row0	0	3	Mr. Owen Harris Braund	male	22	1	0	7.25
Row1	1	1	Mrs. John Bradley (Flo...	female	38	1	0	71.283
Row2	1	3	Miss. Laina Heikkinen	female	26	0	0	7.925
Row3	1	1	Mrs. Jacques Heath (L...	female	35	1	0	53.1
Row4	0	3	Mr. William Henry Allen	male	35	0	0	8.05
Row5	0	3	Mr. James Moran	male	27	0	0	8.458
Row6	0	1	Mr. Timothy J McCarthy	male	54	0	0	51.862
Row7	0	3	Master. Gosta Leonar...	male	2	3	1	21.075
Row8	1	3	Mrs. Oscar W (Elisabe...	female	27	0	2	11.133
Row9	1	2	Mrs. Nicholas (Adele A...	female	14	1	0	30.071
Row10	1	3	Miss. Marguerite Rut ...	female	4	1	1	16.7
Row11	1	1	Miss. Elizabeth Bonnell	female	58	0	0	26.55
Row12	0	3	Mr. William Henry Sau...	male	20	0	0	8.05
Row13	0	3	Mr. Anders Johan And...	male	39	1	5	31.275
Row14	0	3	Miss. Hulda Amanda A...	female	14	0	0	7.854

6.1.1 Satunnaisen metsän malli

Satunnaisen metsän mallissa CSV-tiedoston lukemisen jälkeen käydään lista läpi Row Filter -noodin avulla ja poistetaan mahdolliset tyhjät kentät. Tämä ei tässä aineistossa olisi tarpeen, koska tyhjiä kenttiä ei aineistossa ollut. Tämä on kuitenkin hyvä käytäntö datan valmistelussa analyysiin. Tämän jälkeen ositetaan aineisto kahteen osaan siten, että 70 % aineistosta menee lähdeaineistoksi koneoppimisen noodille ja 30 % jää koneoppivalle mallille käsiteltäväksi. Kun tämä on tehty, työnkulku jakautuu kahtia. Random Forest Learner (Regression) -noodin avulla käsitellään tuo 70 % aineistosta oppimistarkoituksessa. Loppu aineisto menee käsittelyyn Tree Ensemble Predictor (Regression) -noodiin, jossa aiemmin tehdyn koneoppimisen mallin avulla luodaan ennuste henkilöiden mahdollisuudesta selvitä (Kuva 16). Lisäksi verrataan ennustetta varsinaiseen lähdeaineiston totuuteen. Näin mallin onnistumista voidaan arvioida.

Kuva 16 Tree Ensemble Predictor -noodin avulla luotu taulu

Row ID	Survived	Pclass	Name	Sex	Age	Siblings...	Parents...	Fare	SurvivalPrediction	SurvivalPrediction (Prediction Variance)
Row10	1	3	Miss. Marguerite Rut...	female	4	1	1	16.7	0.693	0.163
Row101	0	1	Mr. Richard Frasar ...	male	21	0	1	77.287	0.542	0.185
Row103	0	3	Mr. Anders Vilhelm G...	male	37	2	0	7.925	0.115	0.057
Row106	1	3	Mr. Albert Johan Moss	male	29	0	0	7.775	0.21	0.06
Row109	0	1	Mr. Walter Chamberl...	male	47	0	0	52	0.256	0.051
Row112	0	3	Miss. Katriina Jussila	female	20	1	0	9.825	0.277	0.094
Row114	0	3	Mr. Edvard Pekoniemi	male	21	0	0	7.925	0.171	0.051

Luotu taulu vastaa alkuperäistä taulua, mutta siinä on vain koneoppimisen mallilla luodut rivit ja lisäksi uusina sarakkeina ovat selviytymisen ennuste (SurvivalPrediction) ja selviytymisen ennusteen varianssi (SurvivalPrediction Prediction Variance). Luodun taulun avulla nähdään jo esimerkiksi se, että naisten todennäköisyys selviytyä oli korkeampi kuin miesten. Samoin erityisesti miesten kohdalla matkustajaluokka vaikutti merkittävästi

selviytymismahdollisuuksiin. Tästä taulusta voitaisiin myös tehdä erillinen taulu esimerkiksi Row Filter -noodilla, jossa olisivat vain yli 50 dollaria lipustaan maksaneet tai pelkästään toinen sukupuoli tarkempaa analyysia ajatellen.

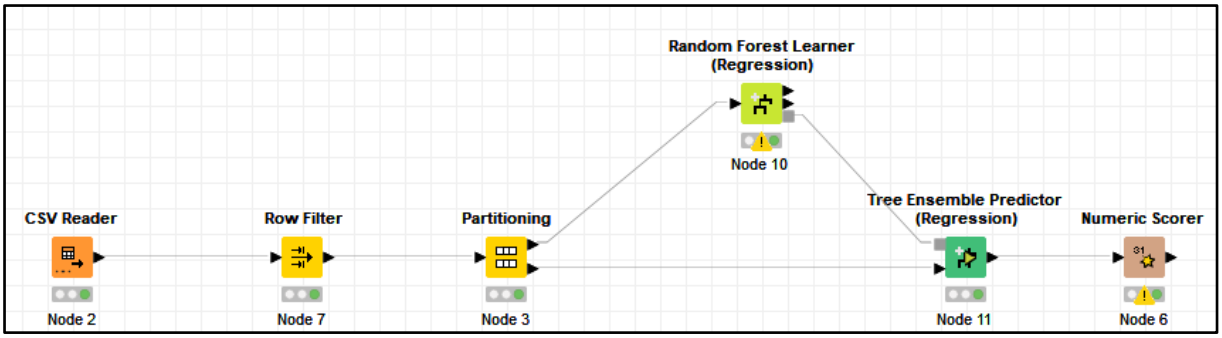
Viimeisenä tehdään vielä numeerinen arvio mallista Numeric Scorer-noodin avulla. Noodin asetuksissa määritellään oikean tiedon sisältävä sarake, sekä koneoppimisen mallilla luotu sarake. Tällä noodilla voidaan arvioida luodun mallin onnistumista laskennallisesti. Noodin tulokset ovat Kuva 17. Ensimmäinen luku on R^2, joka arvioi, kuinka eri muuttujat, kuten matkustajaluokka ja sukupuoli, kertovat selviytymisestä. Tässä tapauksessa luku on 0,492 eli muut tekijät selittävät noin 49 % variaatiosta. Muut taulukon luvut kertovat virheen koosta ja suunnasta. Mean absolute error -sarake kertoo keskimääräisen eroavaisuuden ja ”Mean squared error” on enemmän mallin optimointiin tarvittava luku. ”Root mean squared error” on hieman korkeampi kuin ”mean absolute error” koska sen laskennassa painotetaan enemmän suuria eroavaisuuksia. ”Mean signed difference” kertoo, mihin mallin virheet suuntautuvat, joko yli tai ali todellisen arvon. Koska luku on tässä tapauksessa 0,001, suuntautuvat mallin arviot korkeammiksi verrattuna todelliseen arvoon. Viimeinen arvo on ”mean absolute percentage error” ja se kertoo prosentuaalisesti keskimääräisen eron todelliseen tilanteeseen. Tässä kyseisessä mallissa lukua ei ole, koska osa arvoista on nollia, eikä nollilla voida jakaa. Työnkulku kokonaisuudessaan on Kuva 18. (Knime, n.d.-d)

Kuva 17 Numeric Scorer-noodin tulokset.

Table "Scores" - Rows: 6 | Spec - Column: 1 | Properties | Flow Variables

Row ID	D SurvivedNumeric
R^2	0.492
mean absolute error	0.252
mean squared error	0.119
root mean squared error	0.344
mean signed difference	0.001
mean absolute percentage error	NaN

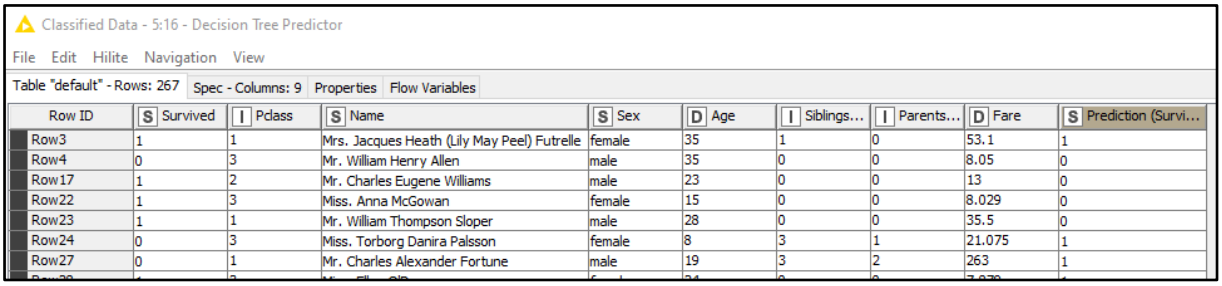
Kuva 18 Satunnaisen metsän mallin työnkulku KNIME-ohjelmassa.



6.1.2 Päätöspuumalli

Päätöspuumallissa CSV reader -noodin jälkeen muutetaan selviytymissarakkeen numeroarvot merkkijonoksi (string). Tämä on tehtävä, koska päätöspuumallin noodi edellyttää merkkijonoa toimiakseen. Satunnaisen metsän mallia mukaillen tehdään myös samanlainen aineiston ositus, jossa 70 % jää mallia varten ja 30 % analysoidaan päätöspuumallilla. Tämän jälkeen siirretään tiedot oikeille noodeille, eli Decision Tree Learner -noodiin mallia varten tarkoitettu aineisto ja Decision Tree Predictor -noodille mallilla analysoitava aineisto. Tämän noodin tulokset ovat Kuva 19.

Kuva 19 Päätöspuumallin Decision Tree Predictor -noodin tulokset taulukkona.


Classified Data - 5:16 - Decision Tree Predictor

File Edit Hilite Navigation View

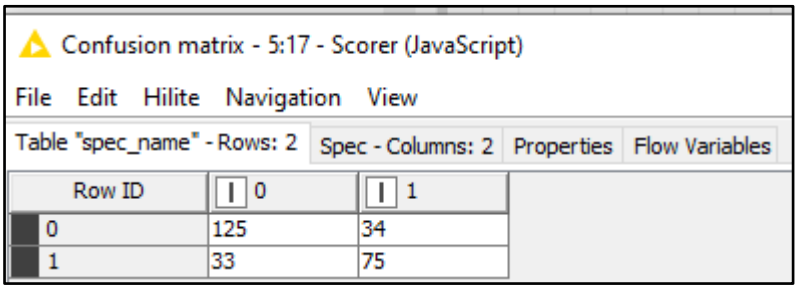
Table "default" - Rows: 267 | Spec - Columns: 9 | Properties | Flow Variables

Row ID	S Survived	I Pclass	S Name	S Sex	D Age	I Siblings...	I Parents...	D Fare	S Prediction (Survi...
Row3	1	1	Mrs. Jacques Heath (Lily May Peel) Futrelle	female	35	1	0	53.1	1
Row4	0	3	Mr. William Henry Allen	male	35	0	0	8.05	0
Row17	1	2	Mr. Charles Eugene Williams	male	23	0	0	13	0
Row22	1	3	Miss. Anna McGowan	female	15	0	0	8.029	0
Row23	1	1	Mr. William Thompson Sloper	male	28	0	0	35.5	0
Row24	0	3	Miss. Torborg Danira Palsson	female	8	3	1	21.075	1
Row27	0	1	Mr. Charles Alexander Fortune	male	19	3	2	263	1

Kun työnkulku on suoritettu, voidaan katsoa koneoppimisen mallilla saatua tulosta ensin viimeisen Decision Tree Predictor -noodin taulukkona. Taulukosta voidaan heti havaita, että selviytymisen ennuste on tässä selkeämmin kokonaisluku 1 tai 0 verrattuna aiempaan satunnaisen metsän mallin desimaalilukuun. Selviytymissarakkeen muuttaminen merkkijonoksi on merkittävin syy tälle. Toiseksi viimeisenä noodina tässä työnkulussa on Scorer-noodi, jossa siis arvioidaan mallin toimivuutta. Tämä noodi poikkeaa melkoisesti

aiemmasta pisteyttävästä noodista. Noodi tuottaa kaikkiaan kolme erilaista taulukkoa: Confusion matrix, class statistics table ja overall statistics table. Näiden lisäksi on tarjolla visuaalisempi interaktiivinen Confusion matrix -näkymä. Confusion matrix -taulukko (Kuva 20) näyttää yksinkertaisesti sen, kuinka monta kertaa malli osui oikeaan tai väärään yksinkertaisilla ”0” ja ”1” riveillä ja sarakkeilla. Jos sarake (ennustearvo) ja rowID (todellinen arvo) ovat samat, niin malli toimi oikein, jos taas eri, malli toimii väärin.

Kuva 20 Confusion matrix -taulukko.



Class Statistics table -taulukko tarjoaa monipuolisemmin tietoa (Kuva 21). Taulukossa käydään kahden eri vaihtoehdon (selviytyi tai ei selviytynyt) oikeat positiiviset (true positives), väärät positiiviset (false positives), oikeat negatiiviset (true negatives) ja väärät negatiiviset (false negatives) tulokset lukumäärinä. Oikea positiivinen tarkoittaa oikein tehtyä positiivista arviota, väärä positiivinen väärin tehtyä positiivista arviota, oikea negatiivinen oikein tehtyä negatiivista arviota ja väärä negatiivinen taas väärää negatiivista arviota. Malli on siis toiminut virheellisesti niiden kohdalla, jotka on arvioinut väärin negatiivisesti ja väärin positiivisesti. Oikeat positiiviset ja oikeat negatiiviset ovat yhteensä 200, joka vastaa oikein arvioitujen määrää. Väärät positiiviset ja väärät negatiiviset taas ovat yhteenlaskettuna 67, joka vastaa väärin arvioitujen kokonaismäärää.

Seuraava sarake taulukossa on recall, joka kertoo kuinka hyvin malli löytää positiiviset tulokset. Selviytymättömyyden osalta (rivi 0) tulos oli parempi n. 79 %, kun taas selviytymisen (rivi 1) kohdalla tulos oli hieman yli 69 prosenttia. Precision kertoo mallin tarkkuudesta merkitä positiiviset tapaukset positiivisiksi. Tässä prosenttiluvut ovat liki identtiset recall-sarakkeeseen verrattuna. Sensitivity -sarake kertoo mallin herkkyydestä tunnistaa positiiviset tapaukset ja merkata ne positiiviseen sarakkeeseen. Myös tässä sarakkeessa prosentit ovat hyvin samankaltaiset ja täysin identtiset recall-sarakkeen kanssa.

Specifity-sarake kertoo mallin tarkkuudesta ja tässä syntyy eroa aiempiin sarakkeisiin. Malli on tarkempi määrittelemään selviytymisen noin 79 prosenttisesti, kun taas selviytymättömyyden tarkkuus on noin 69 prosenttia. F-measure-sarake kertoo harmonisoidun keskiarvon recall- ja precision-sarakkeista. Tämä vastaa hyvin paljon aiempia recall-, precision- ja sensitivity-sarakkeita ja arvot ovat selviytymättömyydelle 79 % ja selviytymiselle 69 %.

Kuva 21 Class statistics table -taulukko.

Class statistics table - 5:17 - Scorer (JavaScript)

File Edit Hilite Navigation View

Table "default" - Rows: 2 | Spec - Columns: 9 | Properties | Flow Variables

Row ID	True Positives	False Pos...	True Ne...	False Ne...	Recall	Precision	Sensitivity	Specificity	F-measure
0	125	33	75	34	0.786	0.791	0.786	0.694	0.789
1	75	34	125	33	0.694	0.688	0.694	0.786	0.691

Viimeinen Scorer-noodista saatava taulukko on overall statistics table (Kuva 22). Sen avulla näkee mallin statistiikkaa kokonaisuutena. Ensimmäinen sarake overall accuracy kertoo mallin tarkkuuden. Tarkkuuden arvo on 0,749 eli 74,9 prosenttisesti malli on osunut arvioissaan oikeaan. Vastaava luku on overall error, joka kertoo virheiden osuuden. Virheiden osuus oli 0,251 eli 25,1 prosenttia. Cohen’s kappa (teoriaa tulossa teoriaosuuteen) kertoo mallin tarkkuudesta -1 ja 1 välillä. Tässä mallissa Cohen’s kappa oli 0,48 eli malli ei tällä perusteella ole kovinkaan tarkka. Jos luku on lähellä yhtä, on malli tarkka. Jos luku on negatiivinen, satunnainen arvonta on mallia tarkempi. Viimeiset kaksi saraketta kertovat mallin oikein ja väärin arvioidut rivit. Tässä mallissa oikeita arvioita oli 200 ja vääriä 67.

Kuva 22 Overall statistics table -taulukko.

Overall statistics table - 5:17 - Scorer (JavaScript) (Mallin arviointi)

File Edit Hilite Navigation View

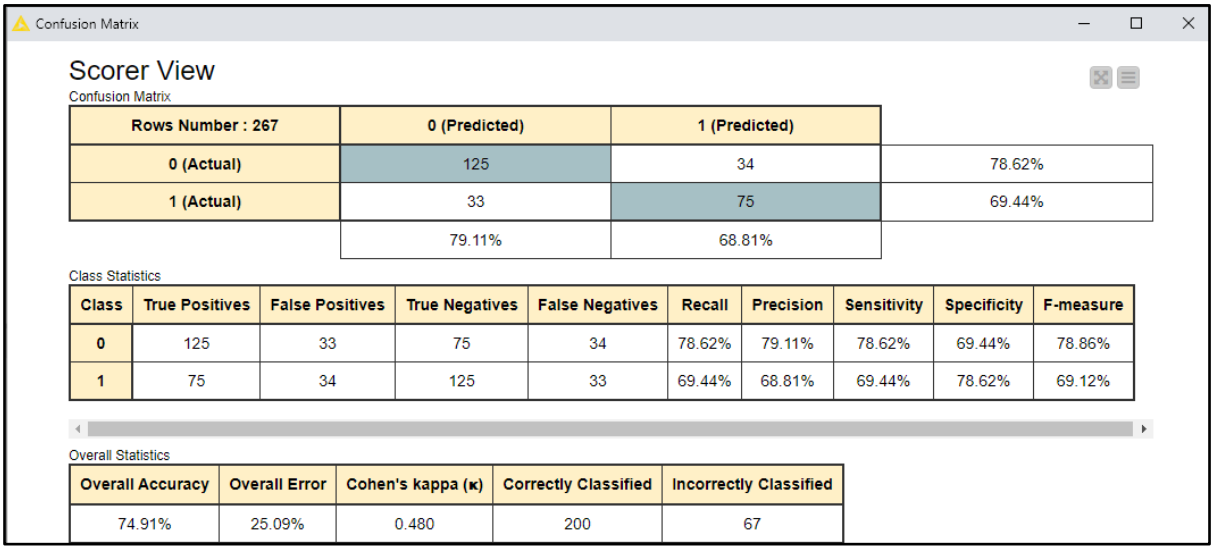
Table "default" - Rows: 1 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	Overall Accuracy	Overall Error	Cohen's kappa	Correctly Classified	Incorrectly Classified
Overall	0.749	0.251	0.48	200	67

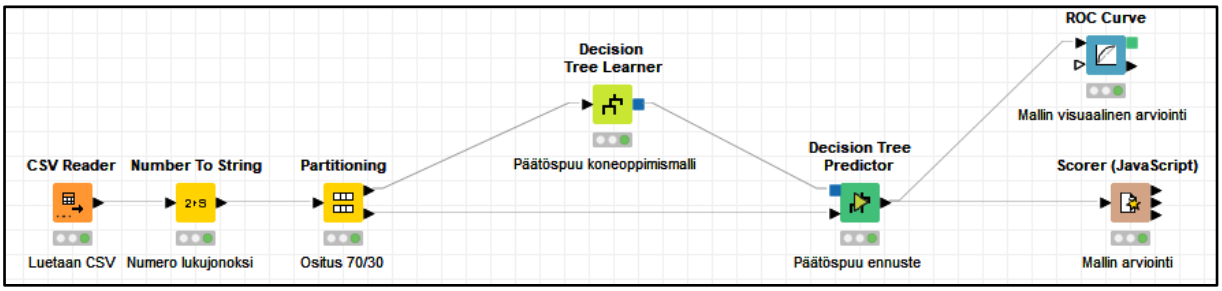
Aiemmin esiteltyjen taulukoiden lisäksi KNIME-ohjelmisto tarjoaa myös visuaalisemman taulukon mallin arviointiin. Tämän taulukon nimi on ”Interactive view: Confusion matrix.” Taulukko on esitelty Kuva 23. Taulukossa on kaikki tiedot samassa visuaalisemmassa muodossa kuin muissa, jo esitellyissä taulukoissa. Lisäksi kaikki luvut on muutettu valmiiksi

prosentuaaliseen muotoon ja taulukon näkymä on kustomoitavissa näytettävien taulukoiden ja yksittäisten tietojen osalta. Koko päätöspuumallin työnkulku on nähtävissä Kuva 24.

Kuva 23 Interactive view: Confusion matrix -visuaalinen taulukko.



Kuva 24 Päätöspuumallin työnkulku KNIME-ohjelmassa.



6.1.3 Lineaarinen regressiomalli

Kolmas käytetty koneoppimisen malli on lineaarinen regressiomalli. Malli on hyvin samankaltainen satunnaisen metsän mallin kanssa. CSV-tiedoston lukemisen jälkeen poistetaan mahdolliset tyhjät rivit ja ositetaan aineisto. Tässä mallissa käytetään ositusta siten, että 80 % aineistosta menee mallin luomiseen ja 20 % analysoidaan mallilla. Käytettävä noodi on nimeltään Linear regression learner. Analyysin tekevä noodi taas nimeltään Regression predictor. Kuten satunnaisen metsän malli, myös tämä tuottaa ennusteen tarkkana desimaalilukuna. Ennen mallina arvioivaa Numeric scorer-noodia, lisäsin tauluun

yhden sarakkeen Round Double -noodilla, jossa pyöristetään ennusteen luvut joko ykköseen tai nollaan. Jos selviytymisen ennuste on yli 0,500, pyöristetään luku yhteen. Jos taas alle 0,5, pyöristetään luku nollaan.

Tulokset lineaarisessa regressiomallissa ovat selkeästi malleista heikoimmat. Varsinkin pyöristettyjä lukuja käytettäessä malli ei vaikuta toimivan lainkaan R^2 luvun ollessa 0,071 eli muut tekijät selittäisivät tässä mallissa vain 7 % selviytymisestä. Tämän taustalla oli voimakas lukujen pyöristäminen. Sen sijaan todelliset mallin tuottamat luvut analysoiva Scorer-noodi (Kuva 25) antaa R^2 -luvuksi 0,361, eli muut tekijät selittäisivät noin 36 % selviytymisestä. Pyöristämättömien lukujen keskimääräinen virhe on 30 % ja mallin tekemät arviot ovat mean signed difference -luvun perusteella pienempiä suhteessa survived-sarakkeen todelliseen tietoon. Myöskään tässä mallissa ”mean absolute percentage error” arvo ei toiminut johtuen arvoista, jotka ovat nolla. Koko lineaarisen regressiomallin työnkulku on esitelty Kuva 26.

Kuva 25 Lineaarisen regressiomallin arvioinnin avainluvut.

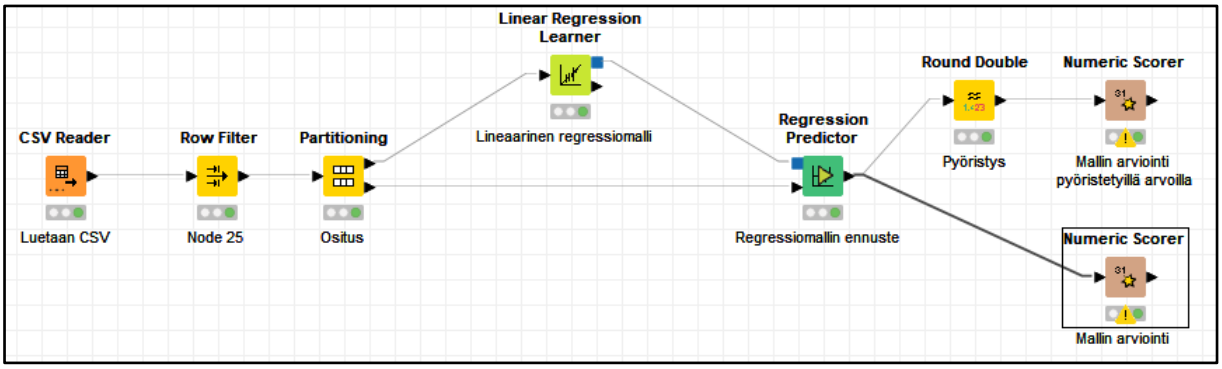
Statistics - 5:27 - Numeric Scorer (Mallin arviointi)

File Edit Hilite Navigation View

Table "Scores" - Rows: 6 | Spec - Column: 1 | Properties | Flow Variable

Row ID	D Predicti...
R^2	0.361
mean absolute error	0.306
mean squared error	0.154
root mean squared error	0.393
mean signed difference	-0.044
mean absolute percentage error	NaN

Kuva 26 Lineaarinen regressiomallin työnkulku KNIME-ohjelmassa

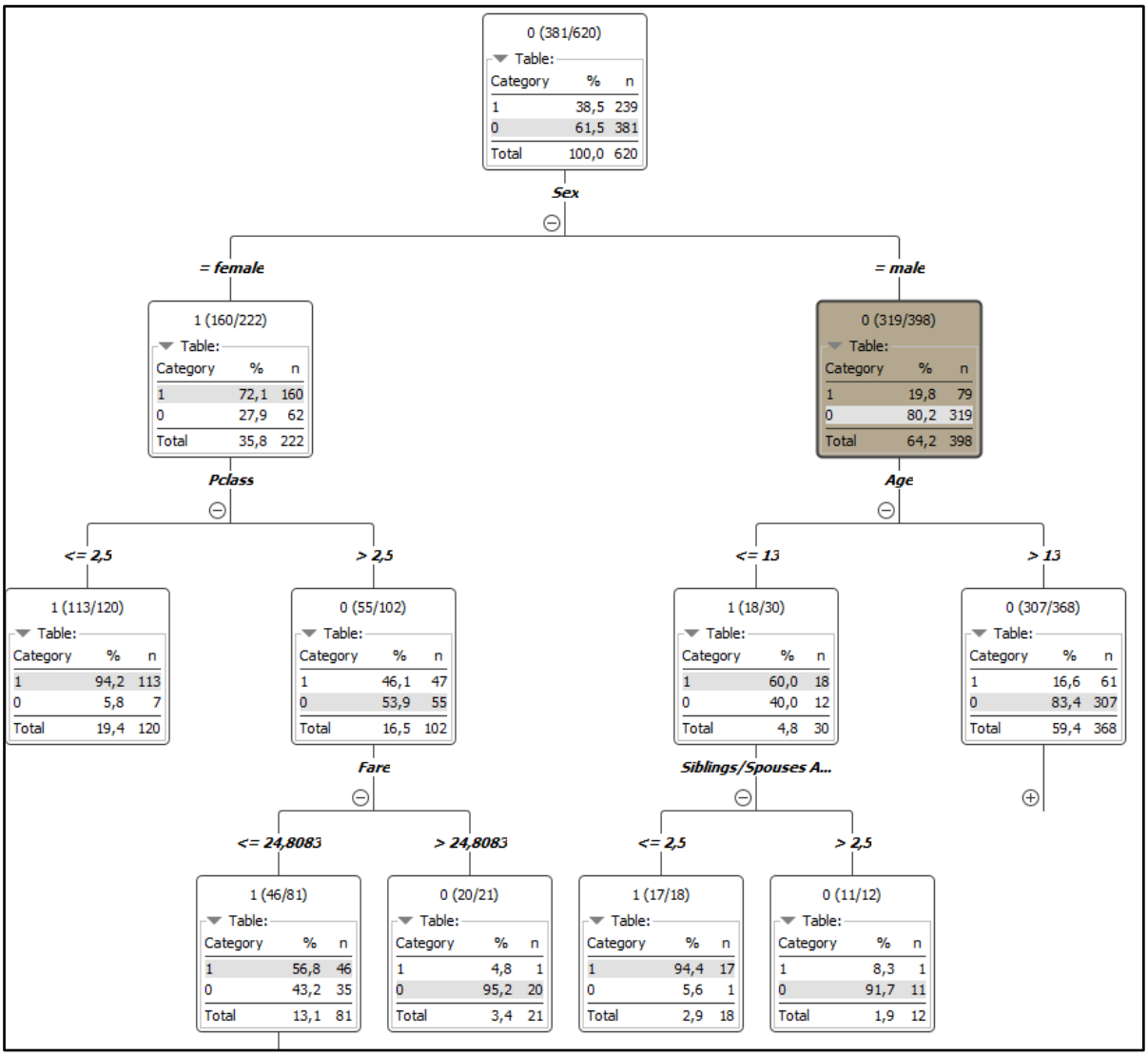


6.1.4 Kolmen koneoppimisen mallin vertailu

Kolmessa erilaisessa mallissa oli paljon samankaltaisuuksia, mutta jokainen oli erilainen tuloksiltaan. Tarkoituksenmukaisin malli tähän Titanicin matkustajista ja heidän selviytymisestään kertovaan aineistoon oli tulosten näkökulmasta päätöspuumalli. Scorer-noodin tulosten perusteella malli myös toimi parhaiten, sillä sen tarkkuus oli 74,9 prosenttia ja malli määritti 267 rivin analysoitavasta aineistosta kaikkiaan 200 tapausta oikein. Mikäli osituksen tekevässä noodissa muutetaan suhdetta 80 % mallille ja 20 % analyysiin, nousee mallin tarkkuus 79,8 prosenttiin. Kun suhde on 90/10 laskee mallin tarkkuus 71,9 prosenttiin. Tällöin myös analysoitava aineisto on todella pieni ja yksittäinen poikkeama laskee tarkkuutta merkittävästi. Myös osituksen uusiminen muuttaa tulosta erityisesti silloin, kun analyysiin menevä osuus aineistosta on pieni. Ositus tehtiin kaikissa malleissa työnkulun satunnaisesti valitsemilla riveillä koneoppivalle mallille ja analyysiin.

Päätöspuumallin varsinainen visuaalinen malli on mahdollista todentaa KNIME-ohjelmassa. Kuva 27 on nähtävissä pieni osa päätöspuumallista. Kokonainen kuva päätöspuumallista on tässä yli 800 rivin aineistossa todella laaja. Päätöspuumallissa arvioidaan kaikki arvioitavaksi asetetut ehdot. Tässä tapauksessa jokainen eri sarake oli mukana analyysissa. Myös matkustajan nimi, vaikkei tällä olekaan mallin kannalta merkitystä. Päätöspuumallissa malli kategorisoi matkustajia kaikkien eri datasta saatavien tietojen perusteella mahdollisimman tarkoituksenmukaisesti. Mallissa erotellaan esimerkiksi ne matkustajat, joiden matkustusluokka on yli 2,5 ja ne, joiden matkustusluokka on alle 2,5. Päätöspuumalli käy läpi kaikki aineistossa olevat ehdot luodakseen mahdollisimman tarkan mallin. Satunnaismetsän ja lineaarisen regressiomallin kohdalla matkustajan nimi otettiin mallista pois, sillä KNIME-ohjelmisto huomautti tästä Learner-noodeissa.

Kuva 27 Pieni osa päätöspuumallista visuaalisessa muodossa.



Ensimmäiseksi käytetty malli oli satunnaisen metsän malli, joka oli R^2 luvun 0,492 perusteella malleista toiseksi paras. Keskimääräinen ero todelliseen selviytymissarakkeeseen oli 25,2 prosenttia. Haastavinta mallissa oli se, että koneoppivan mallin tekemä arvio selviytymisestä on tarkka desimaaliluku. Heikoimmaksi malliksi pisteytyksen perusteella osoittautui lineaarisen regression malli, jossa R^2 luku oli 0,361. Viimeisessä mallissa ositus kuitenkin tehtiin 80:20 suhteessa, joten vertailun vuoksi testasin mallia vastaavalla suhteella kuin muita eli 70:30. Tämä siis tarkoittaa, että oppivalle mallille menee 70 % aineistosta ja 30 % aineistosta analysoidaan. Tämä kokeilu johti hieman parempaan tulokseen mallin osalta, sillä R^2 luku oli näin 0,423 ja siis selvästi aiempaa lähempänä satunnaisen metsän mallin tulosta. Koska aineisto on kaikissa malleissa koneoppimisen näkökulmasta hyvin pieni, yksittäiset poikkeamat aineistossa vaikuttavat tuloksiin. Esimerkiksi malli todennäköisimmin luokittelee

kolmannen luokan miesmatkustajan selviytymisen nollaan tai lähelle sitä. Tämänkaltaisia poikkeamia aineistosta löytyy, joissa kolmannen luokan matkustaja selviytyi.

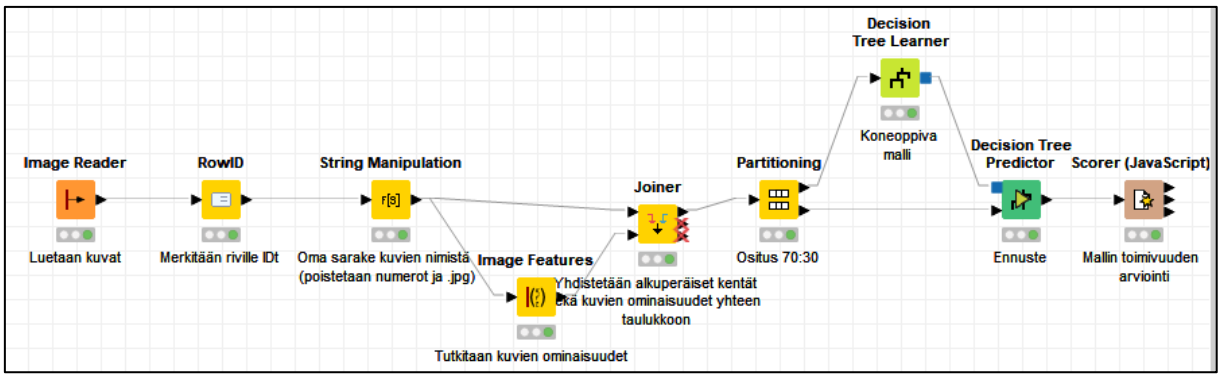
6.2 Valokuvien käsittely koneoppimisen keinoin

Valitsin käyttööni neljää erilaista säätilaa edustavan kuvasarjan, jonka on alun perin luonut Etelä-Afrikan yliopisto. Kuvista koostuva datasarja on saatavissa osoitteesta https://data.mendeley.com/datasets/4drtyfjtfy/1. Kuvia on aineistossa mukana kaikkiaan 1123 neljässä eri kategoriassa: pilvinen, sateinen, auringonpaiste ja auringonnousu. Työnkulku on toteutettu kahdella eri tavalla: alkuperäisten kuvien analyysilla Image features -noodin avulla ja toisessa työnkulussa taas analysoiden myöhemmin esiteltävillä noodeilla muokattuja kuvia. Image features -noodi tuo kuvien erilaisia ominaisuuksia, kuten minimi-, maksimi- ja vaihteluarvoja numeroiksi. Näiden perusteella voi koneoppiva päätöspuumalli tehdä päätelmiä siitä, mihin neljästä ryhmästä kuva kuuluu. Myös aiemmissa työnkuluissa käytetty Scorer-noodi on käytössä lopputulosten analysoimiseksi.

6.2.1 Muokkaamattomien kuvien kategorisointi

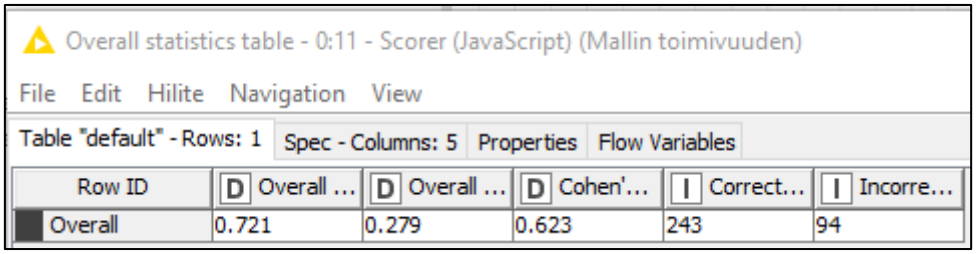
Kun kuvasarja käsitellään muokkaamattomana, työnkulku on hyvin yksinkertainen ja alkaa kuvat taulukkoon lukevasta Image reader -noodista, rivien ID:t lisäävästä RowID -noodista sekä kuvien kategoriat omaksi sarakkeekseen kirjaavasta String manipulation -noodista. Kun taulu on perusteiltaan valmis, voidaan tutkia kuvien ominaisuuksia Image features -noodin avulla. Tämä listaa kuvien ominaisuudet omaksi taulukseen. Tämä noodi on kaikkein kuormittavin tässä työnkulussa ja sen suorittaminen vei aikaa joitakin minuutteja. Varsinkin laajoja kuva- tai datasarjoja tutkittaessa tietokoneen tehokkuus on tärkeää, jotta aikaa työnkulun suoritukseen ei kuluisi kohtuuttomasti. Tässä vaiheessa tauluja on kaksi, joten alkuperäinen taulu kuvineen, sekä kuvien ominaisuudet listaava taulu tulee yhdistää Joiner-noodilla. Tämän myötä aineisto on valmis koneoppivaan malliin. Koko työnkulku on esitelty Kuva 28.

Kuva 28 Koko työnkulku kuvien kategorisoinnin koneoppivassa mallissa.



Koneoppiminen suoritetaan päätöspuumallina ja alkaa aineiston osituksesta. Ositus tehdään suhteessa 70:30, eli 70 % aineistosta menee koneoppivalle mallille aineistoksi ja 30 % jää analysoitavaksi. Tämän jälkeen luodaan mallin ennustenoodi (Decision tree predictor) ja mallin toimivuutta arvioiva Scorer-noodi (Kuva 29). Malli vaikuttaa Scorer-noodin kokonaisstatistiikan perusteella melko toimivalta, sillä kokonaistarkkuus on 0,721 eli 72,1 %. Kaikkiaan malli arvioi 243 kuvaa oikein ja 94 kuvaa väärin.

Kuva 29 Scorer-noodin tulokset koneoppivalle mallille.



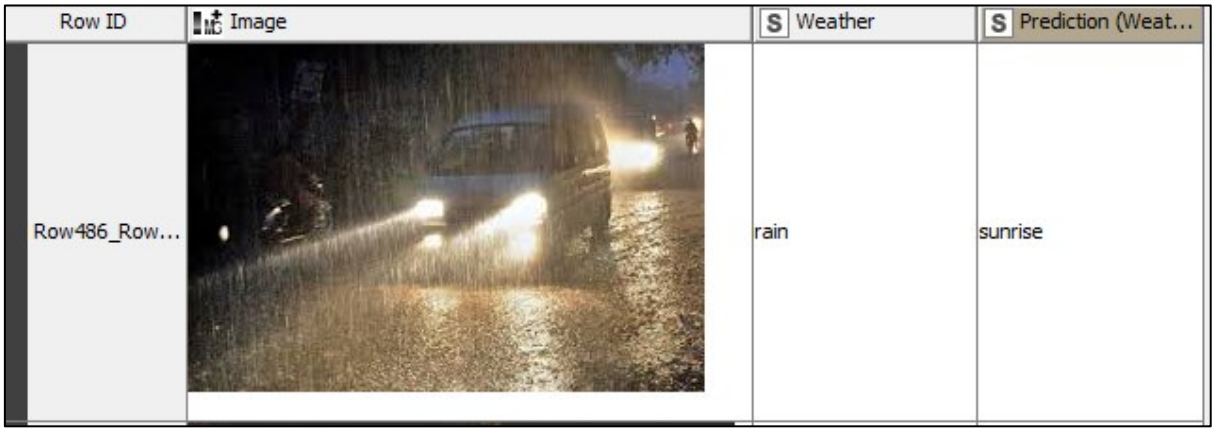
Row ID	Overall ...	Overall ...	Cohen'...	Correct...	Incorre...
Overall	0.721	0.279	0.623	243	94

Tein mallin suorittamisen jälkeen erillisen taulukon ennustenoodin perusteella käyttäen Column filter -noodia, jossa on todellinen arvo, ennuste ja itse kuva tarkempaa tutkimista varten. Kun tarkastelee erityisesti virheitä, osa virheistä tuntuu ymmärrettäviltä, esimerkiksi pilvinen kuva, jossa aurinko paistaa taustalla on merkitty kategoriaan auringonpaiste, mutta koneoppiva malli on tulkinnut kuvan pilviseksi. Myös Kuva 31, jossa auton kirkkaat ajovalot kajastavat sateessa, on malli tulkinnut auringonnousuksi, vaikka kuva on todellisuudessa kategorisoitu sateeksi. Kuvasarja monipuolisuudessaan onkin erinomainen myös todentamaan koneoppivien mallien ajoittaiset heikkoudet. Säätila ei aina ole yksiselitteinen ja kuvassa oleva häiriötekijä voi johtaa siihen, että mallin tekemä päätelmä on väärä.

Kirkkaasta säätilasta huolimatta taivaalla olevat pilvet vievät Kuva 30 suurimman osan kuvan alasta, joten mallin on helppo tehdä tulkinta pilvisestä säästä.

Kuva 30 Auringonpaisteeksi kategorisoitu kuva on mallin mukaan pilvinen.

Kuva 31 Ajovalojen kajastus sekoitti mallia tekemään väärän tulkinnan säätilasta.



Mallin kehittämisen kannalta erityisesti epäonnistuneet ennusteet ovat mielenkiintoisia, mutta on tärkeää tarkastella myös onnistumisia. Erityisesti geneerisemmät kuvat esimerkiksi rannalta, joissa aurinko paistaa pilvettömältä taivaalta ovat onnistuneet, mutta myös pilvisemmät kuvat on tulkittu oikein. Esimerkiksi kaksi aineistossa ollutta talvista auringonpaistetta esittävää kuvaa (Kuva 32) on tulkittu eri tavalla: toinen auringonpaisteeksi ja toinen pilviseksi. Talvisten puiden suurella osuudella kuvan alasta on mahdollisesti suuri vaikutus mallin tekemään päätelmään.

Kuva 32 Ihmiselle samankaltaiset kuvat on tulkittu koneoppivalla mallilla eri tavoin.

Row ID	Image	Weather	Prediction (Weat...
Row663_Row...		shine	cloudy
Row666_Row...		shine	shine

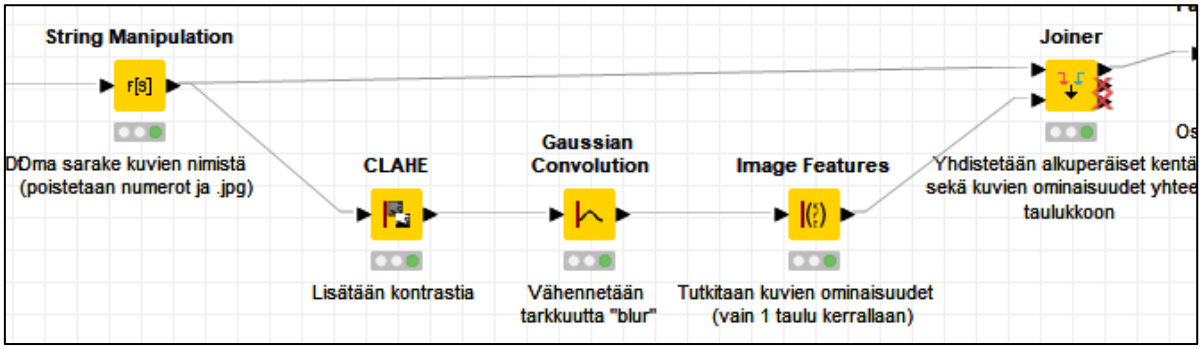
6.2.2 Muokattujen kuvien kategorisointi

Verrattuna työnkulkuun, jossa käsiteltiin muokkaamattomia kuvia, tulee muokkausten myötä käyttöön enemmän noodeja, joten työnkulku pitkittyy. Kuvien muokkaamisen noodit vievät myös useita minuutteja aikaa käsitellä, joten työnkulun läpivienti on myös enemmän aikaa vievää. Muokkaukseen käytettävien noodien valinta oli haastavaa ja joitakin noodeja jouduin jättämään pois, koska ne olivat liian raskaita. Esimerkiksi Global Thresholder -noodilla on mahdollista erotella kuvasta tausta ja etuala. Tätä noodia tulee käyttää yhdessä Connected Component Analysis -noodin kanssa, joka vaatii laitteelta paljon laskentatehoa. Omalla laitteellani oli tämän noodin kanssa ongelmia. Noodi eteni hitaasti ja olisi tarvinnut järkevään käyttöön enemmän suoritintehoa ja muistia, joten jätin nämä noodit kokonaan pois.

Valitsin kuvien muokkaamiseksi kaksi eri noodia. Clahe-noodi lisää kuvan kontrastia ja siten korostaa kuvan ominaisuuksia ja voi helpottaa mallia kategorisoinnissa. Gaussian Convolution -noodi poistaa kuvasta tarkkuutta ja yksityiskohtia. Kuvien ominaisuudet

analysoidaan Image features -noodilla, kuten aiemmin. Tällä noodilla voidaan kuitenkin analysoida vain yksi kuvasarake kerrallaan järkevästi (jottei taulukko paisu liian suureksi), joten kuvien ominaisuuksia tutkiva ajo tulee ajaa kahteen kertaan. Verrattuna aiempaan työnkulkuun kuvien ominaisuuksia muokkaavat uudet noodit sekä kuvien ominaisuuksia tutkivan noodin käyttö on kaikkein hitainta. Työnkulun suorittaminen vei aikaa hieman yli kymmenen minuuttia. Muilta osin työnkulku on samankaltainen muokkaamattomien kuvien kategorisoinnin kanssa. Eriävä osa työnkulusta on nähtävissä Kuva 33.

Kuva 33 Työnkulun eriävä osa muokattujen kuvien käsittelyssä.



Ensimmäisenä työnkulussa käsiteltiin kontrastia lisänneen Clahe-noodin avulla saatu aineisto. Mallin toimivuutta arvioivan noodin (Kuva 34) perusteella kokonaistarkkuus oli 0,739 eli 73,9 %. Luku on hieman korkeampi kuin muokkaamattomien kuvien kategorisoinnissa. Kaikkiaan 249 kuvaa kategorisoitiin oikein ja 88 väärin. Eroavaisuus luvuissa muokkaamattomiin kuviin on kuitenkin niin pieni, että se saattaa selittyä pelkästään yksittäisillä eroavaisuuksilla aineistoissa. Kontrastin lisääminen ei vaikuta ainakaan heikentävästi koneoppivan mallin tulkintoihin.

Kuva 34 Clahe-noodilla muokattujen kuvien tulokset koneoppivassa mallissa.

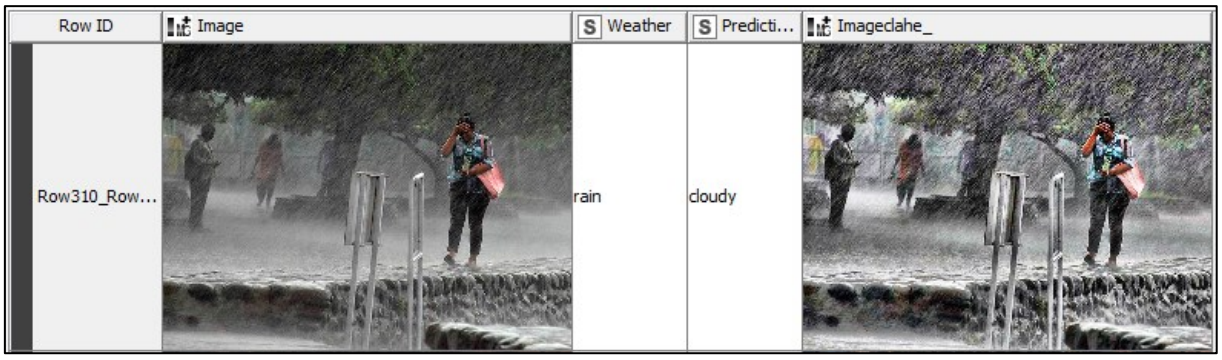
Table "default" - Rows: 1 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	Overall ...	Overall ...	Cohen's kappa	Correct...	Incorre...
Overall	0.739	0.261	0.648	249	88

Kun tulokset olivat valmiina tein havainnollistamiseksi vielä kaksi noodia: Joiner-noodin, jolla yhdistetään Clahe-noodista saadut kuvat ja alkuperäiset kuvat sekä mallin ennusteen

Decision tree predictor -noodista. Yhdistettyä taulu vielä muokattiin Column filter -noodilla siten, että siitä poistettiin tarpeettomat sarakkeet. Tulokset ovat Kuva 35.

Kuva 35 Alkuperäinen kuva, säätila, ennuste ja mallissa muokattu kuva.



Ihmissilmä tulkitsee sateisen kuvan, jossa kontrastia on lisätty helpommin sateeksi tulkittavaksi, koska pisarat erottuvat paremmin. Koneoppiva malli taas pitää kuvaa pilvisenä. Tosin kuvan valoisuus on selvästi korkeampi kontrastisessa kuvassa, joka vaikuttaa myös mallin toimintaan. Myös valkoista väriä on kuvassa havaittavissa selvästi enemmän.

Toisena muokkauksena kuviin tehtiin sumennus. Ennen ajoa työnkulkua muutettiin siten, että Image features -noodiin vaihdettiin tutkittaviksi kuviksi taulu, johon oli aiemmin lisätty sumeat kuvat. Koko työnkulkua ei siis tarvitse ajaa uudelleen, vaan työnkulku suoritetaan Image features -noodista eteenpäin. Tulokset mallin toimivuudessa sumeiden kuvien kohdalla ovat parhaimmat, sillä mallin kokonaistarkkuus on 0,754 eli 75,4 %. Malli arvioi 254 kuvaa oikein ja 83 kuvaa väärin.

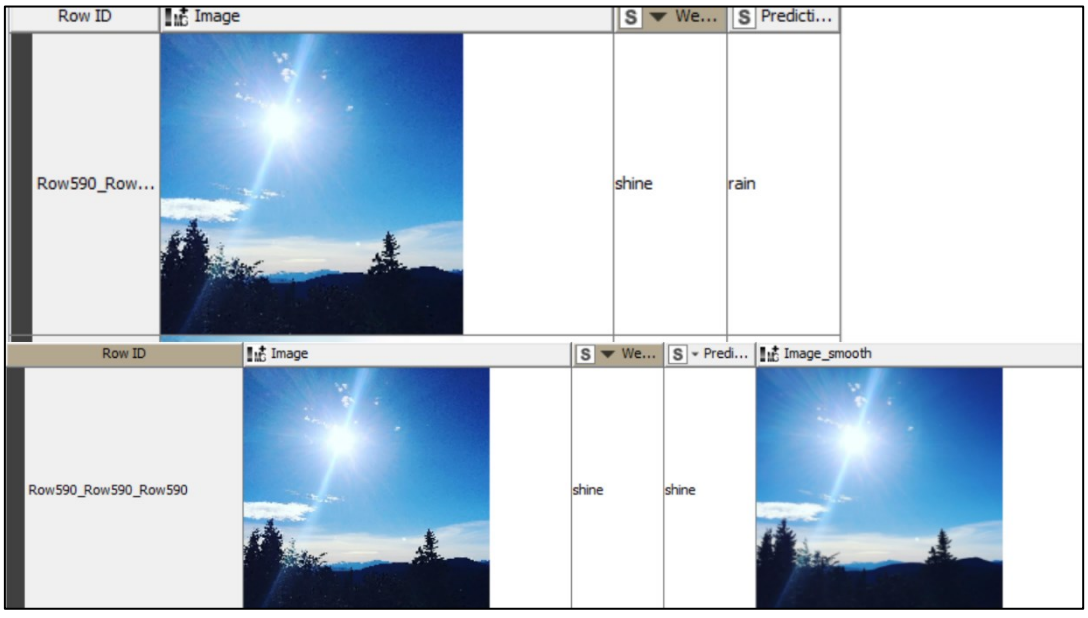
Taulukko 1 Sumennettujen kuvien mallin virheiden jakaumaTaulukko 1 on esitelty sumennettujen kuvien virheiden jakautuminen todellisen säätilan ja mallin tekemän ennusteen välillä. Eniten virheitä tuli pilvisten kuvien tulkinnoissa, joita tulkittiin esimerkiksi sateisiksi tai auringonpaisteeksi. Myös sateisissa kuvissa malli teki 14 kertaa väärän tulkinnan niin, että kuva olisi pilvinen. Auringonpaisteen ja auringonnousun osalta virheitä oli selvästi vähemmän.

Taulukko 1 Sumennettujen kuvien mallin virheiden jakauma

Sää	Ennuste	Virheiden lukumäärä
Pilvinen	Sateinen	**15**
Pilvinen	Auringonpaiste	**15**
Pilvinen	Auringonnousu	**6**
Sateinen	Pilvinen	**14**
Sateinen	Auringonpaiste	**3**
Sateinen	Auringonnousu	**4**
Auringonpaiste	Pilvinen	**7**
Auringonpaiste	Sade	**5**
Auringonpaiste	Auringonnousu	**2**
Auringonnousu	Pilvinen	**6**
Auringonnousu	Sateinen	**4**
Auringonnousu	Auringonpaiste	**2**
		83

Vertailin tuloksia alkuperäiseen työnkulkuun ja esimerkiksi Kuva 36 rivin 590 kuvan koneoppimisen malli muokkaamattomien kuvien työnkulussa teki arvionsa väärin, arvioiden sen sateeksi. Sumeaksi muokatun kuvan kanssa arvio oli oikea eli auringonpaiste. Mallien suora vertailu ei kuitenkaan ole mahdollista, sillä käytetyissä aineistoissa on eroja, koska ositus tehtiin molempiin työnkulkuihin erikseen.

Kuva 36 Koneoppimisen mallien eroavaisuus rivin 590 kuvan osalta.



6.2.3 Kuvien kategorisoinnin työnkulkujen vertailu

Kokonaisuutena kuvien tulkinnat onnistuivat kaikissa malleissa hyvin. Olin itse yllättynyt siitä, että melko vähän muokattujen kuvien käsittely auttoi koneoppivaa mallia tulkinnoissaan sekä siitä, että kaikkein parhaimpaan tulokseen päästiin sumentamalla kuvat. Ennen työnkulkujen suoritusta olisin olettanut, että alkuperäisillä kuvilla tai kontrastin korotuksella olisi päästy parhaimpiin tuloksiin. Tämä kuvaa hyvin sitä, että tässä työnkulussa koneoppiva malli tulkitsee kuvia ihmiseen verrattuna eri tavalla.

Muokattavien kuvien osalta kaikkia haluamiani muokkaamisen noodeja en voinut käyttää tietokoneen rajallisen tehon takia. Tästä huolimatta työnkulku todisti sen, että valmistelemalla kuva-aineistoa ennen koneoppivan mallin luontia, voi päästä parempaan lopputulokseen verrattuna kuvien käyttämiseen sellaisenaan. KNIME-ohjelmisto tarjoaa monipuolisia vaihtoehtoja kuvien muokkauksiin, erilaisten objektien erotteluun sekä taustan ja etualan tunnistamiseen.

7 KNIME-ohjelmiston tarjoamien kurssien arviointi

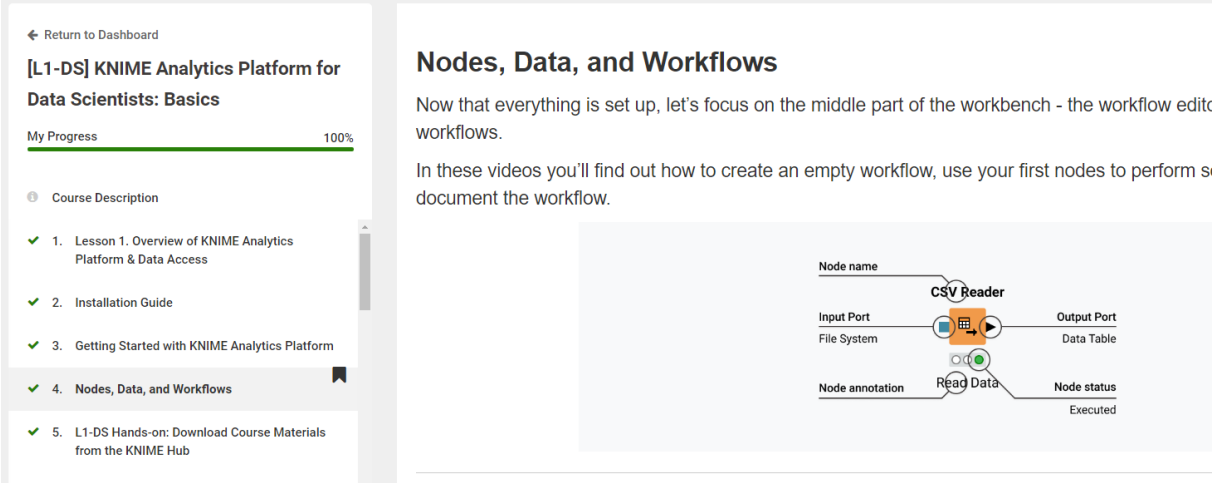
KNIME-ohjelmiston kurssit jakautuvat kappaleen neljä mukaisesti neljään tasoon L1-L4. Kurssit onkin suositeltavaa suorittaa siten, että ensimmäisenä suorittaa L1 eli perustason kurssin tai molemmat kurssit. Sitten etenee L2 kurssille, jossa syvennetään L1 kurssin osaamista. Kun myös L2 tasoinen kurssi on suoritettu, on mahdollista valita L3 tai L4 tasoinen kurssi, omien tai organisaation tarpeiden mukaisesti.

KNIME-ohjelmiston kurssit ovat laajoja kokonaisuuksia ja esimerkiksi perustason (L1) kurssien arvioitu kesto on kahdeksan tuntia. Todellisuudessa kurssiin käytettävä aika oli itselläni hyvin lähellä arviota ainakin ensimmäisen suorittamani kurssin osalta, jolloin pohjatietoa ei ollut lainkaan. Suoritusaikaan pidentävästi vaikuttivat pienet haasteet kurssin tehtävien kanssa sekä se, että katsoin joitakin opetusvideoita useamman kerran. Kurssia ei tarvitse suorittaa kerralla alusta loppuun, vaan sen voi keskeyttää ja jatkaa samasta kohdasta myöhemmin.

7.1 Perustason (L1) koulutukset

L1-DS-kurssi (KNIME Analytics Platform for Data Scientists: Basics) koostuu kaikkiaan 38 eri osasta. Osien sisältö on vaihtelevaa, jotkin osat ovat teoriapainotteisempia, osa harjoitteita ja osa lyhyitä n. viiden kysymyksen monivalintatenttejä. Tenttejä voi yrittää useita kertoja. Kurssin kokonaisuuden hahmottaminen on helppoa etenemispalkin (”My Progress”) ja suoritusprosentin ansiosta (Kuva 37). Lisäksi sivulla on nähtävissä kurssin eri vaiheet numeroituna.

Kuva 37 Etenemistä hyvin havainnollistava numerointi ja etenemispalkki



Kurssien teoriaosuus koostuu sekä kirjoitetusta tekstistä, kuvista että videoista. Pääpaino sisällössä on videoilla. Videot ovat onnistuneita, sillä ne ovat pituudeltaan sopivia (alle 10 minuuttia) ja ne käyvät asian tiiviisti ja käytännönläheisesti läpi. Videoita katsoessa suurin haaste oli se, että puhuja vaihtui eri videoilla ja osittain puhujien aksentti oli voimakas. Videoilla on kaikki puhuttu asia myös tekstinä, joten tästä ei ole suurempaa haittaa. Muutamissa KNIME-ohjelmiston ominaisuuksissa L1-DS-kurssilla oli visuaalisia eroja videon ja itse ohjelmiston välillä. Tämän aiheutti ainakin se, että video oli kuvattu joitakin vuosia aiemmin ja esimerkiksi joidenkin noodien konfigurointinäkymä oli ohjelmistopäivitysten myötä muuttunut. Nämä muutamat erot ohjeiden ja käytännön välillä aiheuttivat pieniä haasteita. Perusperiaate kaikilla käytetyillä noodeilla ja niiden toiminnoilla oli kuitenkin sama.

L1-DS-kurssin sisältöön kuului myös 11 harjoitustehtävää, joihin oli saatavilla kurssin alussa työnkulun pohjat, harjoitusten ratkaisut sekä käytettävät datatiedostot. Lisäksi visuaalinen ratkaisu oli saatavilla myös koulutussivustolla. Harjoitukset ovat mitoitukseltaan sopivia ja vievät n. 10 minuuttia aikaa, mikäli asian on omaksunut hyvin. Harjoituksiin annetut pohjat myös tukevat harjoitteiden suorittamista kohtuullisessa ajassa. Harjoitusten lisäksi kurssilla oli jokaisen asiakokonaisuuden jälkeen pienimuotoinen tentti, jota kutsutaan nimellä ”Knowledge check”. Tämä tentti koostuu viidestä monivalintatehtävästä, joissa vaaditaan 50

prosenttia kysymyksistä oikein. Tenttejä voi yrittää kuinka monta kertaa tahansa ja ne eivät vaikuta koulutuksen läpipääsyyn.

L1-DW-kurssin (KNIME Analytics Platform for Data Wranglers: Basics) sisältö oli L1-DS-kurssiin verrattuna näkökulmaltaan enemmän datan käsittelemiseen keskittyvä, mutta siinä oli myös paljon samoja sisältöjä kuin L1-DS-kurssissa. Arviolta samankaltaisuuksia oli noin kolmasosan verran. Suoritin L1-DW-kurssin läpi neljässä tunnissa ja 32 minuutissa jakaen suorituksen kahteen eri suorituskertaan. Koulutusportaalin antama arvioitu kesto kurssille oli 8 tuntia, mutta koska esimerkiksi tehtävissä oli täysin samoja tehtäviä ja samoja koulutuksen osuuksia kuin aiemmin, pystyin suorittamaan kurssin nopeammin. Kurssi oli laajuudeltaan hieman lyhyempi kuin L1-DS-kurssi ja se jakautui kaikkiaan 33 osioon ja yhdeksään harjoitustehtävään. Myös tällä kurssilla oli osana sisältöä KNIME-ohjelmiston asennusohje. Siten kumman tahansa L1-tasoisen kurssin voi valita aloittaessaan ohjelmaan perehtymisen.

7.2 Syventävät (L2) koulutukset

L2-DS-kurssin (KNIME Analytics Platform for Data Scientists: Advanced) sisältö oli selvästi haastavampaa kuin L1-tason kurssien. Kurssi koostui 26 osasta, kahdeksasta harjoituksesta sekä neljästä monivalintatehtävästä. Jotkin harjoitukset tuntuivat kovin hankalilta siksi, että oikean noodin valitseminen tehtäviin oli suuresta joukosta välillä kovin hankalaa. Neljäs harjoitus oli esimerkiksi sellainen, jota en saanut suoritettua loppuun asti oikein. En myöskään jäänyt perehtymään liian pitkäksi aikaa tehtävään, vaan tyydyin siihen, että tekemäni tehtävä oli oikean ratkaisun kaltainen. Myös toistoihin liittyvät tehtävät (loop) olivat melko haastavia ja vaativat perehtymistä. Myös toistotehtävissä jouduin käyttämään valmista ratkaisua apuna tehtävän suorittamiseen. Tehtävien suorittamista auttaisi esimerkiksi se, jos tehtävänannossa kerrottaisiin mitä noodia missäkin vaiheessa tulisi käyttää. Varsinkin jos KNIME-ohjelmistoon on asennettu paljon lisäosia, noodien määrä voi olla erittäin suuri ja juuri oikean löytäminen hankalaa. Kurssi oli sisällöltään monipuolinen, opetusvideot sopivan mittaisia ja materiaalit hyviä. Kurssin suorittamiseen kului aikaa

yhteensä neljä tuntia 58 minuuttia kolmella eri kerralla. Kurssin arvioitu kestoaika on kahdeksan tuntia.

7.3 Ehdotelma koneoppimisen kurssin sisällöksi

Koneoppimiseen liittyvän kolmen opintopisteen kurssin toteutus on mielestäni mahdollista käyttäen KNIME-ohjelmiston tarjoamia kursseja. L1-DS-kurssi (KNIME Analytics Platform for Data Scientists: Basics) on erinomainen lähtökohta ohjelmistoon tutustumiseen. Peruskurssilla tutustutaan myös koneoppimisen päätöspuumalliin ohjelmistossa. Tämän kurssin läpikäynti kesti itselläni noin kahdeksan tuntia useassa pienemmässä osassa. Kurssi on kyllä mahdollista suorittaa nopeamminkin, mutta varsinkin tehtäviin voi saada kulumaan aikaa enemmänkin, sillä tehtäviä on 11 kappaletta. Kaikkien tehtävien ratkaisut ovat saatavissa kurssipaketin mukana, joten itse harjoitustehtävien palauttaminen tai arvioiminen eivät ole tarpeen.

Toinen kurssille sopiva KNIME-ohjelmiston kurssi on L2-DS-kurssi (KNIME Analytics Platform for Data Scientists: Advanced). Tässä kurssissa syvennetään osaamista esimerkiksi muuttujien, tietokantojen ja työnkulun hallinnan osalta. Kurssin on selvästi L1-kurssia haastavampi, joten oikeat ratkaisut ovat tärkeitä oppimisen tueksi. Ne löytyvätkin ladattavasta koulutuspaketista. Tärkein osa kurssia on koneoppimisen tunti, joka on kurssilla neljäs. Tässä tutustutaan satunnaismetsän mallin käyttöön, joka on erinomainen koneoppimisen malli ja toimi hyvin mm. aiemmassa Titanicin matkustaja-aineiston käsittelyssä.

Tarjolla on myös KNIME-ohjelmisto kursseja datakäsittelijöille (Data wranglers), mutta osittain ne menevät sisällöltään päällekkäin datatieteilijöille (Data scientists) tarkoitettujen kurssien kanssa. Siksi jättäisin ne pois kurssikokonaisuudesta. Koneoppimiseen liittyvät asiat käsitellään myös juuri datatieteilijöille suunnatuilla kursseilla.

Kahden KNIME-ohjelmiston kurssin lisäksi kolmen opintopisteen kurssiin voisi sisältyä yksi tai useampi koneoppimisen harjoitustehtävä, jossa luodaan koneoppiva malli kaikille yhteisen harjoitusdatan perusteella. Lisäksi tehtävän ja luodun työnkulun kautta saatuja tuloksia voidaan esitellä lyhyellä raportilla, joka arvioi käytetyn mallin toimivuutta ja esittelee

tulosten avainluvut. Valmiita koneoppimisen harjoituksiin sopivia harjoitusaineistoja on saatavissa esimerkiksi osoitteessa https://data-flair.training/blogs/machine-learning-datasets/. Jotta prosessin luonti ja suorittaminen eivät veisi kohtuuttomasti aikaa on syytä valita jokin pienehkö aineisto. Sivustolla on saatavissa esimerkiksi ostoskeskuksen asiakasdataa ja Bostonin kaupungin asuntotietokanta. Myös tässä työssä käytetty RMS Titanicin matkustajaluettelo on saatavilla sivustolta.

Valitusta aineistosta ja tutkimuskohteesta riippuen voidaan opiskelijoita myös ohjeistaa siihen, mitä koneoppimisen menetelmää harjoituksessa tulisi käyttää, jotta sopimatonta mallia ei tulisi vahingossa käytettyä. L1- ja L2-tason suoritetut kurssit antavat hyvän pohjan koneoppimiseen keskittyvän harjoituksen tekemiseen erityyppisten aineistojen kanssa.

8 Johtopäätökset ja pohdinta

Opinnäytetyön tekemisen myötä yksi havainto koneoppimiseen liittyen oli se, että lähdeaineiston muotoon ja käyttötarpeisiin sovittamiseen on syytä käyttää tarpeeksi aikaa. Alun perin tarkoituksenani oli käyttää koneoppimisen työnkuluissa World bankin dataa, mutta koska data on todella valtavaa ja keskittyy jaottelemaan tiedot maittain, on käyttö koneoppimisen prosesseissa tarpeettoman hankalaa. Toisteinen ja helposti hahmotettava data oli sopivampi tähän tarkoitukseen, kuten RMS Titanicin matkustaja- ja selviytymistiedot.

Joitakin haasteita tuli työtä tehdessä vastaan. Valokuvien työnkulussa muutama valokuva oli tiedostomuodoltaan ”.jpeg” ja osa taas ”.jpg”. Tämä vaikeutti säätilojen nimen (joka ennen tiedostotyyppiä) siirtämistä omaan tauluunsa. Tämänkin ongelman pystyi kiertämään vain valmistelemalla aineiston erikseen ennen KNIME-ohjelmistoon lukemista. Toki ohjelmisto tarjoaa ratkaisuja myös vastaaviin ongelmiin, esimerkiksi säännöllisiä lausekkeita (Regular expression) hyödyntäen.

Koneoppimisen työnkulkuihin liittyvän aineiston hankinta oli haastavaa, varsinkin valokuvien osalta. Esimerkiksi biologian ja lääketieteen aloilta kuvakokoelmia oli paljon saatavilla, mutta halusin työhön helposti ymmärrettäviä kuvia. Erilaisista säätiloista kertova kuvasarja oli helposti lähestyttävä ja ei vaatinut suurempaa alustusta. Työn kannalta riitti, että kävi aineiston ja sen ominaisuudet lävitse. Monet löydetyt muut koneoppimiseen soveltuvat kuva-aineistot olivat kooltaan useiden gigatavujen kokoisia sisältäen jopa kymmeniä tuhansia kuvia. Näiden käsittely olisi ollut selvästi enemmän aikaa vievää, vaikkakin koneoppimisen prosessille optimaalisempaa. Tein kuitenkin ratkaisun tehdä työnkulut kevyemmillä aineistoilla.

KNIME-ohjelmisto on kätevä työkalu koneoppimisen tarpeisiin erityisesti dataa tai kuvia käsiteltäessä. Ohjelmistolle on tehty esimerkiksi laboratorion kuvia analysoivia koneoppivia työnkulkuja, joissa se pystyy löytämään haluttuja elementtejä. Mikäli tarkastellaan suurempaa datamäärää erityisesti kuvien osalta, on tärkeää huomioida, että vaatimukset käytettävälle tietokoneelle ovat korkeat. Muistia ja prosessoritehoa tarvitaan ajoittain todella paljon. Lisäksi tehokas kone varmistaa sen, että työnkulku valmistuu mahdollisimman nopeasti. Itse jouduin varsinkin valokuvia käsiteltäessä toteamaan tämän vaatimuksen

tehokkaasta tietokoneesta ja jättämään pois joitakin noodeja. Datan käsittelyssä ei ongelmia ollut tietokoneen tehon kanssa.

KNIME-ohjelmiston koulutusten läpikäynti osana työtä oli tärkeää ja vaikkei koulutukset olekaan edellytys ohjelman käyttämiselle, voi niistä saada hyviä vinkkejä datan käsittelyn eri vaiheisiin. KNIME-ohjelmiston kaikkiin kursseihin käyttämäni kokonaisaika oli alhaisempi kuin kurssin tekijän arvio käytettävästä ajasta. Oppimisen osana suoritettavien tehtävien tekemisen nopeus luultavasti tähän eniten vaikuttanut asia. Laadultaan koulutukset olivat hyviä, kunhan hyväksyy joidenkin videoiden voimakkaan murteen. Kaikki puhuttu asia on myös tekstimuodossa, joten suurta ongelmaa murre ei aiheuta.

Työhön ei varsinaisesti sisälly käyttöönottoa, mutta pidän työtä kuitenkin hyvänä pohjana tulevan koneoppimiseen ja KNIME-ohjelmistoon keskittyvän kurssin suunnittelemiseen. Kun opiskelija osaa luoda koneoppivan mallin, on sitä mahdollista hyödyntää myöhemmin monissa eri yhteyksissä opiskeluissa tai työelämässä. Työn pohjalta toimeksiantaja on päättänyt toteuttaa kolmen opintopisteen KNIME-ohjelmistoon pohjautuvan koneoppimisen kurssin.

Tulevaisuudessa tulen varmasti käyttämään KNIME-ohjelmistoa, sillä ohjelmisto tarjoaa paljon myös mahdollisuuksia, joita tässä työssä ei tarkemmin käsitellä. Tällaisia ovat mm. mahdollisuus tuoda dataa Twitteristä tai tietokannoista ja käsitellä sitä.

Opinnäytetyöprosessi oli kokonaisuutena melko työläs, koska koulutukset ja eri työnkulkujen luonnit veivät kaikkiaan useita henkilötyöpäiviä aikaa. Prosessi oli samalla myös erittäin opettava ja antoi erinomaiset valmiudet KNIME-ohjelmiston käyttöön ammattimaisesti työelämässä todellisen datan kanssa.

9 Yhteenveto

Tutkimuksen myötä KNIME-ohjelmiston ominaisuudet ja käytettävissä olevat välineet tulivat tutuiksi. Vaikka työ keskittyikin pääasiassa koneoppimiseen, on työssä hyödynnetty myös kuvan muokkauksen, visualisointien, datan valmistelun ja datan yhdistelyn erilaisia työvälineitä. Työkaluna KNIME on visuaalinen, eikä vaadi esimerkiksi koodaustaitoa. Ohjelmiston tarjoamat perus- ja syventävän tason kurssit antavat hyvän pohjan ohjelmiston käyttöön, mutta ne eivät ole välttämättömiä. Ohjelmiston verkkosivuilla on aktiivinen englanninkielinen yhteisö, jolta voi ongelmatilanteessa kysyä apua.

Tutkimuskysymykset käsiteltiin työssä mahdollisimman kattavasti. Tutkimuskysymyksistä haastavin oli ehdotelma koneoppimiseen liittyvän kurssin toteutuksesta KNIME-ohjelmistolla ja sen koulutuksilla. Näkemys perustuukin täysin omiin kokemuksiini ja ajatuksiini suoritettujen kurssien ja erilaisten koneoppimisen mallien tekemisen myötä. Lopputulos on mielestäni riittävän monipuolisen hahmotelma siitä, millainen kurssi voisi olla. Tutkimuskysymyksistä itse ohjelmistoon liittyviin kysymyksiin vastaaminen olikin sitten suoraviivaisempaa.

Opinnäytetyön myötä opin runsaasti koneoppimisen perusteista sekä käytännön toteutuksesta KNIME-ohjelmistolla. Lisäksi läpikäymieni koulutusten myötä pääsin kokeilemaan myös mm. tietokantojen tai pilvipalveluiden integroimista työnkulkuihin. Tämä on todennäköinen tapa tuoda tietoa data-analytiikan käyttöön ohjelmistoon. Erillinen Excel-tiedoston tuottaminen data-analytiikan tarpeisiin olisi turhan työlästä. Itse aion jatkaa opinnäytetyön tekemisen jälkeen ohjelmiston käyttöä.

KNIME-ohjelmiston kehitys jatkuu yhä ja opinnäytetyöprosessin aikana ohjelmistoon tuli useita päivityksiä. Aktiivinen yhteisö kehittää ohjelmistoon myös omia lisäosiaan, jotka koostuvat yhdestä tai useammasta noodista. Yhä paremmiksi kehittyvät noodit auttavat työnkulkujen kehittämisestä yhä autonomisemmiksi ja loppukäyttäjän, eli esimerkiksi data-analyytikon, työtä helpottavammaksi. Kun data-analyysi on onnistunutta, myös tietojen loppukäyttäjä eli päätöksentekijä saa parhaan mahdollisen tuen päätöksilleen.

Lähteet

Anon. (2021). *KNIME Quickstart Guide*. https://docs.knime.com/2021-06/analytics_platform_quickstart_guide/analytics_platform_quickstart_guide.pdf

DSPA. (n.d.). *CRISP-DM - Data Science Process Alliance*. Retrieved February 20, 2022, from https://www.datascience-pm.com/crisp-dm-2/

Kananen, H. (2019). *Tekoäly : bisneksen uudet työkalut*. Alma Talent Oy. https://bisneskirjasto-almatalent-fi.ezproxy.hamk.fi/teos/BAXBBXATCBIED#/kohta:TEKO((c4)LY((20)-((20)Bisneksen((20)uudet((20)tyokalut/piste:tU

Kelleher, J. D., Tierney, B., & Pietiläinen, K. (2021). *Datatiede*. Terra Cognita.

Knime. (n.d.-a). *Application: Car Counting – KNIME Hub*. Retrieved February 8, 2022, from https://hub.knime.com/knime/spaces/Examples/latest/99_Community/01_Image_Processing/03_Applications/04_Car_Counting~6W_4CLjiCAykbVe0

Knime. (n.d.-b). *External KNIME Courses | KNIME*. Retrieved February 8, 2022, from https://www.knime.com/external-courses

Knime. (n.d.-c). *KNIME Certification Program | KNIME*. Retrieved November 9, 2021, from https://www.knime.com/certification-program

Knime. (n.d.-d). *Numeric Scorer Node - YouTube*. Retrieved March 22, 2022, from https://www.youtube.com/watch?v=243VC3qkM-A&t=224s

Knime. (n.d.-e). *Self-Paced Courses List | KNIME*. Retrieved November 7, 2021, from https://www.knime.com/knime-self-paced-courses

Knime. (2020). *KNIME Analytics Platform Creating Data Science*. www.knime.com

KvantiMOTV. (n.d.). *Regressioanalyysi - KvantiMOTV*. Retrieved March 24, 2022, from https://www.fsd.tuni.fi/menetelmaopetus/regressio/analyysi.html

Lee, W.-M. (2019). *Python Machine Learning*. John Wiley & Sons, Incorporated.

Marr, B. (2015). *Big Data*. John Wiley & Sons, Incorporated.

Merilehto, A. (2018). *Tekoäly : matkaopas johtajalle*. Alma Talent.

Sedkaoui, S. (2018). *Data Analytics and Big Data*. Wiley-ISTE.

Widmann, M. (n.d.). *Cohen's Kappa: what it is, when to use it, how to avoid pitfalls | KNIME*. Retrieved March 22, 2022, from https://www.knime.com/blog/cohens-kappa-an-overview

Liite 1: Aineistonhallintasuunnitelma

Kehitysprojektin aikana pidetään päiväkirjaa (aineisto), johon kerätään teknistä tietoa projektista. KNIME-koulutusten ajankäytöstä ylläpidetään osana päiväkirjaa lisäksi Excel-tiedostoa, joka säilytetään päiväkirjan mukaisesti. Tämä tieto analysoidaan opinnäytetyötä varten. Päiväkirjaa ja muuta dokumentaatiota säilytetään tekijän tietokoneen C-asemalla, ja siitä tehdään säännöllisesti varmuuskopioita OneDrive -pilvipalveluun sekä erilliselle Usb-muistille. Päiväkirjaa säilytetään C-asemalla ainakin vuoden verran opinnäytetyön valmistumisesta. Varmuuskopiot poistetaan työn valmistutuuttua.