Haaga-Helia
ammattikorkeakoulu Oy

Koneoppimisen regressioalgoritmin soveltaminen Azure Machine Learning-palvelussa

Pontus Lindman

Opinnäytetyö
Tietojenkäsittelyn koulutusohjelma
2017

Tiivistelmä

Haaga-Helia
ammattikorkeakoulu Oy

Tekijä(t)
Pontus Lindman

Koulutusohjelma
Tietojenkäsittelyn koulutusohjelma

Raportin/Opinnäytetyön nimi
Koneoppimisen regressioalgoritmin soveltaminen Azure Machine Learning-palvelussa

Sivu- ja liitesivumäärä
33 + 6

Tämän tutkimuksen aiheena oli selvittää miten koneoppimisen tekniikoita voi käyttää laskurin tuloksen laskemiseksi ilman työlästä perinteisten algoritmien kehittämistä manuaalisesti. Koneoppiminen laajempana ilmiönä tarjoaa mahdollisuuden löytää merkityksiä laajoista tietomassoista, joihin perinteisin menetelmin vastaava on vaatinut hyvin usein useamman vuoden kestävän kehitysprosessin. Pitkään nämä koneoppimisen menetelmät ovat olleet ainoastaan pidempään koneoppimista opiskelleiden asiantuntijoiden käytettävissä. Viimeisten vuosien aikana teknologiatoimittajat ja pilvipalvelujen tarjoajat ovat kuitenkin tuoneet markkinoille palveluja, joiden avulla koneoppimisen menetelmät ja teknologiat ovat helpommin kaikkien saatavilla.

Tutkimuksen ensimmäisessä osassa selvitetään mistä koneoppimisessa on kyse ja miten koneoppimista käyttävät tietotekniset ratkaisut toimivat käytännössä. Tässä osassa esitellään koneoppimisen vahvuuksia perinteisiin menetelmiin verrattuna, siihen liittyvä yleinen toimintaperiaate, ja pureudutaan syvemmin piirteiden suunnitteluun ja algoritmien käyttöön koneoppimisessa.

Tutkimuksen toisessa osassa tutustutaan CRISP-DM-menetelmään, joka on yleisin tiedonlouhinnan ja koneoppimisen prosessimalli. CRISP-DM-menetelmään kuuluu 6 vaihetta, joissa liiketoiminnan ymmärtämisen ja käytettävän datan ymmärtämisen kautta päästään julkaisemaan toimiva koneoppimisen malli sitä hyödyntäviin palveluihin ja sovelluksiin.

Tutkimuksen kokeet toteutettiin Azure Machine Learning palvelun avulla. Azure Machine Learning on MLaaS-tyyppinen pilvipalvelu, jossa ilman omille koneille asentamista voi suoraan kehittää koneoppimisen malleja ja toteuttaa erilaisia siihen liittyviä kokeita. Kappaleen alussa esitellään hyvän MLaaS-palvelun arviointikriteerejä ja verrataan miten hyvin tämä palvelu vastaa näihin kriteereihin. Kappaleen toisessa osassa esitellään itse palvelu ja sen koneoppimisen mallin kehittämisen kannalta olennaisimmat moduulit.

Työn käytännön osassa suoritin käytännön kokeena yleisesti tunnetun FINRISKI-laskurin toteuttamisen koneoppimisen keinoin CRISP-DM-menetelmän mukaisesti. Ensimmäisenä vaiheena tässä osuudessa oli selvittää ja kuvata laskurin toiminta, jonka jälkeen hain avoimista tietolähteistä tutkimukseen soveltuvaa aineistoa. Tutkimuksen aineistolle tein laadunvarmistuksen kahdessa vaiheessa tutkimuksessa kuvattujen menetelmien mukaisesti ja tällä aineistolla vertailin ja arvioin palvelun regressio-algoritmeja. Laatuvarmistetun datan ja soveltuvimman algoritmin avulla toteutin FINRISKI-laskurin ja varmistin sovellettavan menetelmän keinoin mallin toimivuuden. Lopuksi julkaisin opetetun mallin ulkopuolisten sovellusten käytettäväksi palvelun Web Services-rajapintojen avulla.

Asiasanat
tekoäly, koneoppiminen, ohjattu oppiminen, algoritmit, Azure Machine Learning, CRISP-DM

Sisällys

1 Johdanto

Tämän tutkimuksen aiheena oli selvittää, miten koneoppimisen tekniikoita voi käyttää laskurin tuloksen laskemiseksi ilman työlästä perinteisten algoritmien kehittämistä manuaalisesti. Koneoppiminen laajempana ilmiönä tarjoaa mahdollisuuden löytää merkityksiä laajoista tietomassoista, joihin perinteisin menetelmin vastaava toteutus on vaatinut hyvin usein useamman vuoden kestävän kehitysprosessin. Pitkään kuitenkin nämä koneoppimisen menetelmät ovat olleet ainoastaan pidempään koneoppimista opiskelleiden asiantuntijoiden käytettävissä. Viimeisten vuosien aikana teknologiatoimittajat ja pilvipalvelujen tarjoajat ovat kuitenkin tuoneet markkinoille palveluja, joiden avulla koneoppimisen menetelmät ja teknologiat ovat helposti kaikkien saatavilla. Minusta oli mielenkiintoista selvittää ja tutkia miten helposti nämä uudet palvelut ja työkalut ovat hyödynnettävissä.

Koneoppimisen käyttö eri toimialoilla on yleistynyt hyvin nopeasti. Tällä hetkellä merkittävä osa teknologiauutisista kertoo siitä, miten koneoppimista on hyödynnetty hyvinkin erilaisissa yhteyksissä. Esimerkkejä paljon julkisuutta saaneista käyttökohteista ovat Otaniemen itseohjautuvat bussit, Trumpin vaalikampanjassa käytetyt menetelmät kohdentaa viestejä äänestäjille sosiaalisen median tiedon perusteella ja koneoppimisen menetelmien hyödyntäminen kyberrikollisuuden torjumisessa. (Espoo 2017; Grassegger & Krogerus 2017; Kanowitz 2017)

Kaikissa näissä esimerkeissä yhteistä on se, että innovaation mahdollistava koneoppiminen ei näy loppukäyttäjille asti. Itse palvelu tuntuu käyttäjistä samanlaiselta kuin ennenkin, mutta nyt se vain jotenkin toimii paremmin kuin ennen. Koneoppimiseen liittyykin vahvasti, että se on huomaamaton taustalla. Siksi on helppoa olla huomaamatta, miten vahvasti se vaikuttaa ja tulee vaikuttamaan sekä yhteiskuntaan laajasti että jokaisen meidän arkeen. Tekoälyllä ja koneoppimisella tuleekin olemaan paljon isompi vaikutus yhteiskuntaan kuin monilla perinteisillä teknologiainnovaatioilla on tähän mennessä ollut.

Koneoppimisen arkipäiväistyessä oli mielenkiintoista tutkia ja oppia, miten uudet palvelut, jotka mainostavat tekevänsä koneoppimisen menetelmien käytöstä helppoa, toimivat käytännössä. Keskeisinä tavoitteina tässä tutkimuksessa oli selvittää, miten helposti ja hyvin koneoppimisen keinojen avulla pystyy toteuttamaan laskurin sekä miten hyvin tämän toteutus onnistuu MLaaS-palvelussa ilman, että tarvitsee asentaa tarvittavat ympäristöt ja työkalut omille tietokoneilleen.

1.1 Käsiteluettelo

Ennustava analytiikka	Tilastotieteen menetelmä, jossa otoksesta pyritään löytämään tietämys, jonka avulla pystytään ennustamaan tulevia tuloksia ja trendejä.
Koneoppiminen	Tekoälyn osa-alue, jonka tarkoituksena on saada ohjelmisto toimimaan pohjatiedon perusteella ilman että ihminen on määrittänyt toimintamenetelmän jokaista tilannetta varten, vaan ohjelmisto oppii itsenäisesti päätymään haluttuun lopputulokseen.
Ohjattu oppiminen	Koneoppimisen menetelmä, jossa opetusaineiston avulla muodostetaan malli, jolla luokiteltava aineisto voidaan luokitella. Opetusaineisto koostuu syötteistä ja tuloksista, jotka syötteistä tulisi seurata eli haluttu tulos tunnetaan etukäteen.
Koneoppimisen malli	Koneoppimisen algoritmilla opetusjoukosta muodostettu malli, joka ennustaa pohjatiedosta algoritmin mukaisen tuloksen.
Algoritmi	Säännönmukainen, mekaaninen laskumenetelmä.
Opetusjoukko	Otoksesta erotettu joukko, jota käytetään koneoppimisen algoritmin opettamiseen.
Testijoukko	Otoksesta erotettu joukko, jota käytetään koneoppimisen mallin arviointiin.
Otos	Yhteinen nimi eri tarkoituksiin käytettäville tietojoukoille.
Otosalkio	Otoksen yksi yksittäinen kappale, tietue, rivi.
Piirre	Koneoppimisessa otoksen yksittäinen muuttuja. Mikä tahansa käytettävistä tietolähteistä muodostettu suure, jonka voi syöttää koneoppimisen algoritmille.
Keskineliövirheen neliöjuuri	Tilastollinen tunnusluku, jota käytetään kuvaamaan regressiomallin ennusteiden laatua.
Selitysaste	Tilastollinen tunnusluku, joka mittaa regressiomallin selittämää osuutta selitettävän muuttujan y havaittujen arvojen kokonaisvaihtelusta ja on hyvä regressiomallin hyvyyden ja yhteensopivuuden asteen mittarina.
CRISP-DM	Cross-Industry Standard Process for Data Mining. Yleinen tiedonlouhinnan ja koneoppimisen prosessimalli.
MLaaS	Machine Learning as a Service. Koneoppiminen pilvipalveluna.

2 Koneoppiminen

Tässä tutkimuksessa perehdytään koneoppimiseen ja sen käyttöön. Koneoppimisella tarkoitetaan yleisesti sitä tietotekniikan alalajia, jossa tietokone tai ohjelma pystyy annetusta aineistosta tekemään ennusteita tai päätöksiä ilman ohjelmointia. Ensimmäisenä tämän määritelmän julkaisi tekoälyn pioneeri Arthur Samuel vuonna 1959 (Samuel 1959). Kohavi ja Provost (1998) ovat määritelleet koneoppimisen tarkoittavan sellaisten algoritmien tutkimista ja tuottamista, jotka oppivat ja pystyvät tekemään ennusteita datasta.

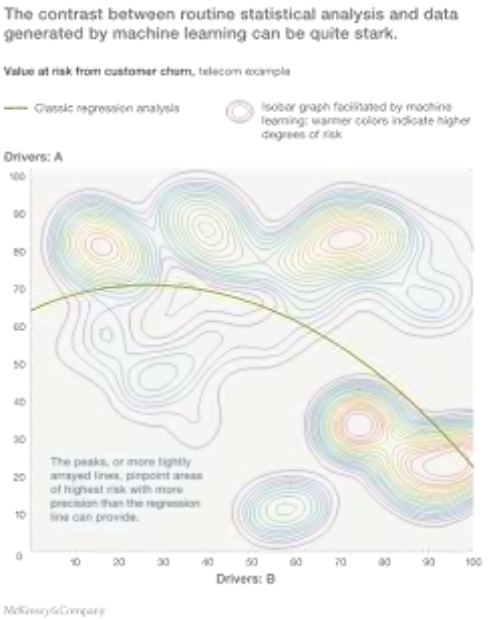
Koneoppimisella on vahva yhteys tilastotieteeseen. Harringtonin mukaan tilastotiedettä käytetään perinteisesti ongelmiin, joihin ei löydy täydellistä ratkaisua tai ongelmiin, joiden ratkaisemiseen olemassa oleva laskentateho ei yksinkertaisesti riitä. Esimerkkinä tällaisesta ongelmasta hän mainitsee ihmisten onnellisuuden arvioimisen ja mittaamisen. Asiat, mitkä tekevät henkilöstä onnellisen, vaihtelevat eri henkilöiden välillä ja tämän mallintaminen ja mittaaminen täydellisesti, on kerta kaikkiaan mahdotonta. Mm. sosiaalitieteissä on yleistä, että jo osittainen oikeassa oleminen lasketaan onnistumiseksi, koska on mahdotonta olla objektiivisesti täysin oikeassa. Laskentatehon tarpeesta Harrington käyttää esimerkkinä Internet of Things-ilmiötä ja siihen liittyvien erilaisten reaaliaikaisten sensoreiden keräämän valtavan datamäärän käsittelyä. (Harrington 2012, 5-6)

Pyle ja San Jose nostavat tilastotieteen tilastollisen päättelyn merkityksen koneoppimisen algoritmien ja menetelmien perustana. He huomauttavat kuitenkin, että viimeisten vuosisatojen aikana kehitetyt perinteisen tilastotieteen menetelmät luotiin huomattavasti pienemmille tietomäärille kuin mitä tänä päivänä on käytössämme. Tämän takia nämä eivät sovellu sellaisenaan nykypäivän ongelmien ratkaisemiseen. Koneoppimisen läpimurto omana tieteenlajina tapahtui 1990-luvulla, kun tutkijoiden käyttöön tuli riittävä laskentateho. Tämä mahdollisti koneoppimisen menetelmien toteuttamisen käytännössä ja oppivien algoritmien käytön tiedon käsittelyssä. (Pyle & San Jose 2015)

Tilastotieteeseen ja analytiikkaan liittyy Kelleher, Mac Namee ja D’Arcyn mukaan vahvasti datan muuttaminen informaatioksi ja tietämykseksi. Tämä on erityisen tärkeää nykyaikana, kun meillä on historiallisesti ennennäkemätön määrä dataa käytettävissä. Modernin organisaation menestyminen perustuu yhä enemmän ja vahvemmin siihen, että organisaatio osaa ja pystyy jalostamaan kaikesta hallussaan olevasta datasta oivalluksia ja näiden oivallusten perusteella tekemään päätöksiä. Kun data-analytiikassa keskeiseen datan jalostamiseen käytetään koneoppimisen menetelmiä, kutsutaan tätä yleisesti ennustavaksi analytiikaksi. (Kelleher, Mac Namee, D'Arcy 2015, 1.)

2.1 Koneoppimisen vahvuudet

Koneoppimisen keskeisin vahvuus perinteisiin menetelmiin verrattuna on Pyle ja San Josen mukaan sen kyky käsitellä todella isoja tietoaineistoja automaattisesti. Käytettävissä olevat tietomäärät ovat nykyään niin valtavat, ettei kukaan ihminen pysty niitä mitenkään käsittelemään järkevästi manuaalisesti. Vaikka koneoppimisen algoritmeja ei vielä voi kuvata erityisen älykkäinä, ovat ne jo nyt erittäin hyviä käsittelemään rajatonta määrää dataa ja tämän datan piirteiden mitä tahansa yhdistelmiä. (Pyle & San Jose 2015)



Kuva 1. Koneoppimisella tarkempi tulos (Pyle ja San Jose 2015)

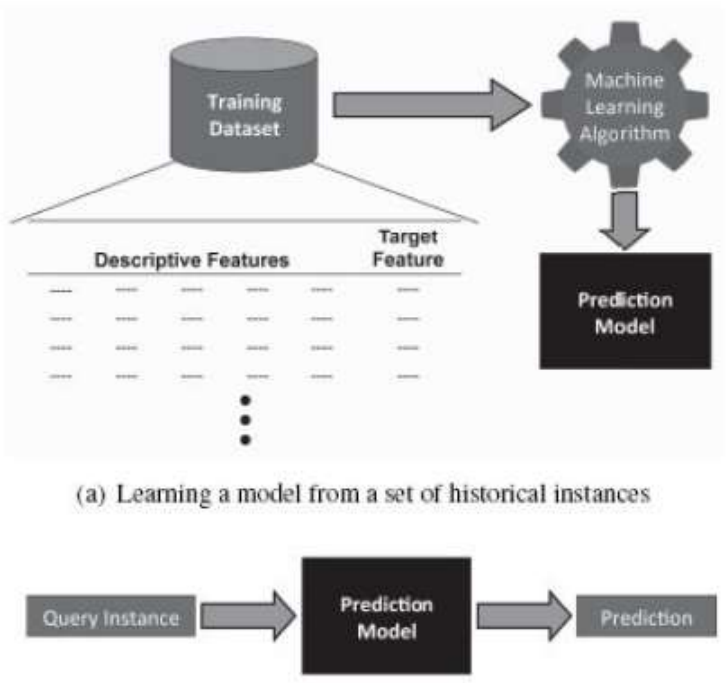
Kuvassa 1 on havainnollistettu, miten koneoppimisen menetelmien hyödyntäminen tiedon analysoimisessa parantaa tarkkuutta verrattuna perinteisempiin tilastotieteen menetelmiin. Siinä vihreän värinen viiva kuvaa sitä tarkkuutta, millä esimerkin tiedot pystytään erottelemaan perinteisillä tilastotieteen regressio-menetelmillä. Syheröiset ääriviivat taas kuvastavat, miten koneoppimisen menetelmin voidaan muodostaa epäsymmetrisiä ja monimutkaisia yhteyksiä piirteiden ja ennustettavan tuloksen välillä. (Pyle & San Jose 2015)

Mathworksin koneoppimisen oppaassa on kiteytetty koneoppimisen menetelmien käytön vahvuudet. Sen mukaan koneoppimisen käyttöä kannattaa erityisesti harkita kolmessa tapauksessa. Ensimmäisenä mainitaan se, jos sääntöjen ja algoritmien tekeminen käsin on liian monimutkaista. Puheentunnistus on hyvä esimerkki tällaisesta ongelmasta. Puheentunnistuksen yhteydessä koneoppimisen menetelmien vahvuutena on kyky muodostaa tarvittavat säännöt ilman, että ihminen määrittelee nämä säännöt. Toisessa tapauksessa säännöt, joilla päästään haluttuun tulokseen, muuttuvat jatkuvasti. Esimerkkinä tällaisesta on luottokorttien väärinkäytön havaitseminen. Koska koneoppimista käyttämällä säännöt

muodostuvat automaattisesti, voi käytetty malli sopeutua muuttuneeseen tilanteeseen, kun sitä opetetaan säännöllisesti uudestaan. Kolmannessa tapauksessa käytössä oleva tieto muuttuu jatkuvasti. Esimerkiksi kaupassa tuotteet vaihtuvat tiuhaan ja ostoskäyttäytymisen ennustaminen vaatii ratkaisua, joka toimi uusilla tuotteilla ilman työläitä muutoksia sovelluksiin ja niiden toteuttamiin sääntöihin. Koneoppimisen vahvuutena on sen kyky yleistää niin, että se toimii myös uudella tiedolla eikä pelkästään opetusjoukkona käytetyn datan kanssa. (Mathworks 2016, 8.)

2.2 Koneoppimisen toimintaperiaate

Koneoppimisen käytössä keskeisenä piirteenä on, että koneoppimisen malli oppii itsenäisesti, miten datasta päästään haluttuun tulokseen ilman, että ihminen määrittelee käytettävän algoritmin. Koneoppimisen yleinen toimintaperiaate on, että se muodostaa datasta ja sen piirteistä joukon toimintamalleja, joiden joukosta se valitsee sen, joka parhaiten yleistää - ei pelkästään opetusjoukkona käytetyn datan suhteen haluttuihin tuloksiin - vaan yleisemmin sen, miten mistä tahansa datasta voidaan ennustaa haluttu tulos.



Kuva 2. Koneoppimisen yleinen toimintaperiaate (Kelleher ym. 2015, 3)

Koneoppimisen yleisenä toimintaperiaatteena on kuvan 2 mukaisesti muodostaa opetusjoukon ja valitun koneoppimisen algoritmin avulla ennustamiseen käytettävän mallin. Algoritmi päättelee induktiivisen päättelyn oletuksien avulla opetusjoukon piirteiden ja ennustettavan tuloksen välisen yhteyden. Mallia käytettäessä siihen syötetään uutta tietoa ja

malli laskee opetetun mallin avulla ennustetun tuloksen. Esimerkkinä toimintaperiaatteesta on sen käyttö lainanhakijan luottokelpoisuuden arvioinnissa. Piirteinä voisi käyttää esimerkiksi hakijan ikää ja ammattia sekä haettavan lainan määrä suhteessa hakijan palkkaan. Ennustettavana tuloksena voisi olla se, pystyykö hakija maksamaan lainan takaisin kokonaisuudessaan vai ei. Opetettua mallia voidaan sitten käyttää esimerkiksi lainaneuvottelussa hakijan kanssa, kun virkailijan tarvitse arvioida hakijan tiedoilla voidaanko hakijalle myöntää laina.

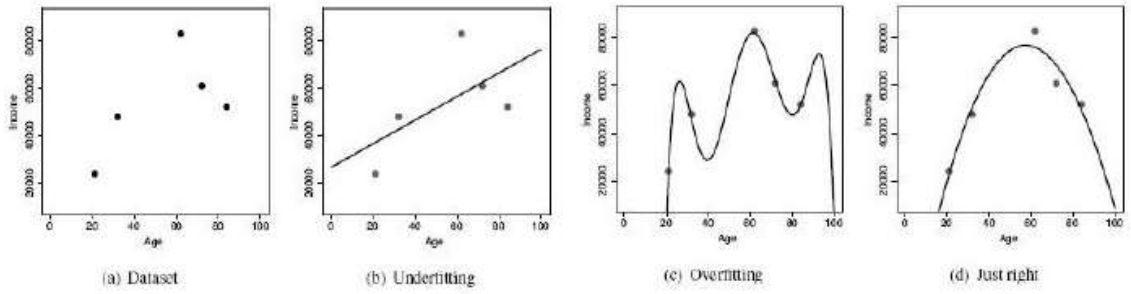
Opetettua koneoppimisen mallia kutsutaan ristiriidattomaksi, jos se ennustaa oikein kaikilla opetusjoukon otosalkioilla. Koneoppimisen algoritmit toimivat siten, että ne muodostavat useita malleja ja hakevat tästä joukosta sen mallin, joka parhaiten ennustaa opetusjoukon piirteiden ja ennustettavan tuloksen välisen yhteyden. Tietojenkäsittelyn perinteisissä menetelmissä ongelmiin haetaan yleensä ristiriidatonta mallia eli sellaista, joka aina tuottaa kaikesta datasta täysin oikean tuloksen. Isoilla tietomäärillä tällainen ristiriidaton malli ei kuitenkaan ole enää paras mahdollinen. Käytettävässä datassa isoilla tietomäärillä esiintyy lähes poikkeuksetta kohinaa ja virheitä, jotka ristiriidattomassa mallissa aiheuttavat oikeassa käytössä virheellisiä ennustettuja tuloksia, koska isoilla tietomäärillä opetusjoukko on vain pieni otos kaikesta datasta.

Tämän takia koneoppimisesta sanotaan, että se ratkaisee huonosti asetettuja ongelmia, joihin ei ole yhtä ainoaa oikeaa ratkaisua käytettävissä olevan tiedon perusteella. Esimerkkinä tällaisesta ongelmasta on ruokakaupan ostokset, joissa mahdollisten ostosten yhdistelmien määrä on todella iso. Näistä yhdistelmistä ihmisten on käytännössä mahdoton löytää sellaista mallia, joka aina muutaman ostoksen perusteella pystyy täydellisesti ennustamaan henkilön kaikki ostokset. Vaikka käytettävästä opetusjoukosta onnistuisimme tällaisen täydellisen mallin luomaan, ei se todellisuudessa toimi täydellisesti muulla aineistolla.

Koneoppimisen menetelmille hyvän mallin tekeminen tällaiseen käyttöön on toisaalta helppoa. Hyvän koneoppimisen mallin yksi keskeinen ominaisuus on se, että se pystyy ennustamaan myös sellaisia tapauksia, jotka eivät esiinny opetusjoukossa. Koneoppimisen malli siis yleistää suhteen opetusjoukon piirteiden ja ennustettavien tulosten välillä. Opetusjoukon datalla opetettua ristiriidatonta mallia ei voi käyttää ennustamiseen, koska ristiriidaton malli oppii vain ulkoa opetusjoukkona käytetyn datan. Mitään varsinaista oppimista ei tapahdu, koska ristiriidaton malli ei etsi eikä löydä syvällisempiä yhteyksiä piirteiden ja ennustettavan tuloksen välillä. Koska koneoppimisen käyttäminen on hyödyllistä vain, jos opetettu malli toimii opetusjoukon lisäksi myös muulla datalla, ei ristiriidattoman

mallin etsiminen ole järkevää koneoppimista käytettäessä. Koneoppimisessa tavoitteena onkin löytää se malli, joka yleistää parhaiten piirteiden ja tuloksen todellisen yhteyden.

Kun koneoppimista käytetään, voidaan valita useamman eri algoritmin välillä. Jokaiseen algoritmiin liittyvät erityiset säännöt, miten se päättelee yhteyden piirteiden ja tulosten välillä. Käytännössä tämä tarkoittaa sitä, että eri algoritmit toimivat eri datalla paremmin tai huonommin. Tämän takia on tärkeää kiinnittää huomiota sopivimman algoritmin valintaan kuhunkin tapaukseen erikseen. Yleisesti ei ole yhtä algoritmia, joka toimii kaikissa tapauksissa parhaiten. Taito valita oikea algoritmi onkin koneoppimista käyttävän data-analyytikon yksi olennaisimpia kykyjä.



Kuva 3. Koneoppimisen mallin ali- ja ylisovittuminen (Kelleher ym. 2015, 13.)

Väärän algoritmin seurauksena tuloksena saadussa mallissa voi esiintyä kahdenlaista ongelmaa: alisovittuminen tai ylisovittuminen. Alisovitettu malli on todelliseen dataan verrattuna liian yksinkertainen, eikä ole tämän seurauksena riittävän herkkä todellisessa datassa olevaan vaihteluun. Ylisovitettu malli taas seuraa opetusjoukon dataa liian tiukasti ja on tämän seurauksena liian herkkä todellisessa datassa olevaan kohinaan. Kuvassa 3 on havainnollistettu näitä ongelmia. Kohdassa (b) opetettu malli on selvästi alisovitettu eikä pysty reagoimaan riittävän tarkasti, kun taas kohdassa (c) opetettu malli on ylisovitettu ja seuraa liian tarkasti opetusjoukon dataa. Vain kohdan (d) malli on pystynyt löytämään valitun piirteen ja ennustettavan tuloksen välisen todellisen yhteyden ja toimii yleisesti oikein myös muulla kuin opetusjoukon datalla. (Kelleher ym. 2015, 3-13)

2.3 Piirteiden valinta ja suunnittelu

Koneoppimisen mallissa käytettävät piirteet suunnitellaan vastaamaan mahdollisimman hyvin mallin kohteena olevaa oikeaa maailmaa ja sen käsitteitä. Piirteet ovat kuitenkin aina jonkinlainen approksimaatio todellisuudesta. Näiden suunnittelussa data-analyytikko joutuu arvioimaan, miten eri tietolähteitä hyödynnetään parhaiten ja keksimään, miten voidaan arvioida ja laskea sellaisia tietoja, jotka eivät ole mitattavissa.

Piirteiden arvot ovat yleisesti joko numeroarvoja, päivämääriä, aikoja, rajallisen arvojoukon arvoja, loogisia arvoja tai merkkijonoja. Hyviä esimerkkejä piirteissä käytettävistä arvoista ovat hinta, syntymäaika, tapahtuma-aika, maakoodit, sukupuoli, nimet ja osoitteet. Koneoppimisessa piirteiden tyypit on ryhmitelty kahteen luokkaan: jatkuviin ja luokitteleviin tyyppeihin. Jatkuviin tietotyyppeihin lasketaan numeeriset arvot ja aikatiedot, kun taas luokitteleviin lasketaan rajalliset arvojoukon arvot, loogiset arvot ja merkkijonot. Luokittelevilla piirteillä mahdollisten arvojen joukko on aina rajallinen.

Koneoppimisen mallissa käytettävä piirre voidaan joko ottaa suoraan tietolähteestä tai se voi olla johdettu piirre. Suoraan tietolähteestä otettavia piirteitä voivat esimerkiksi olla henkilön ikä ja sukupuoli. Johdetut piirteet muodostetaan käytössä olevista tiedoista esimerkiksi yhdistelemällä useamman tietolähteen tietoja. Johdettuja piirteitä voivat esimerkiksi olla asiakkaan keskimääräiset ostokset kuukaudessa tai kirjautumisten lukumäärä kuukaudessa.

Johdetut piirteet ovat yleisimmin joko aggregaatteja, loogisia suureita, suhteita tai kartoituksia. Aggregaatti on numeerinen suure, jossa tietoja ryhmitellään ja näistä lasketaan esimerkiksi alkioiden lukumäärä, summa tai keskiarvo. Aggregaatti voi myös olla esimerkiksi joukon minimi- tai maksimiarvo. Esimerkki aggregaatista on henkilön ostosten kokonaissumma edellisen kolmen kuukauden aikana. Looginen suure kertoo, onko tiedoissa jokin ominaisuus vai ei. Tällainen suure voi esimerkiksi olla tieto onko henkilöllä ollut tilinylityksiä. Suhde on numeerinen suure, joka arvioi kahden piirteen arvojen välistä suhdetta. Tämä voi esimerkiksi olla henkilön lainan määrän suhteessa henkilön palkkaan. Kartoitus on suure, joka on kartoittamalla johdettu muusta tiedosta. Tätä voidaan käyttää esimerkiksi vähentämään tiedossa olevien vaihtoehtojen määrää pienemmäksi. Esimerkiksi henkilöiden palkat voidaan kartoittamalla muuntaa numeerisista luvuista luokkiin matala, keskikorkea tai korkea.

Samasta tiedosta voi johtaa äärettömän määrän eri piirteitä. Henkilön suorittamasta sähkönkäytön maksuista voidaan esimerkiksi johtaa keskimääräinen maksu, minimimaksu ja maksimimaksu eri aikajaksoilla (vuoden ajalta, 3 kuukauden ajalta) tai voidaan muodostaa piirre, joka kertoo, onko asiakas myöhästynyt maksujen kanssa viimeisen vuoden aikana.

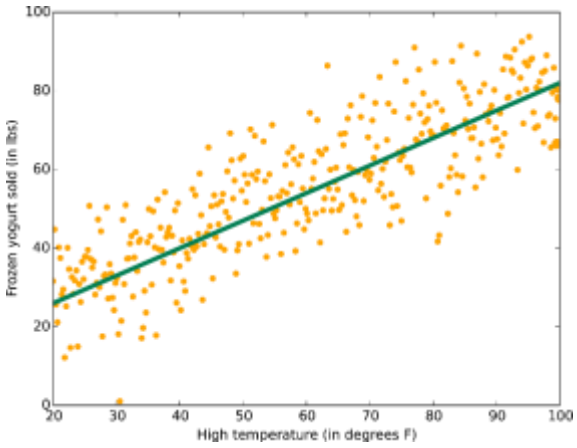
Piirteiden suunnittelussa kolme erityisen tärkeää huomioitavaa asiaa ovat tiedon saatavuus, ajoitus milloin tieto on käytettävissä ja tiedon pitkäikäisyys. Tiedon saatavuus tarkoittaa sitä, että piirteen muodostamiseen tarvittava tieto täytyy olla saatavilla ja käytettävissä. Jos esimerkiksi piirre perustuu historiatietoon, täytyy tietolähteessä säilyttää riittä-

vän pitkältä ajalta tiedot. Tiedon ajoituksella tarkoitetaan sitä, että tiedon tulee olla käytettävissä ennen kuin sitä tarvitaan ennustamiseen. Esimerkiksi jalkapallo-ottelun katsojamäärä voisi olla hyvä piirre, kun ennustetaan ottelujen tuloksia, mutta tämä tieto on käytettävissä vasta ottelujen jo alettua. Tämän takia sitä ei voida käyttää ottelujen tulosten ennustamiseen. Tiedon pitkäikäisyys tarkoittaa sitä, että piirteenä käytettävän tiedon tulisi säilyä mahdollisimman muuttumattomana, jotta käytettävä algoritmi pystyy löytämään luotettavasti yhteyden piirteen ja ennustettavan tuloksen välillä. Esimerkiksi lainanhaun yhteydessä henkilöiden palkat muuttuvat ajan kuluessa, jolloin parempi ja kestävämpi piirre on palkan suhde haettavan lainan määrään. Yleinen tapa parantaa piirteen kestävyyttä on käyttää suoraan tietolähteen tietojen sijasta näistä tiedoista johdettua piirrettä. (Kelleher ym. 2015, 32-37)

2.4 Yleisimmät algoritmityypit

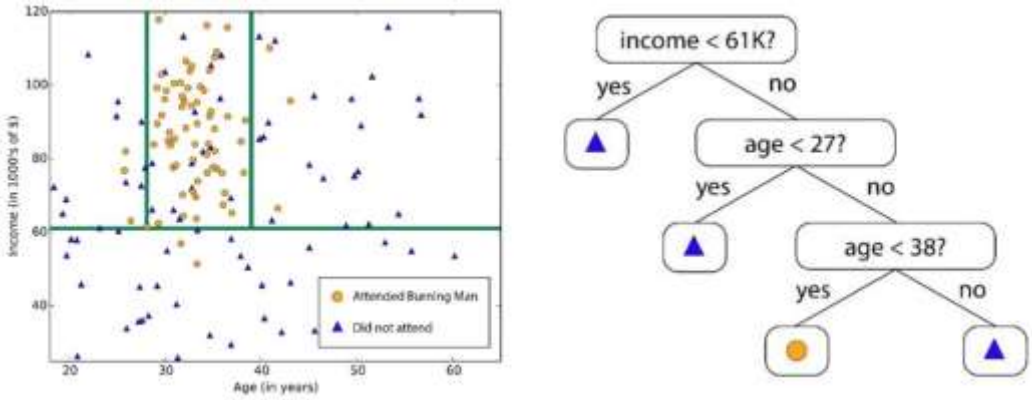
Marsland on kuvannut koneoppimisen algoritmien kuuluvan yleisesti kolmeen pääryhmään: ohjattu oppiminen, ohjaamaton oppiminen tai vahvistusoppiminen. Ohjatulla oppimisella tarkoitetaan sitä, kun opetusjoukosta tiedetään etukäteen haluttu tulos ja koneoppimisen malli opetetaan luomaan yhteys opetusjoukon piirteiden ja halutun tuloksen välille. Tämä on yleisimmin käytetty algoritmityyppi. Ohjaamattomassa oppimisessa taas haluttuja tuloksia ei määritellä etukäteen, vaan käytetty algoritmi etsii opetusjoukosta yhteisiä piirteitä ja ryhmittelee datan automaattisesti. Vahvistusoppiminen on näiden kahden välimuoto, jossa mahdolliset tulokset tiedetään etukäteen, mutta ne eivät ole oppimisalgoritmin tiedossa. Oppimisen yhteydessä algoritmille kerrotaan vain, jos tulos on positiivinen tai negatiivinen, ja algoritmi etsii, kunnes löytää parhaan mahdollisen tuloksen. (Marsland 2015, 5-6)

Tässä tutkimuksessa käytettiin ohjatun oppimisen menetelmää. Marsland on edelleen jakanut ohjatun oppimisen algoritmit sen tuottaman tuloksen tyypistä riippuen kolmeen pääryhmään: regressio, luokittelu ja poikkeamien havaitseminen. Regressio-algoritmit tuottavat tuloksena numeerisen tuloksen. Esimerkkinä regressio-algoritmeille sopivasta käyttötapauksesta on tuotteen hinnan ennustaminen tuotteen ominaisuuksien perusteella. Kolme yleistä regressio-algoritmia ovat lineaarinen regressio, päätöspuut ja neuraaliverkot. Luokittelu-algoritmit jakavat tuloksen kahteen tai useampaan luokkaan. Esimerkkinä luokittelu-algoritmeille sopivasta käyttötapauksesta on kerätyn asiakastiedon perusteella asiakkaiden jakaminen asiakassegmentteihin. Poikkeamien havaitsemisessa algoritmi pyrkii oppimaan normaalit tapaukset ja ilmoittaa, jos tulos on tästä merkittävästi poikkeava. Esimerkkinä tämä algoritmityypin käytöstä on luottokorttien poikkeuksellisen suuren käytön havaitseminen korttien käyttötiedoista. (Marsland 2015, 6-9)



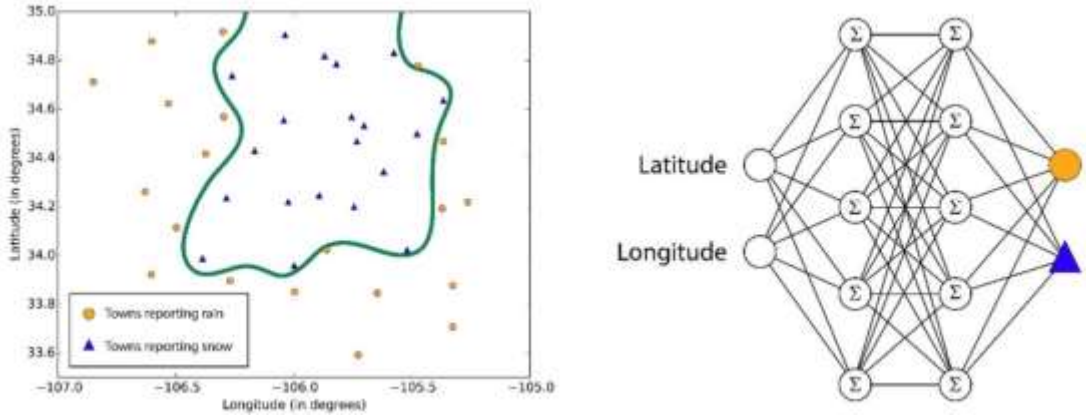
Kuva 4. Lineaarinen regressio (Microsoft 2017a)

Lineaarinen regressio on yksinkertaisin koneoppimisen regressio-algoritmeista. Siinä aineiston piirteistä muodostetaan lineaarinen funktio halutun jatkuvan vastemuuttujan laskemiseksi. Tämä voidaan piirteiden määrästä riippuen havainnollistaa janana, tasona tai hypertasona. Lineaarinen regressio on yksinkertainen ja nopea käyttää, mutta on samalla moneen käytännön käyttöön liian karkea ja epätarkka. Lineaarinen regressio on sen yksinkertaisuuden vuoksi usein ensimmäinen algoritmi, jota kokeillaan ennen kuin muita algoritmeja siltä varalta, että se olisi riittävä kyseiseen ongelmaan. (Microsoft 2017a)



Kuva 5. Päätöspuun toimintaperiaate (Microsoft 2017a)

Päätöspuu-algoritmeja on useita erilaisia, mutta ne toimivat kaikki saman periaatteen mukaisesti. Niissä aineiston piirteet jaetaan mahdollisimman yhdenmukaisiksi alueiksi ja algoritmi muodostaa hierarkkisen puun päättelysääntöjä varten. Päätöspuut ovat siitä erikoisia, että niitä voi käyttää sekä regressioon että luokitteluun. (Microsoft 2017a)



Kuva 6. Neuraaliverkon toimintaperiaate (Microsoft 2017a)

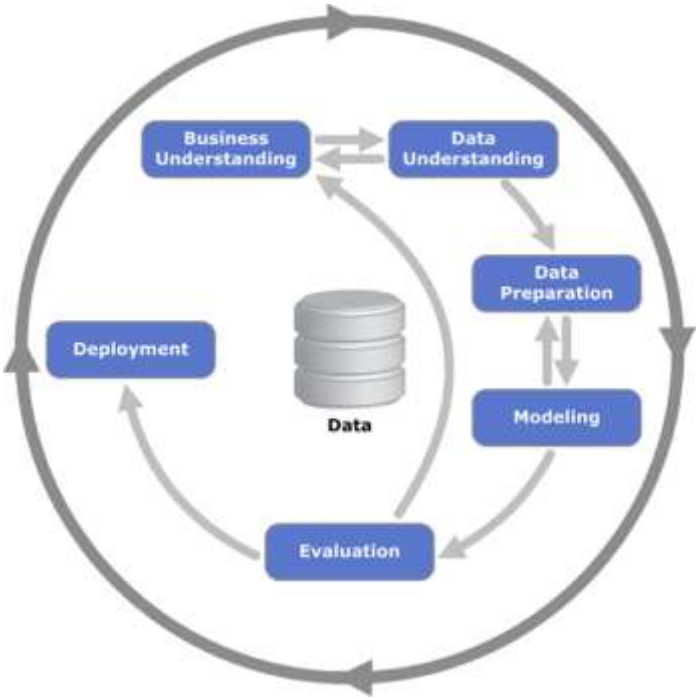
Neuraaliverkko on aivojen toimintaa jäljittelevä oppimisalgoritmi, jota voi käyttää sekä luokitteluun että regressioon. Kuvan 6 mukaisesti tulotiedot syötetään algoritmiin, joka käyttää nämä useammassa kerroksessa olevien painotettujen solmujen läpi ja lopuksi viimeisessä kerroksessa muodostaa tuloksen. Neuraaliverkot ovat yleisesti tarkkoja, mutta toisaalta tämän saavuttamiseksi algoritmin oppiminen on monia muita algoritmeja hitaampaa. Neuraaliverkoilla on lisäksi tavanomaista enemmän parametreja ja säätövaraa, mikä lisää hyvän mallin kehittämiseen kuluvaa aikaa. (Microsoft 2017a)

3 Koneoppimisen menetelmät

Ennustavassa analytiikassa ja koneoppimisessa eniten käytetty menetelmä on Cross Industry Standard Process for Data Mining (CRISP-DM) menetelmä. Menetelmän ensimmäinen versio kehitettiin vuonna 1996. Menetelmän avulla voi hallita tiedonlouhinnan projektia koko sen elinkaaren ajalta samalla tavalla kuin esimerkiksi ohjelmistotuotannossa elinkaarta hallitaan. Menetelmä pyrkii liittämään ennustava analytiikan ja koneoppimisen mallien kehittämisen muuhun organisaation ammattimaiseen toimintaan mm. sisällyttämällä menetelmään ulkoisen asiakkaan ja ottamalla mukaan muitakin vaiheita, kun itse analytiikan ja mallin kehittämisen. (Román 2016)

Menetelmää ei kehitetä enää aktiivisesti ja se on varsin vanha, mutta se soveltuu edelleen hyvin ylätasolla käytettäväksi. Ainoa laajemmin tunnettu vaihtoehtoinen ennustavan analytiikan menetelmä on SEMMA, mutta sen käyttö on ollut vähäistä verrattuna CRISP-DM-menetelmään. IBM on aloittanut vuonna 2015 CRISP-DM pohjautuvan ASUS-DM menetelmän kehittämisen. Tästä menetelmästä ei kuitenkaan löydy juurikaan tietoa, joten ei ole vielä selvää, miten tämä menetelmä saa jalansijaa ennustavaa analytiikkaa ja koneoppimista käyttävien parissa. (Wikipedia 2017a)

3.1 CRISP-DM-menetelmä



Kuva 7. CRISP-DM-menetelmän vaiheet (Ferguson 2016)

Ferguson (2016) on kuvannut CRISP-DM-menetelmään kuuluvan 6 vaihetta:

1. Liiketoiminnan ymmärtäminen
2. Datan ymmärtäminen
3. Datan valmistelu
4. Mallin kehittäminen
5. Mallin arviointi
6. Mallin julkaisu

Ensimmäisenä vaiheena CRISP-DM-menetelmässä on liiketoiminnan ymmärtäminen. Tässä vaiheessa kuvataan käyttötapaus, määritellään tavoitteet ja onnistumisen kriteerit sekä tehdään mallin kehittämiseen ja käyttöön liittyvän projektin suunnittelu. Tätä vaihetta seuraa datan ymmärtämisen vaihe, jossa selvitetään mitä omia ja ulkoisia tietolähteitä on käytössä ja mitä tietoa näistä löytyy. Koneoppimisen käytössä on olennaista tietää, miten paljon tietoa on käytettävissä. Lisäksi on tärkeää selvittää missä muodossa tieto on ja onko tieto tallennettu tiedostoihin, tietokantoihin vai onko tietolähteenä reaaliaikainen suoratoisto. Tähän vaiheeseen kuuluu myös tiedon tilastollinen analyysi, jossa siitä selvitetään tilastolliset tunnusluvut, haetaan mahdollisia vääristymiä ja visualisoidaan esimerkiksi histogrammien avulla. Tässä vaiheessa on tärkeää ymmärtää missä määrin data on puutteellista, virheellistä tai epäjohdonmukaista.

Kolmantena vaiheena on datan valmistelu. Tämän vaiheen tarkoituksena on käsitellä tieto siten, että sen voi syöttää tietona koneoppimisen algoritmeille mm. suodattamalla (virheellisten tietueiden poistaminen), siivoamalla (puuttuvien arvojen lisääminen) ja suorittamalla erilaisia konversiota ja muunnoksia. Hyvin tärkeä tehtävä tässä vaiheessa on datasta käytettävien piirteiden valinta. Tällä pyritään valitsemaan datasta ne datan piirteet, joilla on suurin merkitys ja samalla poistamaan vähiten merkitsevät datasta (turhat piirteet hidastavat oppimista). Neljäntenä vaiheena on koneoppimisen mallin kehittäminen. Ohjatun oppimisen mallin kehittämiseen kuuluvat algoritmin valinta, sen parametrien virittäminen, mallin opettaminen osalla datasta ja mallin testaaminen lopulla datalla. Yleinen nyrkkisääntö on käyttää 75 % datasta opettamiseen ja 25 % mallin arviointiin. Ohjaamattoman oppimisen tapauksessa algoritmit analysoivat dataa eri tilastollisin keinoin niin, että siitä voidaan löytää yhteyksiä ja ryhmiä.

Viidentenä vaiheena on mallin arviointi. Mallin arvioinnissa tavoitteena on ymmärtää miten tarkasti malli ennustaa tuloksen eli miten usein opetettu malli ennustaa oikein ja miten usein väärin. Luokittelun yhteydessä voidaan laskea oikean ja väärien positiivisten ja negatiivisten määrät ja regressiota käytettäessä hyviä tilastollisia suureita ovat neliövirhe ja selitysaste. Viimeisenä vaiheena on mallin julkaisu. Kun malli on valmis ja täyttää sille asetut vaatimukset, julkaistaan se käytettäväksi. Tämä voi olla yksinkertaisimmillaan mallin käyttö taulukkolaskentaohjelmassa kuten Excel ja toisessa ääripäässä mallia käytetään reaaliaikaisesti esimerkiksi Apache Spark-ympäristössä.

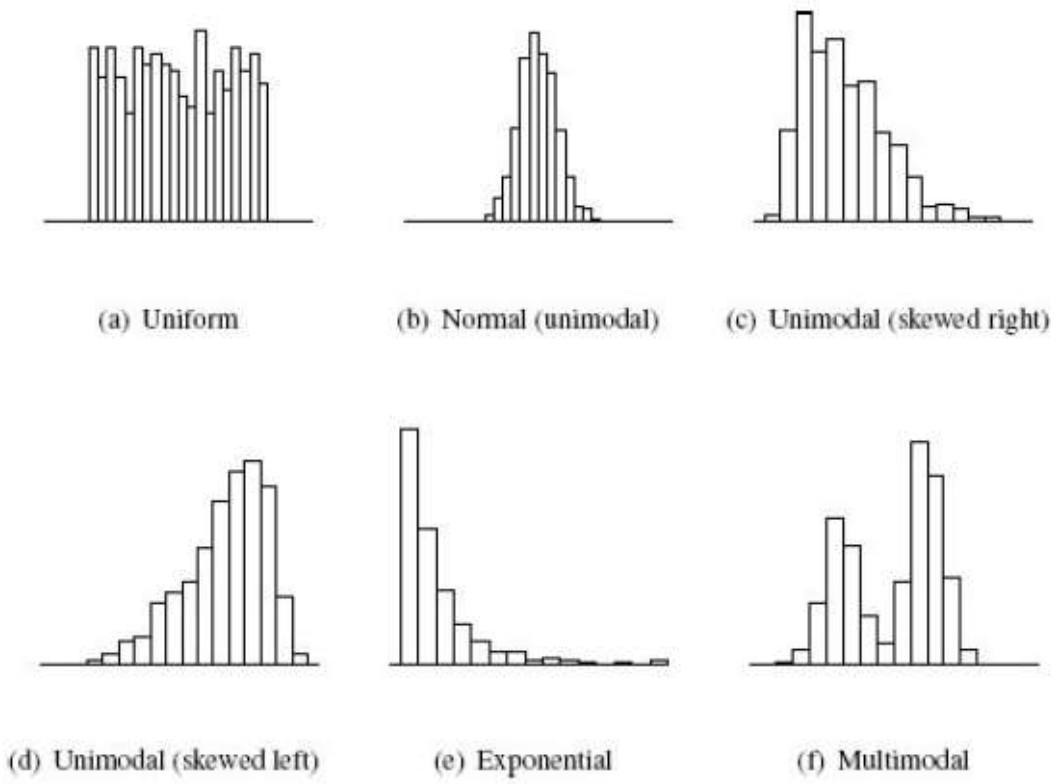
3.2 Datan ymmärtäminen

Koneoppimisen menetelmiä käytettäessä on tärkeää, että käytettävä data on mahdollisimman laadukasta, jotta opetettava malli löytää siitä todelliset piirteiden väliset yhteydet. Tätä varten datan ymmärtämisessä ja käsittelyssä tuotetaan datalaadun raportti, tunnistetaan raportin avulla mahdolliset laatuongelmat ja lopuksi käsitellään havaitut laatuongelmat.

Datan ymmärtämisessä ensimmäisenä tehtävänä on muodostaa datalaadun raportti. Tämä raportti on tärkeä työkalu käytettävän datan ymmärtämisessä. Raportissa tehdään erikseen taulukot datan jatkuville piirteille ja sen luokitteleville piirteille. Taulukoissa kuvataan kunkin piirteen keskeiset tilastolliset ominaisuudet kuten keskiarvo, tyyppiarvo ja mediaani ja vaihtelua kuvaavat tunnusluvut kuten keskihajonta ja varianssi. Taulukoita voidaan täydentää piirteiden jakaumia visualisoivilla kuvaajilla kuten histogrammit, laatikkokuviot ja hajontakuviot.

Jatkuvien piirteiden osalta taulukkoon kuuluvat tunnusluvuista minimi, alakvartiili, keskiarvo, mediaani, yläkvartiili, maksimi, keskihajonta, otoskoko, puuttuvien arvojen määrä ja kardinaliteetti eli uniikkien arvojen määrä. Luokittelevien piirteiden osalta taulukkoon kuuluvat tyyppiarvo ja tämän frekvenssi, puuttuvien arvojen määrä ja kardinaliteetti. (Kelleher ym. 2015, 51.)

Datalaadun raportti antaa monipuolisesti tietoa aineistona käytettävästä tiedosta. Tutkimalla sitä voidaan tunnuslukujen avulla ymmärtää dataa ja havaita siinä olevia virheitä. Jatkuvista piirteistä tarkistetaan raportista piirteen keskiarvo ja keskihajonta, minimi- ja maksimiarvot sekä arvioidaan histogrammin avulla datan jakauma. Luokittelevista piirteistä tunnistetaan raportin avulla dominoiko jonkin piirteen arvo. Tämä onnistuu piirteen tyyppiarvon ja sen frekvenssin avulla sekä histogrammia katsomalla.



Kuva 8. Yleiset histogrammista havaittavat jakaumat (Kelleher ym. 2015, 62.)

Histogrammiin liittyy usein jokin yleinen jakauma ja tämä kannattaa yrittää tunnistaa datasta, koska tästä on hyötyä laatuongelmien tunnistamisessa ja käsittelyssä. Yleisiä histogrammista tunnistettavia jakaumia ovat tasajakauma, normaalijakauma, positiivinen tai negatiivinen vino jakauma, eksponenttijakauma ja monihuippuinen jakauma.

Tasajakauma tarkoittaa, että piirteen arvot ovat tasaisesti jakautuneet koko arvovälille. Tällainen piirre ei yleensä tarjoa mitään mielenkiintoista tietoa mallille. Monesti tasajakautuneessa aineistossa on kyse esimerkiksi otosalkioille muodostetusta tunnistetiedosta, jossa jokaisella alkiolla on eri arvo.

Piirteet, joiden arvot noudattavat normaalijakaumaa, tunnistaa siitä, että arvot keskittyvät keskiarvon ympärille symmetrisellä vaihtelulle tämän keskiarvon molemmin puolin. Moni luonnollinen ilmiö kuten satunnaisesti valittujen mies- tai naishenkilöiden pituudet tai painot jakaantuvat yleensä normaalijakauman mukaan. Normaalijakauman tunnistaminen on yleensä hyvä asia, koska moni koneoppimisen algoritmi toimii hyvin normaalijakautuneella datalla. Vinossa jakaumassa datassa esiintyy normaalijakaumasta poiketen joko hyvin matalia tai suuria arvoja. Esimerkiksi henkilöiden palkoissa on tavallista, että joillekin henkilöille maksetaan hyvin korkeaa palkkaa ja tämä aiheuttaa vinoutuman muuten normaalijakautuneessa datassa.

Eksponenttijakautuneessa datassa matalien arvojen esiintyvyys on suuri, mutta hyvin nopeasti arvojen suuretessa tippuu niiden lukumäärä datassa. Koska esimerkiksi hyvin harva henkilö menee monta kertaa naimisiin, jakautuu tämä tieto eksponenttijakauman mukaan. Eksponenttijakautuneessa datassa esiintyy usein poikkeavia arvoja, jotka pitää tarkistaa, kun data käsitellään.

Monihuippuinen jakauma tarkoittaa käytännössä sitä, että datassa on useampi toisistaan riippumattomat arvovälit, joille arvot keskittyvät. Esimerkiksi satunnaisten henkilöiden pituusmittauksissa on todennäköistä, että miesten ja naisten arvot keskittyvät eri arvoväleihin. Monihuippuinen jakauma voi olla sekä huolestuttava että hyvä asia. Huonona puolena on, että keskeiset tunnusluvut kuten keskiarvo voivat sijaita huippujen välissä eivätkä kuvasta dataa oikealla tavalla. Toisaalta, jos datan eri ryhmät ovat merkitseviä ennustettavan osalta, toimii kyseinen piirre hyvin. Jos esimerkiksi pituusarvoista yritetään ennustaa henkilön sukupuolta, on eduksi, että pituusarvot jakautuvat monihuippuisen jakauman mukaan. (Kelleher ym. 2015, 61-63)

3.2.1 Laatuongelmien tunnistaminen

Data ymmärtämisessä ja käsittelyssä toisena vaiheena on tunnistaa mahdolliset laatuongelmat datassa. Laatuongelmat havaitaan tutkimalla edellisessä vaiheessa muodostettua datalaadun raporttia. Laatuongelma voi johtua joko virheellisestä datasta tai myös virheettömästä datasta. Esimerkiksi poikkeava arvo voi johtua datan muodostamisessa tapahtuneesta virheestä, mutta toisaalta kerättyyn tietoon voi luonnollisesti kuulua toisistaan

poikkeavia arvoja. Jos data on virheellistä, tulee se korjata välittömästi sekä datalaadun raportti muodostaa uudestaan. Jos taas laatuongelma ei johdu virheestä, ei ongelmaa tarvitse välttämättä korjata paitsi, jos käytettävä algoritmi sitä edellyttää. Tietyt algoritmit ovat esimerkiksi herkkiä puuttuville arvoille ja tällaista algoritmia käytettäessä tulee data korjata ennen datan käyttöä mallin opettamisessa.

Kaikki havaitut laatuongelmat kirjataan datan laatusuunnitelmaan. Jokaisesta laatuongelmasta kirjataan mikä piirre kyseessä, mikä laatuongelma kyseessä ja myöhemmin lisätään mahdolliset tavat korjata tai käsitellä ongelma. Yleisimmät laatuongelmat datassa ovat puuttuvat arvot, odottamaton kardinaliteetti ja poikkeavat arvot.

Ensimmäisenä tutkitaan, onko datassa piirteitä, joilla esiintyy merkittävästi puuttuvia arvoja. Datalaadun raportissa puuttuvien määrän sarakkeen avulla on helppo tunnistaa ne piirteet, joiden osalta arvoja puuttuu merkittävästi. Puuttuvien arvojen kohdalla täytyy tunnistaa kunkin piirteen kohdalla syy miksi arvoja puuttuu. Esimerkiksi käsin syötetyn datan kohdalla kyse voi olla syöttövirheestä, mutta toisaalta erityisesti luottamuksellisen ja sensitiivisen tiedon osalta nämä tiedot on saatettu jättää tietoisesti pois esimerkiksi, jos henkilö ei ole antanut lupaa tietojen tallentamiseen. Toinen yleinen syy puuttuville arvoille on se, että kyseistä tietoa ei alun perin ole kerätty ja sitä esiintyy datassa vasta tietyn ajanhetken jälkeen kerätyissä alkioissa.

Toinen tutkittava asia on, esiintyykö piirteissä odottamattomia kardinaliteetteja. Kardinaliteetilla tarkoitetaan uniikkien arvojen määrää tietyn piirteen tiedoissa. Odottamaton kardinaliteetti on merkki datassa piilevästä ongelmasta. Ensimmäisenä on tarkistettava, löytyykö datassa piirteitä, joilla kardinaliteetti on 1. Tämä tarkoittaa sitä, että datassa tämän piirteen osalta kaikilla alkioilla on sama arvo, eikä kyseisen piirteen osalta datassa ole mitään hyödyllistä informaatiota. Tällaisten piirteiden osalta tarkistetaan ja korjataan mahdolliset virheet, mutta mikäli tämä ei johdu virheestä, poistetaan piirre datasta. Seuraavaksi tarkistetaan, ettei datassa ole virheellisesti jatkuvina merkittyjä piirteitä. Jatkuvilla piirteillä arvojen määrä on yleensä lähes yhtä suuri kuin alkioiden määrä. Jos jatkuvalla piirteellä uniikkien arvojen lukumäärä on pieni, tulisi tämä tutkia mahdollisten virheiden varalta ja tarvittaessa korjata. Lasten lukumäärä on esimerkki normaalista tilanteesta, jossa jatkuvalla piirteellä on vain vähän uniikkeja arvoja. Toisaalta sukupuoli merkitään usein arvoilla 0 ja 1 ja tällöin piirre merkitään helposti jatkuvaksi piirteeksi, vaikka kyseessä onkin oikeasti luokitteleva piirre.

Kolmas tutkittava asia on, löytyykö datassa yllättävän korkeita kardinaliteetteja. Esimerkiksi sukupuolen kohdalla suuri kardinaliteetti voi tarkoittaa sitä, että sama tieto on vahingossa merkitty datassa eri tavoilla (esimerkiksi nainen merkitty sekä koodeilla f, F että female). Tällaiset virheet datassa pitää korjata ja yhdenmukaistaa luokittelussa käytetyt koodit. Viimeisenä tutkitaan, löytyykö piirteitä, joilla kardinaliteetti on erittäin suuri (yli 50). Vaikka tämä olisi datassa normaalia, ovat tällaisten piirteiden käsittely vaikeaa monelle koneoppimisen algoritmille ja tällaiset piirteet tulee merkitä datan laatusuunnitelmaan.

Viimeisenä ryhmänä analysoidaan, onko datassa merkittävästi poikkeavia arvoja. Poikkeavilla arvoilla tarkoitetaan sellaisia arvoja, jotka eroavat merkittävästi piirteen arvojen keskiarvosta. Virheelliset poikkeavat arvot datassa johtuvat yleisesti datan keruussa esiintyvästä kohinasta tai siinä tapahtuneesta virheestä. Näppäilyvirhe, jonka seurauksena esimerkiksi luku 1000 tallentuu lukuna 100000, on esimerkki tällaisesta virheestä. Virheettömät poikkeavat arvot ovat toisaalta arvoja, jotka ovat harvinaisempia datassa, mutta niissä ei ole sinänsä mitään epänormaalia. Miljonäärin tulot ovat esimerkiksi poikkeavia verrattuna keskimääräisiin tuloihin, mutta eivät johdu virheestä.

Poikkeavia arvoja voidaan havaita datassa kahdella tavalla. Ensimmäinen tapa on tarkistaa ovatko piirteen minimi- ja maksimiarvot odotetut. Tämä vaatii ymmärrystä ja tietämystä tutkittavasta piirteestä. Esimerkiksi negatiivinen minimiarvo henkilön iälle on hyvin todennäköisesti korjausta vaativa virhe. Tällaiset virheet joko korjataan käytettävässä datassa tai kyseinen tieto poistetaan datasta. Lisäksi poikkeavia arvoja voidaan löytää tarkistamalla ovatko mediaanin, minimin, maksimin ja kvartiilien erot samankokoiset. Jos esimerkiksi yläkvartiilin ja maksimin ero on suuri, on maksimiarvo todennäköisesti poikkeava. Toinen tapa havaita tällaisia poikkeavia arvoja on tarkastella arvojen jakauman histogrammeja. Jos histogrammi osoittaa eksponentti- ja vinojakauman, on se merkki siitä, että datassa on poikkeavia arvoja. Tällaiset poikkeavat arvot eivät yleensä johdu virheestä ja tällaiset havainnot kirjataan tiedoksi datan laatusuunnitelmaan. (Kelleher ym. 2015, 66-70)

3.2.2 Piirteiden väliset yhteydet

Piirteiden analysoinnissa on myös olennaista selvittää löytyykö piirteitä, joiden välillä on vahva yhteys. Jos datassa kahden piirteen välillä on vahva yhteys, riittää yleensä, että ennustamisessa käytetään vain toista näistä. Tässä kappaleessa esitellään keinoja havaita ja havainnollistaa yhteys jatkuvien piirteiden välillä, yhteys luokittelevien piirteiden välillä, yhteys jatkuvan ja luokittelevan piirteen välillä sekä hyödyntää laskettua korrelaatiota.

Kahden jatkuvan piirteen välisen yhteyden voi arvioida käyttämällä hajontakuviota. Mikäli kuvaajassa pisteet keskittyvät kuvaajan diagonaalin ympärille, on piirteiden välillä vahva yhteys. Toinen, laskennallinen, tapa selvittää kahden jatkuvan piirteen välinen yhteys on laskea niiden välinen korrelaatio. Piirteiden välinen laskettu korrelaatio on arvoasteikolla [−1, 1], jossa arvot lähellä ääripäitä merkitsevät yhteyttä piirteiden välillä ja arvo lähellä nollaa merkitsee että piirteet ovat toisistaan riippumattomat. On kuitenkin tärkeä tiedostaa, että korrelaatio kuvaa vain lineaarista yhteyttä piirteiden välillä eikä paljasta ollenkaan muunlaisia yhteyksiä. Korrelaatiota käytettäessä on olennaista myös muistaa, että siitä ei välttämättä seuraa kausaliteetti.

Kahden luokittelevan piirteen välisen yhteyden visualisointi on hiukan monimutkaisempaa. Ensin ensimmäisen piirteen luokkien esiintyvyys visualisoidaan pylväskaavioiden avulla. Tämän jälkeen samat pylväskaaviot muodostetaan siten että datasta käytetään vain toisen piirteen tietyn luokan omaavat alkiot. Näitä uusia kuvaajia verrataan ensimmäisiin kuvaajiin. Jos kuvaajien jakaumat näyttävät samanlaiselta, tarkoittaa tämä, ettei piirteiden välillä ole yhteyttä ja jos jakaumat poikkeavat merkittävästi toisistaan, on piirteiden välillä yhteys.

Jatkuvan ja luokittelevan muuttujan yhteyden selvittämisessä toimitaan melko samalla lailla kuin kahden luokittelevan piirteen tapauksessa. Jatkuvasta muuttujasta muodostetaan ensin histogrammi kaikilla alkioilla ja tämän jälkeen suodatetut histogrammit jokaiselle luokittelevan piirteen luokalle. Jos kaikki histogrammit näyttävät samankaltaista jakaumaa, ei piirteiden välillä ole yhteyttä ja jos jakaumat poikkeavat merkittävästi toisistaan, on piirteiden välillä vahva yhteys. (Kelleher ym. 2015, 77-92)

3.3 Datan ja laatuongelmien käsittely

Kun datan laatuongelmat on tunnistettu ja kirjattu laatusuunnitelmaan, voidaan suorittaa muutama yleinen toimenpide datan laadun parantamiseksi. Seuraavassa käsitellään puuttuvien ja poikkeavien arvojen käsittelyä lyhyesti.

Puuttuvien arvojen käsittelyssä yleisiä keinoja ovat piirteen poistaminen kokonaan, uuden laskennallisen piirteen muodostaminen, alkioiden poistaminen, imputointi ja puuttuvien arvojen laskeminen koneoppimisen keinoin. Yksinkertaisin toimenpide näistä on poistaa kokonaan piirteet, joilla puuttuu arvoja. Tämä on kuitenkin hyvin karkea ja sen seurauksena menetetään monesti turhan paljon dataa.

Toinen tapa on muodostaa uusi laskennallinen piirre alkuperäisen piirteen pohjalta. Uusi piirre kertoo esimerkiksi, oliko datassa alkuperäistä piirrettä vai ei. Yleensä tätä toimenpidettä käytettäessä alkuperäinen piirre poistetaan datasta. Datasta voidaan myös poistaa ne alkiot, joilta puuttuu arvoja. Tämäkin keino voi aiheuttaa turhan suuren datamäärän menetyksen ja sen riskinä on myös, että data vääristyy, mikäli arvojen puuttuminen ei ole satunnaista. Imputointi tarkoittaa sitä, että puuttuvat arvot korvataan uusilla riittävän uskottavilla ja tarkoilla arvoilla. Yleisiä tapoja ovat käyttää jatkuville piirteille piirteen keskiarvoa tai mediaania ja luokitteleville piirteille piirteen tyyppiarvoa. Imputointia ei pidä käyttää, jos puuttuvia arvoja on paljon, koska se vääristää helposti datan liian lähelle piirteen keskiarvoa (hyvä harkita muita keinoja, jos puuttuvia arvoja on yli 30 %).

Yleinen suositus on käyttää mahdollisimman yksinkertaisia keinoja ja vain tarvittaessa siirtyä monimutkaisempiin keinoihin. Imputointi tuottaa yleisesti kelpo tuloksia eikä sen käytössä esiinny samanlaista tiedon hukkaamista kuin poistamalla tietoa, mutta toisaalta imputointi saattaa vääristää tietoa ja estää ennustavaa mallia löytämästä todellista yhteyttä piirteiden ja ennustettavan arvon välillä.

Poikkeavien arvojen käsittelemisen keinoista yleisin on leikata valitun kynnysarvon ylittävät tai alittavat arvot pois datasta. Yleisiä menettelyjä tällöin on käyttää kynnysarvoina kvartiilien arvoja, joihin lisätään niiden erotus sekä käyttää kynnysarvona keskiarvoa, johon lisätään keskihajonta kerrottuna kahdella. Jälkimmäinen toimii hyvin normaalijakautuneella datalla. Yleinen suositus on leikata poikkeavia arvoja datasta vain, jos opetettu malli toimii huonosti poikkeavista arvoista johtuen. (Kelleher ym. 2015, 73-76)

3.3.1 Piirteiden valinnan tekniikat

Koneoppimisessa ja tilastotieteessä piirteiden valinnalla tavoitellaan mallin yksinkertaistamista, nopeampaa oppimisaikaa, moniulotteisuuteen liittyvien ongelmien välttämistä ja ylisovittumisen välttämistä. Yleisesti menetelmien tavoitteena on tunnistaa mallin kannalta olennaisimmat piirteet. Lisäksi näiden menetelmien avulla pyritään tunnistamaan ne piirteet, jotka ovat mallin kannalta turhia ja voidaan poistaa ilman merkittävää tiedon menettämistä. On hyvä tiedostaa, että sinänsä itsenäisenä merkittävä piirre voi olla turha yhdessä muiden piirteiden kanssa, jos piirteiden välillä on vahva yhteys eli korrelaatio. Piirteiden valintaa käytetään yleisimmin tapauksissa, joissa otosjoukon koko on pieni. (Wikipedia 2017b)

Käytettävän menetelmän valinnassa on olennaista huomioida kunkin menetelmän tukemat tutkittavien piirteiden ja tuloksen tyypit, koska tietyt tekniikat toimivat vain tietyillä tietotyypeillä. Numeerisille ja loogisille arvoille soveltuvia tekniikoita ovat Pearsonin korrelaatio, Kendallin korrelaatio, Spearmanin korrelaatio ja Fisherin pisteytys. Keskinäisen informaation pisteytyksen, khii-neliön statistiikan sekä lukumääriin perustuva piirteiden valinnan tekniikoiden kanssa voi käyttää mitä tahansa tietotyyppejä. (Microsoft 2017b)

Näiden lisäksi kaksi yleistä menetelmää ovat lineaarinen erotteluanalyysi (LDA) ja pääkomponenttianalyysi (PCA). Lineaarinen erotteluanalyysi on ohjatun oppimisen tekniikka, jonka avulla voidaan luokitella jatkuvia muuttujia. Tämä tekniikka tunnistaa datasta sen piirteiden kombinaation, joka parhaiten erottelee ryhmät. Pääkomponenttianalyysi on tekniikka, joka tunnistaa datasta piirteet, joissa on eniten informaatiota ja variaatiota ja ovat näin hyödyllisimmät mallin opettamisessa. (Microsoft 2017b)

3.4 Mallin kehittäminen

Koneoppimisen käytössä oikean algoritmin valinta on erittäin tärkeää. Valitun algoritmin tulee olla ongelmaan ja käytettävälle data sopiva, niin että vältetään esimerkiksi ali- että ylisovittumiseen liittyvät ongelmat. Microsoft on suositellut koneoppimisen algoritmin valinnassa huomioitavina tekijöinä mallin tarkkuutta, opetusaikaa, datan lineaarisuutta, piirteiden lukumäärää ja algoritmin parametrien lukumäärää. Mallin tarkkuuden osalta on tärkeää, ettei algoritmi tuota turhan tarkkaa tulosta. Jos tuloksena riittää karkea arvio, on parempi valita vähemmän tarkka algoritmi, jonka opettaminen on nopeampaa. Eri algoritmeilla opetusaika vaihtelee, joten jos käytössä on iso opetusjoukko, voi olla järkevää valita algoritmi, joka oppii nopeammin. Tosin useimmiten nopea oppiminen tarkoittaa vähemmän tarkkaa mallia.

Useimmat algoritmit hyödyntävät lineaarisuutta eli sitä, että datan voi jakaa suoran viivan avulla ja on tärkeää tunnistaa algoritmia valittaessa oman aineiston lineaarisuus. Mallissa käytettävien piirteiden lukumäärällä on merkitystä, koska tietyt algoritmit soveltuvat toisia paremmin datalle, jossa on paljon piirteitä. Tukivektorikone on esimerkki algoritmista, jolle iso määrä piirteitä ei ole ongelma. Algoritmin parametrien avulla voidaan opettamista virittää sekä tarkkuuden että opetusajan suhteen, joten algoritmin valinnassa tulee huomioida tämä. Mitä enemmän parametreja algoritmilla on, sen isompi mahdollisuus käyttäjällä on vaikuttaa mallin hyvyyteen, mutta samalla myös algoritmin virittäminen vaikeutuu. (Microsoft 2017a)

	Lineaarinen regressio	Päätöspuut	Neuraaliverkko
Tarkkuus	Karkea	Tarkka	Tarkka
Opetusaika	Lyhyt	Keskipitkä	Pisin
Lineaarisuus	Kyllä	Ei	Ei
Parametrien lukumäärä	4	5	9

Taulukko 1. Regressio-algoritmien vertailu (Microsoft 2017a)

Taulukossa 1 on esitetty lyhyesti tässä tutkimuksessa käytettyjä regressio-algoritmeja ja miten ne täyttävät kuvatut valintakriteerit. Piirteiden lukumäärä on jätetty taulukosta pois, koska valitut regressio-algoritmit eivät tämän suhteen eroa toisistaan. Lineaarinen regressio on näistä algoritmeista yksinkertaisin mutta samalla nopein. Neuraaliverkko on näistä algoritmeista tarkin ja monipuolisin, mutta samalla sen opettaminen on hitaampaa kuin muiden regressio-algoritmien opettaminen. Neuraaliverkko-algoritmin virittäminen vaatii näistä eniten asiantuntemusta. (Microsoft 2017a)

3.5 Mallin arviointi

Ennustavan mallin arvioinnilla tavoitellaan kolmea asiaa. Arvioinnissa halutaan varmistaa, että malli on soveltuvin tarkoitukseensa, laskea tilastolliset tunnusluvut miten hyvin malli toimii sekä vakuuttaa mallin käyttäjät siitä, että se soveltuu heidän käyttötarkoitukseen. Kahdessa ensimmäisessä on kyse eri mallien toimivuuden mittaamisesta ja vertailusta. Kokonaisarvioinnissa on kuitenkin tärkeää muistaa, että tarkkuuden lisäksi mallin käytön kannalta olennaista on myös, miten nopeasti malli laskee tuloksen ja miten nopeasti se voidaan tarvittaessa opettaa uudestaan. Esimerkiksi lääketieteellisessä käytössä tuloksen tarkkuudella on suurempi merkitys, kun taas markkinoinnissa tarkkuudella ei ole yhtä suuri merkitys. (Kelleher ym. 2015, 398-399)

Tässä tutkimuksessa tutkittiin regressio-algoritmien käyttöä, jolloin ennustavan mallin tuloksena on jatkuva arvo. Yleisin lineaarisen regression yhteensopivuuden asteen tunnusluku on jäännösneliösumma (SSE). Mallin arvioinnissa tämä lasketaan laskemalla mallin ennustama arvo testijoukon alkioille ja vertaamalla näitä testijoukon tosiasiallisiin arvoihin. Laskemalla tästä keskineliövirheen (MSE) saadaan keskimääräinen laskettujen ja todellisten arvojen erotus. Keskineliövirhettä voidaan käyttää useamman mallin vertailussa. Tämän tunnusluvun arvot ovat välillä [0, ∞], ja pienempi arvo tarkoittaa parempaa yhteensopivuutta ja mallin suorituskykyä. Yleinen kritiikki keskineliövirhettä kohtaan on se, että sen arvo ei itsessään tarkoita mitään eikä se kerro miten iso virhe mallin arvossa saattaa olla.

Laskemalla keskineliövirheen neliöjuuren (RMSE) voidaan arvo muuttaa samalle asteikolle kuin itse arvot ja näin tunnusluvun arvot ovat vertailukelpoiset mallin arvojen kanssa. Keskineliövirheen neliöjuuren käytössä pienempi arvo tarkoittaa parempaa yhteensopivuutta ja suorituskykyä. Koska laskennassa virheet on korotettu potenssiin kaksi, korostaa tämä mallin suurimpia virheitä ja samalla laskettua kokonaisvirhettä. Tämä tunnusluvun ominaisuutta voi käyttää hyväksi, koska se voi paljastaa systemaattisen virheen. Vaihtoehtoinen tunnusluku, joka ei ole yhtä herkkä mallissa oleville isoille virheille, on keskimääräinen absoluuttinen poikkeama (MAE). Tämän tunnusluvun arvot ovat myös välillä [0, ∞], ja pienempi arvo tarkoittaa parempaa yhteensopivuutta ja mallin suorituskykyä.

Edellä kuvattujen tunnuslukujen käytössä yhtenä haasteena on se, että niiden tulkitseminen vaatii arvojen arvovälin tuntemista ja ymmärtämistä. Nämä tunnusluvut yksinään eivät kerro vielä miten hyvin malli toimii. Selitysaste (R2) on normalisoitu tilastollinen tunnusluku, jonka arvoväli ei ole riippuvainen mallin arvojen arvovälistä, vaan sen arvo on aina korkeintaan 1. Mitä lähempänä lukua 1 tunnusluvun arvo on, sen paremmin malli selittää datassa olevan variaation. R2-tunnuslukua käytetään kaikkein yleisimmin regressiomallien suorituskyvyn arvioinnissa, koska sen käyttö ei vaadi hyvää datan alkuperän tuntemusta mallin toimivuuden tulkitsemiseksi. R2-tunnusluvun kanssa tulee kuitenkin olla varovainen, koska isoilla määrillä piirteitä se antaa herkästi hyvältä vaikuttavia tuloksia, mutta malli ei silti välttämättä yleistä tulosta hyvin ja saattaa näin toimia huonosti oikeassa käytössä. (Kelleher ym. 2015, 442-447)

Yhteenvetona jatkuvan arvon ennustamisessa yleisimmät tunnusluvut ovat

- Selitysaste (R2)
- Keskineliövirheen neliöjuuri (RMSE)
- Keskimääräinen absoluuttinen poikkeama (MAE)

Regressiomallin arvioinnissa näistä käytetään yleisimmin selitysastetta ja keskineliövirheen neliöjuurta. Hackenbergin mukaan näiden molempien käyttö yhtä aikaa on hyödyllistä, koska ne antavat hiukan eri tietoa mallista ja sen toimivuudesta. Selitysasteen hyvä puoli on sen riippumattomuus arvojen skaalasta ja se on näin helppo ja nopea tulkita. Sen käyttöön liittyy kuitenkin riski huonoista tuloksista varsinkin, jos käytössä on paljon piirteitä ja tämän takia mallia on hyvä arvioida myös keskineliövirheen neliöjuuren avulla. Keskineliövirheen neliöjuuren osalta hyvän arvon voi saavuttaa vain, jos datassa ei ole suuria virheitä eikä mallissa ole systemaattista virhettä. Jos mallin tuloksena on R2-arvo lähellä yhtä ja RMSE arvo on datan arvoasteikolla tulkittuna matala, voidaan malli tulkita hyväksi ja toimivaksi. (Hackenberg 2.9.2016.)

4 Koneoppiminen pilvipalveluna

Machine Learning as a Service (MLaaS) palvelulla tarkoitetaan yleisesti sellaista pilvipalvelua, jolla käyttäjä voi toteuttaa koneoppimisen menetelmiä ilman, että käyttäjän tarvitsee tätä varten asentaa tätä varten ympäristöjä ja työkaluja. MLaaS-palvelu tarjoaa käyttäjälle työkalut tiedon käsittelyyn, koneoppimisen mallien tekemiseen ja näiden käyttämiseen palvelun käyttöliittymän ja rajapintojen avulla. Yksinkertaisimmillaan käyttäjä avaa palvelun selaimessa, lataa oman aineistonsa palveluun tai käyttää palvelussa valmiina olevaa aineistoa, ja palvelun työkalujen avulla suorittaa haluamansa tehtävän. (Techopedia 2017)

4 suurinta MlaaS-palveluntarjoajaa ovat Amazon, Google, IBM ja Microsoft, mutta tällä uudella ja nopeasti kehittyvällä markkina-alueella on myös paljon pienempiä startup-yrityksiä kuten esimerkiksi BigML, Datoin ja Natero. Startup-yritysten liiketoiminta ei yleensä perustu kilpailuun 4 ison palveluntarjoajan kanssa. Useimmiten ne keskittyvät ratkaisemaan hyvin jonkin erikoisemman ongelman, joka ei ainakaan vielä ole optimaalisesti tuettuna isompien yritysten laajemmalle kohdejoukolle tarkoitetuissa palveluissa. MLaaS-markkinan nopean kehitystahdin takia tällä hetkellä on vaikeaa tietää ja arvioida mikä palvelu on paras nyt ja lähitulevaisuudessa, koska markkinatilanne muuttuu jatkuvasti sekä uusien alan teknologiainnovaatioiden että palvelujen ominaisuuksien nopean kehittymisen myötä. (Quora 2017)

4.1 MLaaS-palvelun arviointikriteerit

Rakshith Begane (2017) on kirjoituksessaan Quora-palvelussa ehdottanut keskeisimmät kriteerit MlaaS-palvelun arvioimiseksi. Nämä hän on ryhmitellyt kolmeen pääryhmään:

1. Yleistäminen
2. Erikoistuminen
3. Asiantuntijatuki

Yleistämiseen Begane laskee mukaan seuraavat arviointikriteerit:

- Palvelu tarjoaa hyvän käyttöliittymän mallien tekemiseen, vertailuun ja tulosten visualisointiin kokeiden avulla.
- Palvelu tarjoaa hyvä työkalut käytettävän datan käsittelyyn ja muokkaamiseen.
- Hyvässä palvelussa käyttäjällä voi valita soveltuvimman algoritmin laajasta valikoimasta eikä palvelu rajoita tarjoamalla vain suppeaa valikoimaa algoritmeja.
- Palvelussa on hyvät työkalut tutkia data ja sen laatu sekä löytää siitä oikeat oppimisessa käytettävät piirteet.
- Palvelussa voi julkaista ja käyttää kehittämiään malleja suoraan palvelun tarjoaman rajapintojen avulla.
- Käyttäjä voi tarvittaessa laajentaa palvelun yllä mainittuja ominaisuuksia mm. R- tai Python-skripteillä.

Erikoistumisen osalta Begane esittää 3 arviointikriteeriä:

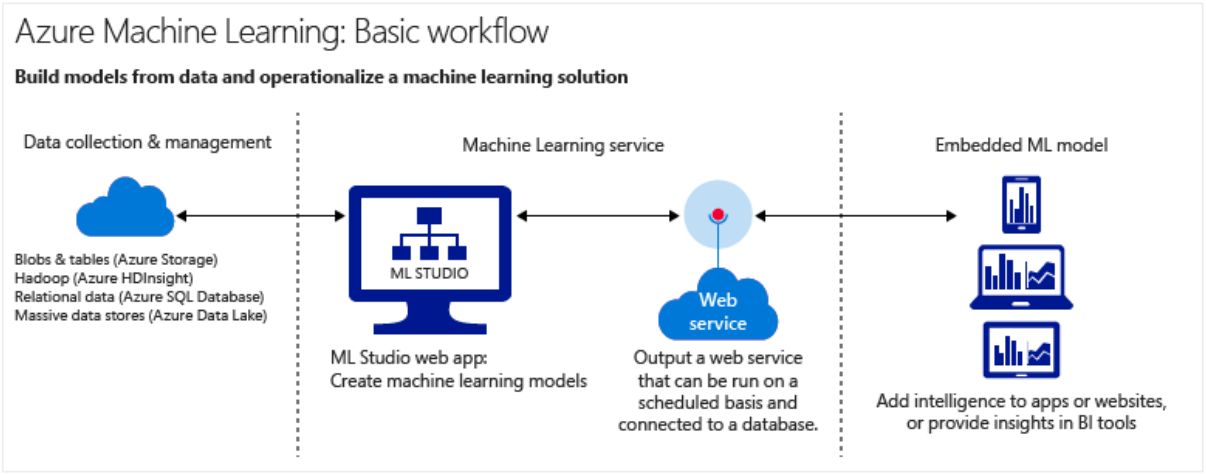
- Palvelussa on valmiit tietoaineistot eri toimialueille.
- Palvelussa löytyy valmiit komponentit yleisesti tunnettujen ongelmien ratkaisemiseksi.
- Palveluntarjoajalla on kyky tarjota toimialatuntemusta ja -osaamista tuetuille toimialoille.

Asiantuntija-avun osalta Begane mainitsee 3 keskeistä tukimuotoa:

- Palveluun kuuluu kehittäjätuki ja/tai -yhteisö.
- Palvelun käyttöön liittyvä tuki on riittävä.
- Palvelun tarjoama dokumentaatio on riittävä.

4.2 Azure Machine Learning

Azure Machine Learning on Microsoftin selainpohjainen palveluympäristö, jolla voi koostaa sen moduulien avulla ennustavan analytiikan ja koneoppimisen menetelmien mukaisia työnkulkuja. Graafisen käyttöliittymän Azure Machine Learning Studio avulla käyttäjä voi helposti ja nopeasti koostaa kokeen yleisimpiin koneoppimisen tehtäviin. Azure Machine Learning Studio-sovellusta täydentää Machine Learning API Service, jonka avulla käyttäjä voi julkaista tekemänsä ennustavan mallin REST-tyyppisenä Web Services-palveluna. (Microsoft 2017c)



Kuva 9. Azure Machine Learning (Microsoft 2017d)

Seuraavissa kappaleissa olen arvioinut miten hyvin Azure Machine Learning-palvelu täyttää Beganen asettamat arviointikriteerit. Yleistämisen osalta Azure Machine Learning täyttää laajasti kaikki edellä kuvatut arviointikriteerit. Azure Machine Learning Studio tarjoaa selainpohjaisen käyttöliittymän, joka ei vaadi lisäosien asentamista. Käyttöliittymän avulla käytännön kokeiden tekeminen on helppoa ja vaivatonta. Tietoja voi esikäsitellä sekä valmiilla peruskomponenteilla että liitettävällä skriptillä (R, Python). Tiedonkäsittelyn tuki on

laaja ja riittävä. Palvelu sisältää kaikki yleisimmin käytetyt algoritmit ja algoritmituki on hyvin laaja ja kattava. Palvelussa on ominaisuuksien valintaan laaja tuki yleisille tilastotieteen menetelmille kuten esimerkiksi lineaarinen erotteluanalyysi, pääkomponenttianalyysi ja erilaisia korrelaatioon perustuvia menetelmiä. Palvelu sisältää tuen ennustavan mallin julkaisemiseksi REST-tyyppisenä rajapintana.

Erikoistumisen osalta tuki on myös laaja. Palvelussa on useita kymmeniä valmiita esimerkkiaineistoja ja tietoaineistoja on helppo lisätä muista ulkoisista tietolähteistä. Cortana Intelligence Gallery tarjoaa hyvin laajan joukon valmiita sekä Microsoftin että palvelun käyttäjien luomia kokeita. Gallerian kokeita pystyy suodattamaan toimialan mukaan ja palvelussa on kattavasti kokeita yleisimmille toimialoille kuten kaupan, valmistuksen, pankin ja terveydenhuollon aloille.

Palvelussa on selvästi panostettu käytön helppouteen ja palvelun sivuilla löytyy paljon dokumentaatiota ja käyttöoppaita. Käyttäjille löytyy sekä käyttäjätuki että käyttäjien oma yhteisö, jolla käyttäjät voivat auttaa toisiaan. Kaikkiin uusimpiin kysymyksiin käyttäjäyhteisössä oli siihen tutustuessani vastattu, joten tämä näyttää olevan hyvin toimiva.

Palvelu täyttää kaikki Beganen asettamat kriteerit jopa yllättävän hyvin ja laajasti. MLaaS-palveluja on jonkin verran kritisoitu siitä, että ne ovat puutteelliset ja vajaat verrattuna itse koottuihin ympäristöihin, mutta tässä arvioinnissa en huomannut mitään sellaista olennaista puutetta tai rajoitetta, joka vahvistaisi tämän kritiikin. Varsinkin kun palvelussa on hyvä tuki oman sisällön ja skriptien käytölle, en löytänyt selkeää perustetta oman vastaavan ympäristön perustamiselle tämän tutkimuksen laajuutta vastaavassa käytössä.

4.3 Azure Machine Learning-palvelun moduulit

Azure Machine Learning-palvelussa kokeen työnkulku kootaan moduulien avulla. Jokainen moduuli suorittaa jonkin tietyn kokeen kannalta olennaisen koneoppimiseen liittyvän osatehtävän. Kokeen moduulit liitetään toisiinsa niin, että edeltävän moduulin tuottama tieto virtaa sisään siihen liitettyihin moduuleihin. Moduuli voi esimerkiksi tuottaa opetetun mallin, laskea tilastollisia tunnuslukuja, muuntaa dataa toiseen muotoon tai tallentaa tietoja ulkoiseen tietovarastoon.

Ohjatun oppimisen regressiomalliin liittyvässä kokeessa tarvitaan ainakin moduulit ”Split Data”, ”Train Model”, ”Score Model” ja ”Evaluate Model”. Lisäksi kokeeseen täytyy valita käytettävä regressio-algoritmi. (Microsoft 2017e)

4.3.1 Datan ymmärtäminen ja käsittely

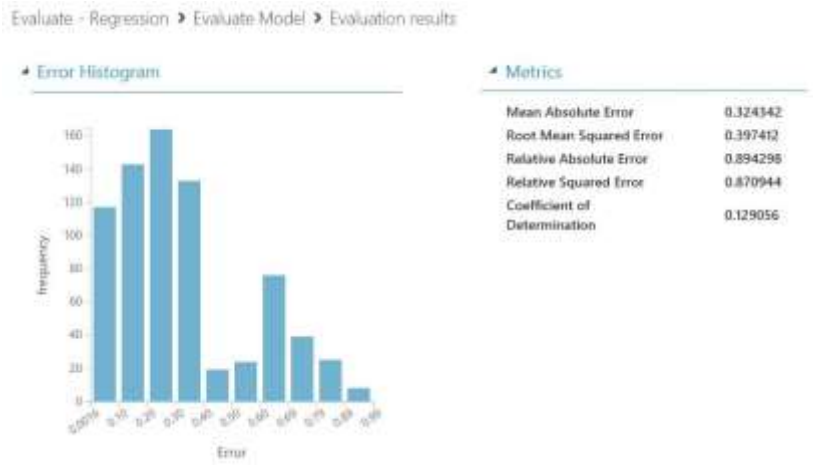
Datan ymmärtämisessä ja käsittelyssä ensimmäisinä tehtävinä on muodostaa datalaadun raportti ja tunnistaa sen avulla mahdolliset laatuongelmat datassa. Tätä varten voidaan käyttää palvelun tilastollisia moduuleja ”Summarize Data” ja ”Compute Elementary Statistics”. ”Summarize Data” tuottaa kattavan raportin, jossa on datan ja sen piirteiden keskeiset tilastolliset tunnusluvut ja kuvaajat. ”Compute Elementary Statistics”-moduulin avulla voidaan datasta laskea kullekin piirteelle valittu, yksittäinen tilastollinen tunnusluku kuten esimerkiksi keskiarvo tai keskihajonta. Piirteiden välistä yhteyttä voi tutkia mm. ”Compute Linear Correlation”-moduulin avulla.

Azure Machine Learning-palvelu sisältää valmiit moduulit yleisimmille piirteiden valinnassa käytettäville menetelmille. Kolme tärkeintä moduulia ovat ”Linear Discriminant Analysis”, ”Filter-Based Feature Selection” ja ”Principal Component Analysis”. Näitä voi sekä käyttää piirteiden tutkimisessa turhien piirteiden tunnistamiseksi moduulien laskemien tulosten avulla että käyttää suoraan moduulin tuottamaa tietojoukkoa mallin opettamisessa. Laatuongelmien korjaamisessa hyviä moduuleja ovat mm. ”Clean Missing Data” ja ”Replace Discrete Values” puuttuvien arvojen korjaamiseen ja ”Clip Values” poikkeavien arvojen poistamiseen.

Yleisiä moduuleja datajoukon muokkaamiseksi ja muuntamiseksi ovat mm. ”Select columns in data set” turhien piirteiden poistamiseksi ja “Edit metadata”, jolla voi esimerkiksi muuttaa jatkuvan piirteen luokittelevaksi. Dataa voi myös muokata joustavasti käyttämällä R- ja/tai Python-ohjelmointikieliä ”Execute R Script”- ja ”Execute Python Script”-moduulien avulla. (Microsoft 2017e)

4.3.2 Mallin arviointi

Mallin arvioinnissa käytetään yleisimmin “Evaluate Model”-moduulia, joka tuottaa yleisimmät tilastolliset tunnusluvut käytetyn algoritmin mukaisesti. ”Cross Validate Model”-moduulia voi käyttää ristiinvalidoinnissa eli moduulin avulla voi arvioida opetettavan mallin tuottamaa ennustevirhettä ja tämän avulla optimoida mallin parametreja. Lisäksi R- ja Python-skripteillä voi luoda omia tai käyttää valmiita kolmannen osapuolen kehittämiä tunnuslukuja mallin arvioinnissa. Arviointimoduulien tuloksen voi joko visualisoida tai tallentaa tietojoukkona jatkokäsittelyä varten.



Kuva 10. "Evaluate Model"-moduulin visualisointi

Regressio-mallien osalta “Evaluate Model”-moduulin tulos on havainnollistettu kuvassa 10. Tulos koostuu lasketuista tunnusluvuista ja kuvaajista. Tuloksen tarkka sisältö riippuu käytettävästä algoritmista. Esimerkissä moduuli on laskenut regressio-mallille mm. selitysasteen (R2), keskineliövirheen neliöjuuren (RMSE) ja keskimääräisen absoluuttisen virheen (MAE). Kuvaajana esitetään virhehistogrammi. (Microsoft 2017f)

5 Käytännön kokeet

Tämän tutkimuksen käytännön kokeessa kehitin FINRISKI-laskurin koneoppimisen keinoin CRISP-DM-menetelmän mukaisesti. CRISP-DM-menetelmän päävaiheet ovat kappaleen 3 mukaisesti liiketoiminnan ymmärtäminen, datan ymmärtäminen ja valmistelu, mallin kehittäminen, mallin arviointi ja mallin julkaisu.

5.1 Liiketoiminnan ymmärtäminen

Liiketoiminnan ymmärtäminen -vaiheessa kuvataan ratkaistava käyttötapaus, määritellään tavoitteet ja onnistumisen kriteerit sekä tehdään mallin kehittämiseen ja käyttöön liittyvän projektin suunnittelu. Käyttötapauksena tässä tutkimuksessa oli FINRISKI-laskurin toteutus koneoppimisen keinoin ja tavoitteena onnistunut toteutus Azure Machine Learning-palvelun avulla. Projektisuunnitelmana käytin tutkimussuunnitelmaa. Koska tässä tutkimuksessa kohteena oli valmis ja hyvin tunnettu laskuri, oli tämä vaihe tässä tutkimuksessa hyvin rajallinen.

5.1.1 FINRISKI-laskuri

FINRISKI-laskuria voi käyttää sydän- ja verisuonitautien kokonaisriskin arvioinnin tukena. Laskuria voivat käyttää sekä terveydenhuollon ammattilaiset että kansalaiset. Laskuri ei kuitenkaan ole diagnostinen työväline (ammattilaiset eivät voi käyttää ainoana välineenä arvioidessaan henkilön terveydentilaa). Laskuri auttaa havainnollistamaan miten erilaiset muuttujat vaikuttavat sairastumisriskiin. FINRISKI-laskuri perustuu vuosina 1982, 1987 ja 1992 FINRISKI-tutkimuksessa tutkittujen henkilöiden riskitekijätietoihin ja sairastuvuuden seurantaan.

FINRISKI-laskuri laskee riskin sairastua sydäninfarktiin, riskin sairastua vakavaan aivoverenkiertohäiriöön, näiden yhteisen riskin eli riskin saada jompikumpi näistä sairauksista seuraavan kymmenen vuoden aikana. Riskin laskennassa käytetään seuraavia tietoja: ikä, sukupuoli, kokonaiskolesteroli, HDL-kolesteroli, tupakoiko tällä hetkellä, systolinen verenpaine sekä tieto siitä, sairastaako henkilö tyypin 2 diabetesta ja onko kumpikaan henkilön vanhemmista saanut sydäninfarktia alle 60-vuotiaana. (THL 2014)

5.2 Datan ymmärtäminen ja käsittely

Datan ymmärtämisen ja käsittelyn vaiheessa selvitetään ensimmäisenä, mitä omia ja ulkoisia tietolähteitä on käytössä ja mitä tietoa näistä löytyy. Koottu aineisto analysoidaan datan laadun ymmärtämiseksi ja mahdollisten laatuongelmien havaitsemiseksi. Data käsitellään ja korjataan niin, että se on riittävän laadukasta mallin opettamiseen. Tähän vaiheeseen kuuluu myös piirteiden valinta niin, että mallin opettaminen olisi mahdollisimman tehokasta. Koska tämän tutkimuksen kohteena oli tunnetun laskurin toteuttaminen, ei tässä tutkimuksessa syvennytty tarkemmin piirteiden valintaan, vaan käytettiin samoja piirteitä kuin mitä nykyinen FINRISKI-laskuri käyttää.

5.2.1 Aineiston suunnittelu

Koneoppiminen ja avoin data -ilmiöiden yleistymisen myötä on maailmalle syntynyt useita hyviä, tähän tarkoitukseen soveltuvia tietoarkistoja. Tämän tutkimuksen koetta varten käytin University of California:n ylläpitämässä palvelussa löytyvää aineistoa. Tästä tietopalvelusta löytyy heidän omien tietojen mukaan 360 eri koneoppimisen tutkimiseen soveltuvaa tietoaineistoa. Tietopalvelu on perustettu 1987 ja on laajasti koneoppimisen tutkijoiden ja opiskelijoiden käytössä. Sen aineistoihin on viitattu tieteellisissä julkaisuissa yli 1000 kertaa. Palvelun aineistoista löytyy esimerkiksi autoihin, henkilöiden tuloihin, metsäpaloihin ja erilaisiin tauteihin liittyviä tietokantoja. (UCI 2017a)

Kokeen tietolähteenä käytin palvelun “Heart Disease Data Set”-aineistoa. Aineistoon on koottu 75 ominaisuutta kustakin aineistossa olevasta henkilöstä sisältäen kaikki tähän kokeeseen tarvittavat tiedot paitsi HDL-kolesteroli. Aineistossa on mukana tietoja unkarilaisista, sveitsiläisistä ja yhdysvaltalaisista henkilöistä. Aineisto on anonymisoitu eli siitä on poistettu henkilötiedot. Kokeessa käytettävät aineiston tiedot on kuvattu liitteessä 1.

5.2.2 Datan laadun varmistaminen

Datan laadun selvittämiseksi käsittelin tutkimukseen kootun datan kappaleessa 3 kuvattujen tekniikoiden avulla puuttuvien, odottamattomien kardinaliteettien ja poikkeavien arvojen tunnistamiseksi. Datassa paljastui korjattavia laatuongelmia. Merkittävin oli aineistoissa käytetty tapa merkitä puuttuvat arvot numerolla -9. Kaikki havaitsemani laatuongelmat merkitsin laaturaportissa korjattaviksi. Analyysissa paljastui myös, että kerätyssä aineistossa henkilöiden ikäväli oli laajempi kuin mitä kehitettävälle laskurille soveltui ja poistin aineistosta sopivan ikävälin ulkopuoliset alkiot. Lisäksi analyysi paljasti muutaman piirteen olevan aineistossa jatkuvina piirteinä ja näiden vaativan muuttamisen luokitteleviksi piirteiksi.

Käsittelin seuraavaksi aineiston laaturaporttiin merkittyjen havaintojen ja toimenpiteiden mukaisesti. Korjaavien toimenpiteiden jälkeen muodostin aineistosta korjatun version ja suoritin laatutarkastuksen uudestaan. Uusintatarkistuksessa en enää havainnut laatuongelmia ja data oli valmis käytettäväksi mallin kehittämiseksi. Tämän vaiheen havainnot ja toimenpiteet on kuvattu tarkemmin liitteessä 2.

5.3 Mallin kehittäminen

Ohjatun oppimisen mallin kehittämiseen kuuluvat algoritmin valinta, sen parametrien viritttäminen ja mallin opettaminen osalla datasta. Yleinen nyrkkisääntö on käyttää 75 % datasta opettamiseen ja 25 % mallin arviointiin.

Algoritmien vertailemiseksi kehitin kokeen, jossa edellä muodostetulla datalla opetetaan useampi algoritmi ja näiden tulokset vertaillaan keskenään. Kokeen tavoitteena oli löytää se algoritmi, joka tuottaa tämän tutkimuksen aineistolla tarkimman tuloksen. Lisäsin edellä laatuvarmistetun datan kokeeseen ja siitä valitsin kokeessa käytettävät piirteet. Jaoin kokeessa datan jaolla 75/25 opetus- ja testijoukoiksi ja asetin jokaisen vertailtavan algoritmin “Train model”-moduulin käyttämään datan risk-piirrettä oppimisessa.

Valitsin tähän kokeeseen vertailtavaksi lineaarinen regressio-, päätöspuu- ja neuraaliverkko-algoritmit. Tässä kokeessa käytin kaikkia algoritmeja niiden oletusarvoilla enkä yrittänyt parantaa algoritmien tuloksia parametreja virittämällä.

	Linear Regression	Boosted Decision Tree Regression	Neural Network Regression
Keskimääräinen absoluuttinen virhe (MAE)	4.91385	3.416778	0.732782
Keskineliövirheen neliöjuuri (RMSE)	6.278349	4.632938	1.136869
Selitysaste (R2)	0.943926	0.969466	0.998161

Taulukko 2. Regressio-algoritmien vertailun tulokset

Kappaleessa 3 käsiteltiin koneoppimisen regressiomallin tunnuslukuja ja niiden merkitystä, jossa todettiin hyvässä mallissa selitysasteen (R2) olevan lähellä arvoa 1 ja keskineliövirheen neliöjuuren (RMSE) ja keskimääräisen absoluuttisen virheen (MAE) olevan niin lähellä arvoasteikon minimiarvoa kuin mahdollista. RMSE- ja MAE-tunnuslukujen vertailussa on huomioitava mallin tulospiirteen arvoasteikko [0,100]. Parhaat tunnuslukujen arvot olivat neuraaliverkko-algoritmilla, jolla selitysaste oli 0,998 ja kaksi muuta tunnuslukua olivat myös parhaat. Päätöspuu-algoritmilla vastaavasti selitysaste oli 0,969 ja lineaarinen regressio-algoritmilla se oli 0,944. Näiden tulosten perusteella valitsin kehitettävälle laskurille neuraaliverkko-algoritmin. Vertailun tulokset on koottu yllä taulukkoon 2.

Laatuvarmistetun datan ja valitun algoritmin avulla seuraava vaihe oli kehittää koneoppimisen malli. Tämä koe noudatti edellä kuvattua mallia, jossa otosjoukko jaetaan opetus- ja testijoukoiksi ja opetettu malli luodaan kytkemällä valittu algoritmi ja opetusjoukko ”Train model”-moduuliin. Opetusmoduulin tuloksena on opetettu malli.

5.4 Opetetun mallin arviointi

Mallin arvioinnissa tavoitteena on arvioida ja ymmärtää miten tarkasti malli ennustaa eli miten usein opetettu malli ennustaa oikein ja miten usein väärin. Regressiota käytettäessä hyviä tilastollisia suureita ovat selitysaste ja keskineliövirheen neliöjuuri.

	Neural Network Regression
Keskimääräinen absoluuttinen virhe (MAE)	0.948679
Keskineliövirheen neliöjuuri (RMSE)	1.402157
Selitysaste (R2)	0.99719

Taulukko 3. Opetetun mallin tuloksen tunnusluvut

Tässä kokeessa kehitetyn mallin arvioinnissa mallin selitysaste oli 0,997 eli hyvin lähellä tavoiteltua arvoa 1. Opetetun mallin arvioinnin tunnusluvut on havainnollistettu taulukossa 3. Tässä kohdassa on hyvä huomata, että tulos ei ole täsmälleen sama kuin algoritmien vertailussa saavutettu tulos. Tämä johtuu siitä, että opetustilanteissa käytettävät tiedot valitaan aineistosta satunnaisesti ja tästä johtuen tulokset ovat eri kerroilla hiukan toisistaan poikkeavat.

5.5 Mallin julkaisu

Kun opetettu malli on valmis ja täyttää sille asetut vaatimukset, on viimeinen vaihe julkaista se käytettäväksi. Kuten kappaleessa 3 todettiin, tämä voi yksinkertaisimmillaan olla mallin käyttö taulukkolaskentaohjelmassa kuten Excel ja toisessa monimutkaisimmassa ääripäässä mallia käytetään reaaliaikaisesti esimerkiksi Apache Spark-ympäristössä.

Kun opetettu malli oli valmis, tein Azure Machine Learning-palvelun vakiotoiminnallisuuksilla opetetusta mallista REST Web Service-palvelun. Tämä vaihe on lähes täysin automaattinen ja ainoa olennainen toimenpide oli tarkistaa tulo- ja lähtötietojen nimet ja muuttaa ne omaan tyyliin sopiviksi. Koneoppimisen keinoin toimiva laskuri oli tämän jälkeen valmis käytettäväksi sovelluspalveluissa riskin ennustamiseen.

Azure Machine Learning-palvelun rajapintojen hallintasivuston avulla pystyin testaamaan mallin käyttöä rajapinnan avulla. Hallintasivusto sisältää hyvät työkalut kokeen mallista luodun Web Services-rajapinnan testaamiseen ja selkeät ohjeet rajapinnan käyttämiseen. Palvelusta voi myös ladata rajapinnan swagger-tiedoston, jolla pystyy helposti luomaan tarvittavan koodin useimpiin nykyaikaisiin ohjelmointikieliin. Lisäksi käyttäjälle tarjotaan useampi ohjattu opas mm. rajapintaa käyttävän sovelluksen tekemiseksi ja opetetun mallin uudelleenopettamiseksi esimerkiksi, kun saatavilla on uudempaa dataa.

6 Pohdinta

Tämän tutkimuksen aiheena oli selvittää, miten koneoppimisen tekniikoita voi käyttää laskurin tuloksen laskemiseksi ilman työlästä perinteisten algoritmien kehittämistä manuaalisesti. Keskeisinä tavoitteina työssä olivat selvittää miten helposti ja hyvin koneoppimisen keinojen avulla pystyy toteuttamaan käytännön osuuden kohteena olleen laskurin sekä selvittää, miten hyvin tämä toteuttaminen onnistuu MLaaS-palvelussa ilman, että tarvitsee asentaa tarvittavat ympäristöt ja työkalut omille tietokoneilleen.

Itse ennustavan mallin tekeminen toteuttaminen oli varsin helppoa ja suoraviivaista, mutta toisaalta datan laadun varmistamisen merkitystä en ollut osannut ennakoida riittävästi. CRISP-DM-menetelmään tutustuminen ja sen noudattaminen olivat todella hyviä asioita tämän työn kannalta. Menetelmän avulla koneoppimisessa erittäin olennaiset datan ymmärtäminen ja datan laadun varmistaminen toteutuivat työssä kunnolla, eikä lopputuloksena syntynyt näennäisesti toimiva malli.

Mallin toimivuuden arvioinnissa käytetään yleisiä tilastotieteen tunnuslukuja. Data-analyytikolle onkin erittäin tärkeä tuntea ja osata tilastotiedettä ja sen käsitteitä ja menetelmiä kunnolla. Suosittelen jokaiselle koneoppimisen opettelua miettivälle ensin varmistaa oma tilastotieteiden osaaminen vähintään tämä työn kappaleessa 3 esitellyssä laajuudessa.

Azure Machine Learning-palvelun osalta oma odotusarvo palvelun tasosta ei ollut korkea, koska usealla asiantuntijafoorumilla on esitetty paljon kritiikkiä MLaaS-palveluja kohtaan. Kritiikissä väitetään usein MLaaS-palvelujen toteuttamien ja tukemien ominaisuuksien olevan vain suppea osajoukko siitä kaikesta, mitä koneoppimisen keinoin on mahdollista. Tästä syystä olin erittäin myönteisesti yllättynyt, koska Azure Machine Learning -palvelu tekee käytännössä kaikista perusasioiden tekemisestä helppoa, mutta sisältää myös tuen tehdä R- ja Python-skriptien avulla kaiken sen, mihin ei vielä ole suoraa tukea itse palvelussa. Minun arvioni on, että Azure Machine Learning ja vastaavan kaltaiset MlaaS-palvelut tulevat muuttamaan vallitsevan tilanteen samalla tavalla kuin on jo käynyt muilla tietotekniikan aloilla siirtymisessä pilvipalveluihin. On hyvä pitää mielessä se tosiasia, että datan ymmärtämisessä toimialatuntemuksella on ja tulee aina olemaan iso merkitys. Jos toimiala-asiantuntija ja koneoppimisen asiantuntija pystyvät kehittämään yhdessä koneoppimisen malleja helppokäyttöisessä palvelussa, helpottaa tämä yhteistyötä merkittävästi.

Työn tekemisen yhteydessä heräsi mielenkiinto jatkaa tutkimista usealla eri tavalla. Mallin arvioinnin osalta kirjallisuudessa on esitetty kritiikkiä ja varoituksia yleisimmin käytettyjä

selitysaste (R2) ja keskineliövirheen neliöjuuri (RMSE) kohtaan. Näiden lisäksi voisi selvittää tarkemmin esimerkiksi muokatun R2-luvun käyttöä tunnuslukuna. Algoritmien parametrien virittäminen esimerkiksi ristiinvalidointia hyödyntäen olisi myös mielenkiintoinen jatkoaihe.

Tässä työssä keskityttiin ohjattuun oppimiseen ja regressio-algoritmien käyttöön. Ohjatun oppimisen puolella löytyy paljon mahdollisia jatkoaiheita, kuten luokittelevien ja poikkeamia tunnistavien algoritmien käyttö. Ohjaamattoman oppimisen puolella algoritmien kyky jäsentää ja ryhmitellä tietoa tuntuu hyvin mielenkiintoiselta aiheelta. Yhteenvetona voi sanoa, että tämä tutkimus tuntuu vasta herkulliselta pintaraapaisulta koneoppimisen aiheeseen ja vielä on paljon tutkittavaa ja opittavaa tämän erittäin mielenkiintoisen ja ajankohtaisen aihealueen tiimoilta.

Lähteet

Begane, R. 2017. Quora/ Rakshith Begane kommentti. Luettavissa: https://www.quora.com/What-are-the-best-machine-learning-as-a-service-MLaaS-companies-and-startups. Luettu: 15.4.2017

Eskelinen, S. 2017. Glukoosi. Duodecim Terveyskirjasto. Luettavissa: http://www.terveyskirjasto.fi/terveyskirjasto/tk.koti?p_artikkeli=snk03091. Luettu: 13.4.2017.

Espoo 2017. Robottibussit liikenteeseen Otaniemessä. Luettavissa: http://www.espoo.fi/fi-FI/Robottibussit_liikenteeseen_Otaniemessa(99596). Luettu: 28.1.2017.

Ferguson, M. 2016. What is machine learning? Making the complex simple. IBM Big Data Hub Blogs. Luettavissa: http://www.ibmbigdatahub.com/blog/what-machine-learning. Luettu: 17.2.2017.

Grassegger, H., Krogerus, M. 2017. The data that turned the world upside down. Luettavissa: http://motherboard.vice.com/read/big-data-cambridge-analytica-brexit-trump. Luettu: 28.1.2017.

Hackenberg, J. 2.9.2016. Stackexchange-kommentti. Luettavissa: https://stats.stackexchange.com/questions/38631/rmse-vs-coefficient-of-determination. Luettu: 15.4.2017.

Harrington, P. 2012. Machine Learning in Action. Manning Publications.

Kanowitz, S. 2017. Machine learning tool helps county detect cyber risks. Luettavissa: https://gcn.com/articles/2017/01/25/darktrace.aspx. Luettu: 28.1.2017.

Kelleher, J., Mac Namee, B. and D'Arcy, A. 2015. Fundamentals of Machine Learning for Predictive Data Analytics. MIT Press.

Kohavi, R., Provost, F. 1998. Glossary of terms. Machine Learning, 30, 271–274. Kluwer Academic Publishers, Boston. Luettavissa: http://ai.stanford.edu/~ronnyk/glossary.html. Luettu: 16.2.2017.

Marsland, S. 2015. Machine Learning: An Algorithmic Perspective, Second Edition. CRC Press.

Mathworks 2016. Introducing Machine Learning. Luettavissa: https://www.mathworks.com/campaigns/products/offer/machine-learning-with-matlab.html. Luettu: 17.2.2017.

Microsoft 2017a. How to choose algorithms for Microsoft Azure Machine Learning. Luettavissa: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-algorithm-choice. Luettu: 13.2.2017

Microsoft 2017b. Feature Selection Modules. Luettavissa: https://msdn.microsoft.com/en-us/library/azure/dn905912.aspx. Luettu: 29.4.2017.

Microsoft 2017c. Azure Machine Learning frequently asked questions. Luettavissa: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-faq. Luettu: 28.1.2017.

Microsoft 2017d. Introduction to Azure Machine Learning in the cloud. Luettavissa: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-what-is-machine-learning. Luettu: 28.1.2017.

Microsoft 2017e. Machine Learning Module Descriptions. Luettavissa: https://msdn.microsoft.com/en-us/library/azure/dn906013.aspx. Luettu: 29.4.2017.

Microsoft 2017f. Machine Learning - Evaluate. Luettavissa: https://msdn.microsoft.com/en-us/library/azure/dn906026.aspx. Luettu: 29.4.2017.

Pyle, D., San Jose, C. 2015. An executive's guide to machine learning. Luettavissa: http://www.mckinsey.com/industries/high-tech/our-insights/an-executives-guide-to-machine-learning. Luettu: 11.02.2017.

Quora 2017. What are the best machine learning as a service (MLaaS) companies and startups? Luettavissa: Quora 2017: https://www.quora.com/What-are-the-best-machine-learning-as-a-service-MLaaS-companies-and-startups. Luettu: 15.4.2017.

Román, J. 2016. CRISP-DM: The methodology to put some order into Data Science projects. Luettavissa: https://data.sngular.team/en/art/40/crisp-dm-the-methodology-to-put-some-order-into-data-science-projects. Luettu 15.2.2017.

Samuel, A. 1959. Some studies in machine learning using the game of checkers. IBM Journal of research and development. Luettavissa: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.368.2254&rep=rep1&type=pdf. Luettu 16.2.2017.

Techopedia 2017. Machine Learning as a Service (MLaaS). Luettavissa: https://www.techopedia.com/definition/32434/machine-learning-as-a-service-mlaas. Luettu. 15.4.2017.

THL 2014. FINRISKI-laskuri. Luettavissa: https://www.thl.fi/fi/web/kansantaudit/sydan-ja-verisuonitaudit/finriski-laskuri. Luettu: 8.2.2017.

UCI 2017a. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. Luettavissa: https://archive.ics.uci.edu/ml/. Luettu: 12.3.2017.

UCI 2017b. UCI Machine Learning Repository Heart Disease Data Set. Irvine, CA: University of California, School of Information and Computer Science. Luettavissa: https://archive.ics.uci.edu/ml/datasets/Heart+Disease. Luettu: 12.3.2017.

Wikipedia 2017a. CRISP-DM. Luettavissa: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining. Luettu 15.2.2017.

Wikipedia 2017b. Feature selection. Luettavissa: https://en.wikipedia.org/wiki/Feature_selection. Luettu 29.4.2017.

Liitteet

Liite 1. Tutkimuksen aineiston kuvaus

Nimi: Heart Disease Data Set
Piirteiden määrä/tyypit: 75. Luokittelevia, Jatkuvia kokonais- ja liukulukuja
Päivämäärä: 1988-07-01

Käytetyt tiedot	**Aineiston kenttä ja tarkempi kuvaus**
Sukupuoli	Kenttä #4 sex, jonka arvo kuvauksen mukaan on 1 jos mies ja 0 jos nainen.
Ikä	Kenttä #3 age, jonka arvo kuvauksen mukaan on yksikössä vuosia.
Kokonaiskolesteroli-arvo	Kentässä #12 chol, jonka arvo kuvauksen mukaan on yksikössä mg/dl. Tieto muunnetaan SI-yksikköön mmol/l.
HDL-kolesteroliarvo	Tietoa ei löydy aineistossa ja tämä jätetään tyhjäksi.
Tupakoiko henkilö	Kenttä #13 smoker. Arvo 1 kyllä tai 0 ei tupakoi.
Tupakkaa/päivä	Kenttä #14 cigs. Numeerinen arvo.
Tupakointivuodet	Kenttä #15 years. Numeerinen arvo.
Systolinen veren-paine	Kenttä #10 trestbps, jossa kuvauksen mukaan ammattilaisen mit-taama systolinen verenpaine yksikössä mmHg.
Korkea paastove-rensokeri	Kenttä #16 fbs, jossa tieto onko paastoverensokeri ylittänyt rajan 120 mg/dl. Arvo on 1 jos ylittänyt ja 0 jollei. Arvo 120 mg/dl on yhtä kuin 6.7 mmol/l. Suomessa jos viitearvo 7.0 ylitetään useammassa toistetussa paastomittauksessa, on kyseessä diabetes (Eskelinen, S. 2017).
Diabetes-historia	Aineiston kenttä #17 dm, jonka arvo kuvauksen mukaan on 1 jos diabetes- historia ja 0 jollei ole.
Sydäntauti-historia	Kenttä #18 family history, jossa arvo 1 jos esiintyy ja arvo 0 jollei ole.

Liite 2. Tutkimuksessa datan laadun varmistamisen kokeet

Koe: ONT-data-0-raw

Data	new.raw.0.csv
Otoskoko	1541
Tavoite	Tarkistaa alkuperäinen aineisto ja löytää korjattavat laatuongelmat.

Aineiston tiedot	**Havainnot**	**Toimenpiteet**
age	Data OK. Arvot välillä [20,78].	
sex	Data numeerisena piirteenä kun pitäisi olla luokitteleva. Kardinaliteetti oikein 2.	Muunnetaan luokittelevaksi piirteeksi.
trestbps	Virheellinen minimiarvo -9.	Data korjattava.
chol	Negatiivinen minimiarvo -0.5	Data korjattava.
smoker	Virheellinen minimiarvo -9. Data numeerisena piirteenä kun pitäisi olla luokitteleva.	Data korjattava. Muunnetaan luokittelevaksi piirteeksi.
smokercigs	Virheellinen minimiarvo -9.	Data korjattava.
smokeryears	Virheellinen minimiarvo -9.	Data korjattava.
fbs	Virheellinen minimiarvo -9. Data numeerisena piirteenä kun pitäisi olla luokitteleva.	Data korjattava. Muunnetaan luokittelevaksi piirteeksi.
dm	Data numeerisena piirteenä kun pitäisi olla luokitteleva. Kardinaliteetti arvolla 0 korkea.	Muunnetaan luokittelevaksi piirteeksi.
famihist	Virheellinen minimiarvo -9. Data numeerisena piirteenä kun pitäisi olla luokitteleva.	Data korjattava. Muunnetaan luokittelevaksi piirteeksi.

hdl	Piirre on jätetty kokonaan tyhjäksi aineistossa.	Poistetaan opetusjoukosta, koska ei ole yhtään arvoa.
risk	20 alkiossa ei ole laskettua riskiarvoa.	Selvitetään miksi näissä muutamassa alkiossa on tyhjä arvo. Todennäköisesti virhe johtuu siitä että laskennassa ei riittävän hyvin tarkisteta käytettävien arvojen oikeellisuutta ja tästä syystä näissä muutamassa laskenta epäonnistuu.

Yhteenveto toimenpiteistä:

- trestbps, chol, smoker, smokercigs, smokeryears, fbs, famihist. Piirteen arvo -9 tulkitaan puuttuvaksi arvoksi ja muunnetaan datassa tyhjäksi arvoksi. Tässä tutkimuksessa ei selvitetä tarkemmin syitä miksi näitä arvoja puuttuu ja yksinkertaisuuden vuoksi rivit, joilla puuttuvia arvoja suodatetaan pois aineistosta paitsi tupakoinnin osalta jonka osalta puuttuvat tulkitaan ei-tupakoiviksi.
- risk. Tutkitaan miksi datassa 20 puuttuvaa arvoa. Tutkinnan tuloksena selvisi että riskilaskennassa kertoimet vain ikävälille [30,74] joten rajataan aineiston tälle ikävälille.
- hdl. Jätetään kokonaan pois aineistosta.
- sex, smoker, fbs, dm, famihist. Piirteet muunnetaan luokitteleviksi piirteiksi.
- Tämän jälkeen suoritetaan uusi datan laadun arviointi.

Koe: ONT-data-1-raw

Data:	new.raw.1.csv
Otoskoko	547
Tavoite	Tarkistaa aineiston ensimmäinen versio ja löytää selkeimmät korjattavat laatuongelmat.

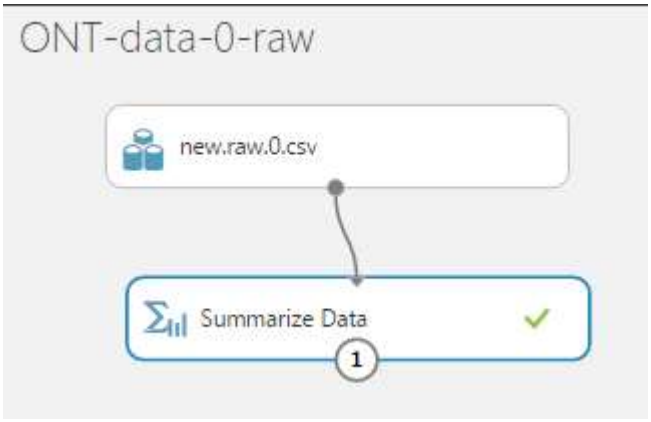
Aineiston tiedot	**Havainnot**
age	Data OK. Arvot välillä [33,74].
sex	Data OK. Kardinaliteetti 2.

trestbps	Data OK. Arvot välillä [92,200].
chol	Arvot välillä [5,56, 31,3]. Histogrammin perusteella poikkeavia arvoja arvojen yläpäässä. Yläkvartiilin ja maksimin erotus on suuri.
diabetes	Data OK. Kardinaliteetti 2.
fbs	Data OK. Kardinaliteetti 2.
dm	Data OK. Kardinaliteetti 2.
infarct	Data OK. Kardinaliteetti 2.
risk	Data OK. Arvot välillä [0.73, 99.83]

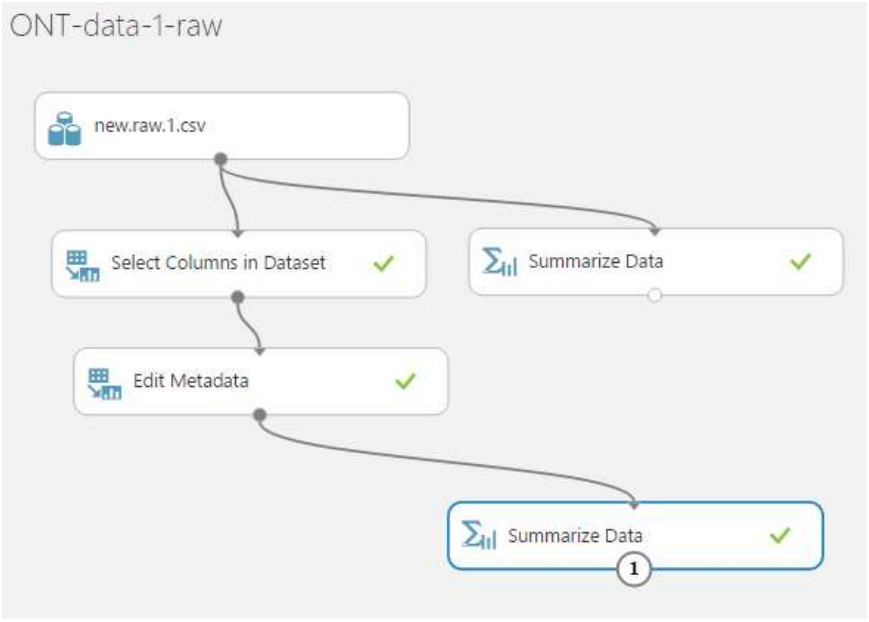
Yhteenveto toimenpiteistä:

Datan laatu on riittävä tutkimuksen mallin opettamiseen eikä dataa tarvitse korjata eikä käsitellä.

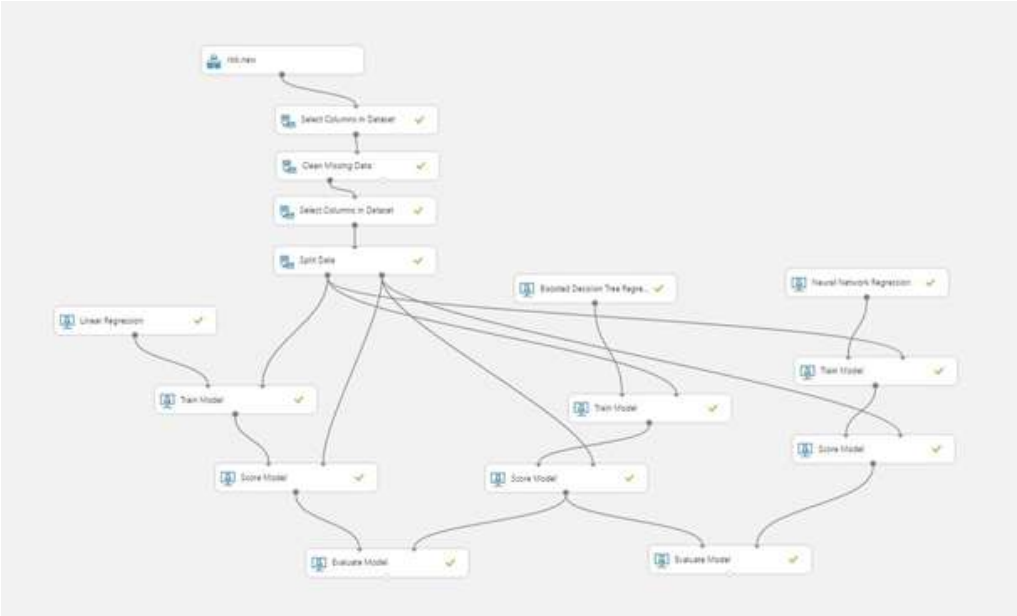
Liite 3. Azure ML-kokeiden havainnekuvat



Kuva 11. Datan laadunvarmistamisen 1. koe



Kuva 12. Datan laadun varmistamisen 2. koe



Kuva 13. Algoritmien vertailukoe

Kuva 14. Mallien opettamisen koe